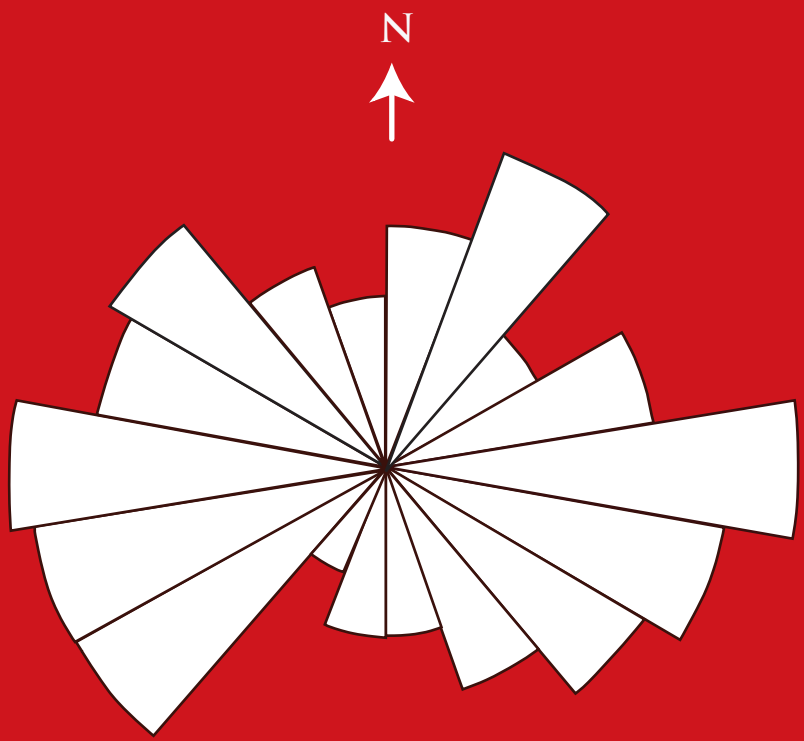
The Epistemology of Statistical Science



Mauritz van Aarde

The Epistemology of Statistical Science

Mauritz van Aarde

The Epistemology of Statistical Science

Published by SUN PReSS, an imprint of SUN MeDIA Stellenbosch, Stellenbosch, 7600
www.africansunmedia.co.za
www.sun-e-shop.co.za

First edition 2009

Revised edition 2010

ISBN: 978-1-920338-32-9
e-ISBN: 978-1-920338-33-6
DOI: 10.18820/9781920338336

Set in 10/12 Constantia
Cover design by SUN MeDIA Stellenbosch
Typesetting by SUN MeDIA Bloemfontein

SUN PReSS is an imprint of SUN MeDIA Stellenbosch. Academic, professional and reference works are published under this imprint in print and electronic format. This publication may be ordered directly from www.sun-e-shop.co.za

Printed and bound by SUN MeDIA Stellenbosch, Ryneveld Street, Stellenbosch, 7600.

For the memory of Oscar Kempthorne

Contents

Preface

In the usage of present-day statistics 'statistical inference' is a profoundly ambiguous expression. In some literature a statistical inference is a 'decision made under risk', in other literature it is 'a conclusion drawn from given data', and most of the literature displays no awareness that the two meanings might be different. This book concerns the problem of drawing conclusions from given data, in which respect we have to ask: Does there exist a need for the term 'statistical inference'? If so, does there also exist a corresponding need for every other science? If so, how does, for example, agronomy then manage to reason in terms of botanical inference, soil scientific inference, meteorological inference, biochemical inference, molecular biological inference, entomological inference, plant pathological inference, etc. without incoherence or self-contradiction? Consider the possibility that agronomy does not reason in terms of such a motley of special kinds of inference. Consider the possibility that, apart from subject matter, botany, soil science, entomology, etc. all employ the same kind of reasoning. If so, must we then believe that statistics, alone among all the sciences, is the only one that requires its own special kind of inference?

Starting with Thomas Bayes (1763) the statistical profession has by and large believed that statistics requires a kind of inference of its very own. However, the belief does not rest on clear agreement as to what precisely the term 'inference' is supposed to mean, and so it has brought about confusion of which it can only be said: There is none so great as a learned one. There are no fewer than four different schools of thought identifiable as advocates of frequentist inference, Bayesian inference, likelihood inference and fiducial inference, respectively. Even amongst these there are further disagreements. All Bayesians for instance proceed from so-called prior probabilities, but are unable to agree as to whether such probabilities are 'logically' determined (Jeffreys 1961) or 'subjectively arrived at' (Savage 1954, 1962; Lindley 1965). Again, Fraser (1968) advocates structural inference, but does not make it clear whether or how that might differ from fiducial inference. And yet again, some frequentists embrace randomised hypothesis tests, whilst such tests are anathema to other frequentists. Then there are statisticians who refuse to admit to the existence of any such confusion. Along these lines a silly campaign has even urged us to be proud that statistics, unlike other sciences, 'is not so simple a subject as to admit only one correct answer to any given question'. Clearly then, some two and a half centuries of debate and development has failed to produce consensus. So it is entirely reasonable to ask of the different schools of thought that instead of dwelling on the disagreements that divide them, they seriously consider whether in fact they might not be united in mutual error.

The present book proceeds from the premise that despite the vast variety of its subject matter all science is based on the same fundamental principles of reasoning. Statistics differs from the rest only in its subject matter, and so must learn from other, much older sciences, how to reason. We must go back to the very outset and carefully, step by step, learn from our customers in the substantive sciences how to proceed. In order to do that, we have to understand that it is the principles of scientific reasoning, rather than mathematical reasoning, that we must grasp. We must be extremely careful not to foist some peculiarly statistical ideas upon the discourse of substantive science. In other words, whatever ideas we try to develop must manifestly originate from all the other sciences together. That is the only way in which we can hope to clear up the confusion into which we have fallen.

Clearly, that will require a discourse that spans the interface between statistics and substantive science. We need to involve, not only statisticians, but also our customers in the substantive sciences, as statistics can serve no purpose other than to be of service to substantive science. Ultimately then, it is our customers who must judge our contribution. With that in mind, the present book tries to involve a wide audience, and so, unavoidably, might then to a statistician seem pedestrian in its attempts to explain statistical matters, and might then to a substantive scientist seem pedestrian in its attempts to explain substantive matters. In this we can but beg the reader's indulgence.

Acknowledgements

It is impossible to achieve anything without relying on other people, and it would be impossible to list all those who, in numerous, and often humble ways enabled me. I can but give a woefully incomplete list.

My interest in the subject matter of the present book was first stimulated as a student, initially by S.J. ('Faantjie') Pretorius, and subsequently by Oscar Kempthorne, both of whom influenced me toward the view it expounds. F.X. Laubscher taught me that a scientist must always have an open mind without indulging defective reasoning.

I am grateful for the support of colleagues over many years: Bill Louw, Jeanne Heyman, John Randall, Ben Eisenberg, Frikkie Calitz, Marietta van der Rhijst and Mardé Booyse.

My deepest gratitude to Professor Jannie Hofmeyr of Stellenbosch University and Professor Christine Thiart of the University of Cape Town for reviewing the manuscript.

Last but not least, I thank my darling wife Inge for her steadfast encouragement and help.

Chapter 1

COMMENCEMENT TESTS

Populations being brought into the human mind

1.1 Introduction

This chapter concerns the situation where we take first steps toward trying to make statistical sense, so to speak, of a given set of raw data. The data will generally be one of two different types. One type takes the form of a sequence of results, where we would then want to establish whether or not the sequence could be represented as the outcome of a specific class of stochastic processes. Suppose for instance that the following sequence is a record of apparent success (S) or failure (F) in nine consecutive responses by a particular animal in a learning trial:

F, F, F, S, F, S, F, S, S. (1.1.1)

We might ask whether the sequence involves a trend or, alternatively, whether it could more simply be represented as a random sample from a specific class of populations. A second type of data has no sequential structure, where we might then more directly ask whether the data could be represented as a random sample from a specific class of populations. Consider for instance the data in Table 1.1.1, giving, for each of two groups of fruit trees, the measured half-life of their fruit. In this case we might ask whether or not each group of measurements could be represented as a random sample from a normal population.

Table 1.1.1: Half-life, in days, of the fruit of ten trees in a completely randomised design, comprising five replications each of a carbaryl treatment and control

Trees treated with carbaryl	Untreated controls
11.9 12.8 13.1 13.1 14.4	8.8 10.8 11.1 11.2 11.4

In trying to deal with these problems we almost always begin by plotting the data in such a way that a proposed representation can be visually judged for its tenability. For instance, when the data given at (1.1.1) are plotted as in Figure 1.1.1 overleaf, a slight trend toward an increased frequency of success is made visually apparent. Similarly, each group of half-life measurements might be ordered from smallest to largest and then plotted against the expected values of the corresponding standard normal order statistics (Figure 1.1.2). The human body is thereby enabled to visually grasp and to analytically judge the tenability of the proposed model. We may ask, for instance, as a matter of visual judgement of the data plots in Figure 1.1.2:

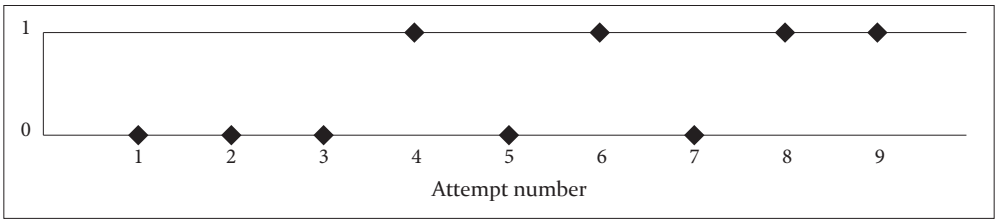


Figure 1.1.1: Success (1) or failure (0) in 9 consecutive attempts in a learning trial

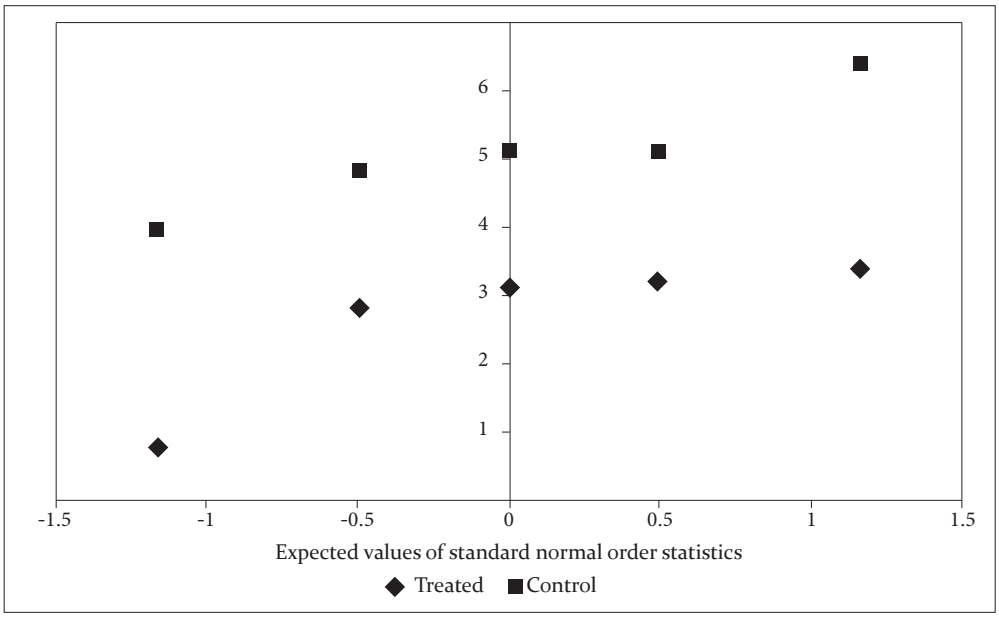


Figure 1.1.2: Order-statistical plot of the half-lives, in days minus 8, of the fruit of 10 trees

Can each plot be represented as a sample of points scattered around a straight line? If so, can the plots be represented as scattered around two parallel lines?

This example shows how physical experience and mathematical reasoning interplay to produce a refinement of primitive method. Here primitive method might try to judge the 'shapes' of the two distributions of half-lives by way of two histograms; but on second thoughts we realise that for so small a data set we need to refine the primitive method.

Clearly then, statistical data analysis concerns the development of statistical models for the representation of certain kinds of data, and very early on data analysts began to experience a need for refined methods to test the adequacy of such models (Arbuthnott, 1710). However, the systematic development of such tests only began in the 20th century. Significance tests originated in the test for isotropic directions of Raleigh (1880), the χ^2 test of Pearson (1900) and the *t* test of Student (1908). During the next 20 years R. A. Fisher developed many significance tests. Subsequently Neyman and Pearson (1933) introduced hypothesis tests. Despite sharing much mathematical common ground, the

two kinds of tests seek to implement fundamentally different ideas, where the difference has not at all been widely understood. Lehmann (1986) has given an excellent account of the formal mathematics of hypothesis tests. Kempthorne and Folks (1971) have given a definitive account of significance tests and how they differ from hypothesis tests.

This chapter takes the first steps toward the development of a new kind of test. The reader will find the development drawing on ideas that originate in Fisher (1970), first published in 1925. The reader should, however, be wary of taking the development to be a re-invention of significance tests, because as subsequent chapters will show, the new tests differ profoundly from significance tests, so much so that the reader will ultimately be compelled to take a stance on a devastating outcome. It will nevertheless be found that significance tests and co-ordination tests, as we will name the new kind of tests, share so much common ground that an economy of presentation is achieved by drawing on existing literature. To that end we draw primarily on Kempthorne and Folks (1971) and Cox and Hinkley (1974).

We will present a variety of examples to motivate the introduction of certain definitions and theorems. We will in fact risk a redundancy of such examples, as the introduction of unfamiliar ideas might well require some repetitiveness for their clarification.

As stated above, in this chapter we will be dealing with the very first steps required for the statistical modelling of given data. The further development, and the uses and usefulness of such models, are discussed in subsequent chapters. However, before proceeding to the development of any ideas about statistical data analysis, we must first examine the nature of the scientific discourse that statistical data analysis is supposed to serve, otherwise we risk trying to foist inappropriate statistical inventions onto the discourse of substantive science. This must be firmly grasped, as the development of modern statistics largely took place in the 20th century, long after the substantive sciences that it wishes to serve were already well developed. The point here is that long before the advent of modern statistics, individuals such as Kepler, Galileo, Newton, Mendel and many others, had already developed a huge body of scientific knowledge. Moreover, a great deal of their work rested on analyses of just the kind of data that we now look upon as requiring the expertise of mathematical statistics. So rather than try to tell our customers from the substantive sciences about the principles of scientific data analysis, we should accept that *they* developed those principles in the first place. That is not to say that we should not try to develop their methods for application to the statistical case, but only that we must try to understand those methods before trying to develop them further. In the next few sections we therefore begin by briefly examining the nature of science, and how the concept of establishing scientific facts must bear upon our development.

1.2 The discourse of science

Science, like any other cultural product, requires an understanding of language, and the present development will require an especially clear understanding of a distinction that separates two different kinds of words, as follows:

Suppose we wish to compile a dictionary of the English language. It might seem at first that we must collect all the words in English, list them in lexicographical order, and then

adjoin to each word a definition of its meaning. On second thoughts, however, that cannot be, as the definitions would be circular; this word would be defined in terms of that word, and that word would be defined in terms of this word. Lexicographers are familiar with this problem. They deal with it by in effect drawing up, not one, but two lists of words; one list comprises definable words and the other list comprises ultimate words.

Ultimate words are not definable since they deal with the first-order experiences of life; the meanings of such words are demonstrable only. 'Red', for instance, is such a word; its meaning can be demonstrated to a normally sighted person by pointing out this, that and the other red object. However, a person who has always been blind is physically (bodily) incapable of grasping such a demonstration, where such physical incapacity cannot be circumvented by definitions; a person who has always been blind simply cannot grasp the physical (bodily) meaning of 'red'.

Having drawn up the two lists of words, the lexicographer must next consider how to explicate the ultimate words. As the dictionary can hardly provide its user with appropriate first-order experiences for the explication of ultimate words, it has to rely on experiences the user has already had. In the case of 'red', for instance, the standard solution is to have the dictionary declare 'red is the colour of blood', where that is not a definition, it is an evocation of a first-order experience of life. The dictionary relies on a childhood memory, in which a finger points and a voice says: 'This is blood. See, it is red'.

The word 'red' is an ultimate of physical science, as physical science is the discourse that concerns the world as experienced by the human body. When used in this sense, the term 'physical science' embraces basic sciences, such as physics, chemistry and biology, as well as applied sciences, such as agriculture, engineering and medicine. Many people would hold that the qualification 'physical' as used here is redundant, since they maintain that what we call 'physical science' is simply science. Others might disagree because they might want to distinguish physical science from, for instance, what they call 'normative science'. Such disagreements need not concern us here. We need not establish the valid usage of the word 'science'. We need only make it clear that unless explicitly stated otherwise, we are concerned with science in the sense of the discourse of physical experience (bodily experience), much of which concerns the development of two complementary, but fundamentally different questions formulated in Definitions 1.2.1 and 1.2.2.

Definition 1.2.1:

'How might these bodily experiences have come about?' is the definitive question of scientific investigation. In science it proclaims the discourse of the pursuit of knowledge.

Definition 1.2.2:

'How might such bodily experiences be brought about?' is the definitive question of scientific technology. In science it proclaims the discourse of the use of knowledge.

This chapter concerns the pursuit of knowledge, rather than the use of knowledge. The bodily experiences then *to be explained* are usually referred to as 'data'. Hence, the definitive question of investigative science can be put into the form *'How might these data*

be explained?' We note in passing that the discourse of scientific technology sometimes involves 'data' in the different sense of bodily experiences *to be responded to*.

1.3 Establishing scientific facts

Scientific facts are those that can compel agreement by appealing to the experiences of the human body. Oenology, for instance, uses a variety of special terms to identify certain tastes, odours and colours that might characterise a wine. Most people can learn to detect those characteristics. For example, wines made from Pinot Gris vines grown in the Western Cape of South Africa were found to occasionally have a paraffin-like taste that is undesirable and that a panel of tasters were trained to detect. These tasters were then used by way of 'blind' tasting to establish whether or not, and to what degree, certain experimental wines had the paraffin-like taste. This example shows how science establishes physical facts, that is to say, facts that the human body can be compelled to grasp, as when the oenologist, if challenged, can say 'Taste these for yourself'.

Again, recall Galileo's law on the acceleration of falling bodies. Consider dropping two iron balls – one large, one small. To the human mind it might seem 'logical' that the heavier ball would accelerate faster than the lighter one. So Galileo had to trick his opponents into watching him drop two such balls from the leaning tower of Pisa. He had to circumvent their 'logic' in order to compel their bodies to physically grasp the contrary.

Once again: consider the drafts required to draw ploughs at speeds commonly attained by tractors. To the human mind it might seem 'logical' that the regression of draft, Y, on speed, X, should include the origin, as there would seem to be no draft when the plough is stationary. However, a plot of recorded (X, Y) data pairs will compel the human body to grasp that 'inertia must be overcome' before the plough will move (Figure 1.3.1 overleaf).

The issue is crucial: the ultimate facts of science are those that can compel agreement by appeal to the human body as the ultimate arbitrator of science. Anyone who would try to make 'logic' circumvent such an appeal is either being obstinate or silly.

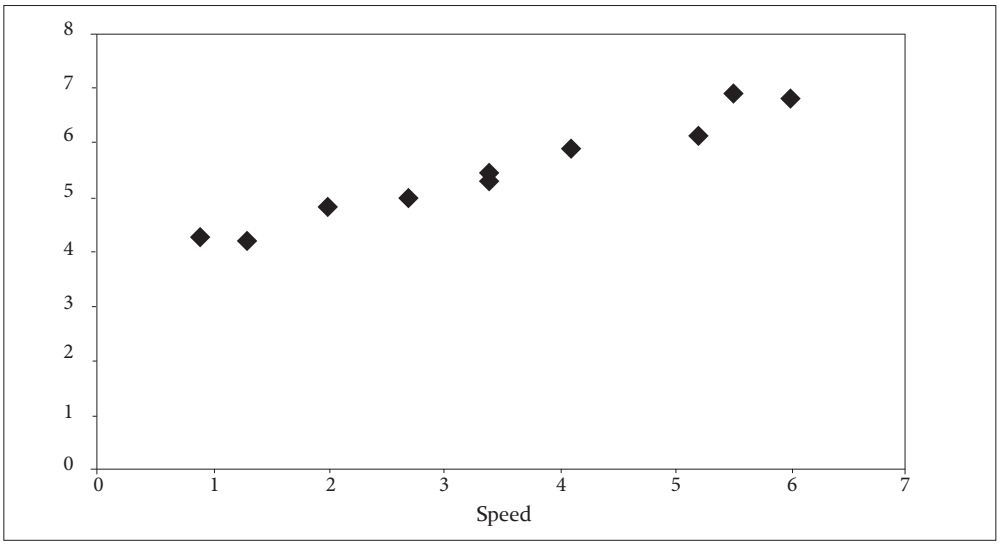


Figure 1.3.1: Draft (pounds times 0.01) and speed (miles per hour) of plows drawn by tractors (Source: Snedecor 1956, p. 142)

It should be obvious from the foregoing discussion that we find it convenient to use the expression 'physical experience' ambiguously; sometimes we refer to an actual experience, and sometimes we refer to a representation (a record) of that experience. This is not important as long as it is perceived and understood.

1.4 The ultimate words of statistics

A spoon sent spinning into the air can land with its bowl facing either up (u) or down (d). The following sequence of outcomes was obtained from just 35 spins of a spoon:

duddu uuudu uudud uduuu uuuud uuudd uuudu.

By plotting the relative frequencies of the two different outcomes against the number of spins as in Figure 1.4.1, we can compel the human body to grasp the concept called long-run frequency (theoretical frequency). It is an ultimate concept of science – of genetics, of statistical mechanics and of mathematical statistics. It is in fact one member of an inseparable pair of ultimate concepts, the other being the one called sampling, as physically demonstrable by spinning a spoon, rolling a die, flipping a coin, or shuffling cards.

We note in passing that it is not uncommon for ultimate concepts of science to occur in inseparable pairs. Euclid's geometry, for instance, is a theory of physical space where perpendicular and parallel amount to such a pair. One of the members of such a pair is often operational and the other one is perceptual. The simplest forms of these are found in looking to see, licking to taste, and listening to hear.

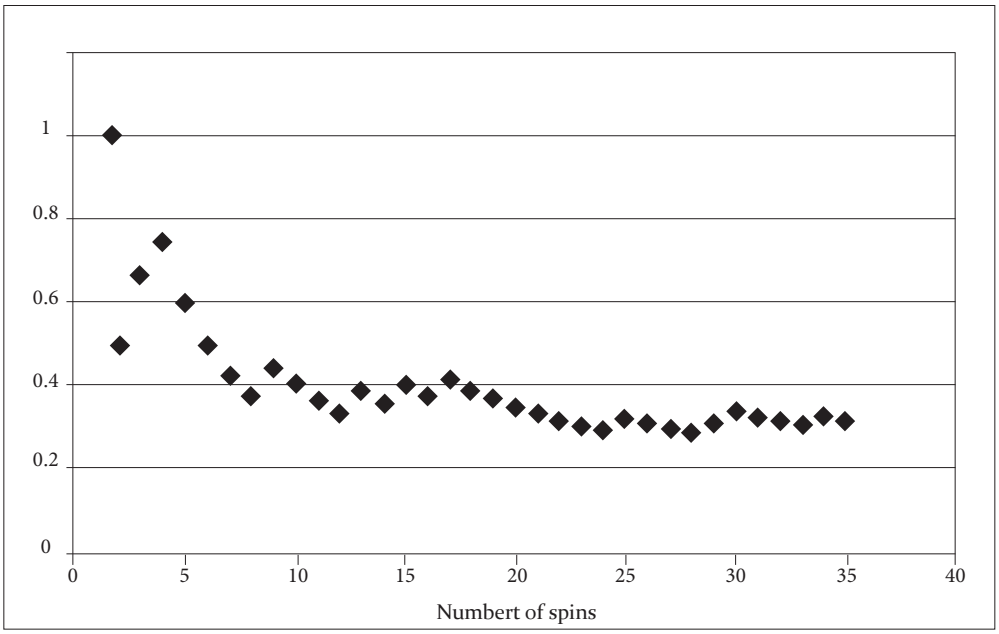


Figure 1.4.1: Frequency of outcome 'bowl down' when spinning a spoon

We also note that probability cannot be defined as long-run frequency; the latter is not definable, it is demonstrable only. Mathematical probability is used to describe the physical experiences we associate with long-run frequency, but can also be used to describe other experiences. Thus, for instance, of a cocktail made of equal proportions of vermouth, gin and lemon juice, it can be said, correctly, and in terms of standard notation:

$$\Pr(\text{gin}) = \tfrac{1}{3}\,, \text{ and } \Pr(\text{gin} \mid \text{alcoholic beverage}) = \tfrac{1}{2}\,.$$

Mathematical probability is really just the mathematics of proportional constituency. However, unless stated otherwise, we use the term 'probability' to mean 'theoretical frequency' only.

1.5 Mathematical forms and physical meanings

This book will ask statisticians to revise deeply entrenched ways of thinking. We urge the reader to constantly bear the following fact in mind:

> *The same* mathematical forms can be used to convey *different* physical meanings; *physical meanings* therefore cannot be derived from *mathematical forms* as such.

This fact is exemplified in Table 1.5.1 by a finite geometry developed by Miss Evelyn Rosenthal in a book for the parents of school children (Rosenthal, 1965, p. 204). In order to prove that the axioms of such a geometry *as such* is a consistent set, we must find at least one model that provides a *physical (bodily)* proof that they work, because, as explained by Miss Rosenthal, it is impossible to provide a mathematical proof of such

consistency. For the present geometry she develops, not one, but two, quite different physical proofs by way of the two different diagrams in Figures 1.5.1(a) and (b). She says, 'If you check the three axioms and the theorem in each diagram you will find that they work'. Hence, by pointing at just one of the two diagrams, she compels the human body to grasp that the formal mathematics of the present example can be made to convey a system of physical meanings. Next, by pointing at the other diagram, she compels the human body to grasp that the selfsame formal mathematics can be made to convey another, very different, system of physical meanings. She thus proves *inter alia* that the mathematical forms *per se* are devoid of those meanings.

Table 1.5.1: A logic to which different scientific meanings can be adjoined

Undefined terms: *rudd*; *vory*. 'A rudd *joins* two vories' means the same as 'Two vories *are on* a rudd'.
Axioms: (1) There are exactly four rudds on each vory. (2) There are exactly two vories on each rudd. (3) Every vory is joined to every other vory by exactly two rudds.
Among the theorems that can be deduced is:
Theorem: There are exactly six rudds and three vories.

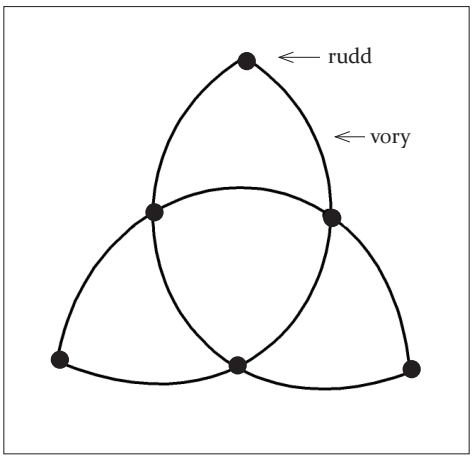


Figure 1.5.1(a): In this figure, rudds are points and vories are lines

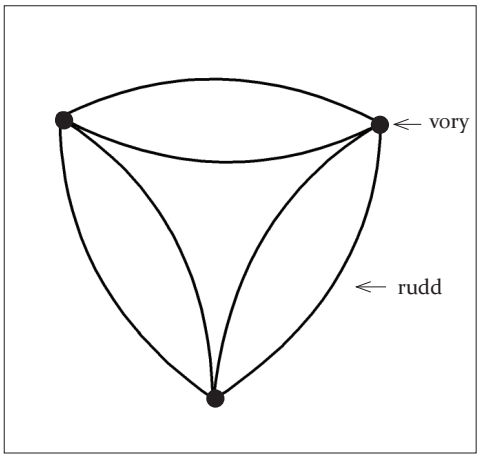


Figure 1.5.1(b): In this figure, vories are points and rudds are lines

The distinctions between significance tests, hypothesis tests and co-ordination tests are very much like the distinctions between Miss Rosenthal's finite geometries, in that the selfsame mathematical forms will ultimately turn out to be conveying very different physical meanings. Unfortunately, much of the present book has to be read before the different meanings will have been fully developed. Nevertheless, it will here serve our immediate purposes to take a first step in that direction by way of considering the following version of the so-called *mixed sampling problem*:

Suppose that an unknown value, μ, can be measured precisely by Instrument A and can be measured imprecisely by Instrument B, as follows: a measurement made by A can be represented as a realisation of X whose distribution is given by

$$\Pr(X = \mu) = 1.$$

A measurement made by B can be represented as a realisation of Y whose distribution is given by

$$\Pr(Y = \mu + \varepsilon) = \frac{1}{5} \text{ for } \varepsilon = -2, -1, 0, +1, +2.$$

If we flip a balanced coin to pick an instrument to make a measurement, the result can be represented as a realisation of Z whose distribution is given by

$$\Pr(Z = \mu+\delta) = \frac{6}{10} \text{ for } \delta = 0, \text{ and } \Pr(Z = \mu+\delta) = \frac{1}{10} \text{ for } \delta = -2, -1, +1, +2.$$

The precision of the various measurements is described by

$$\text{Variance } (X) = 0. \text{ Variance } (Y) = 2. \text{ Variance } (Z) = 1.$$

We now ask the following questions:

> If a measurement from A was obtained by flipping the coin, must it be represented as a sample from the population whose variance equals 0, or must it be represented as a sample from the population whose variance equals 1? (1.5.1)

> If a measurement from B was obtained by flipping the coin, must it be represented as a sample from the population whose variance equals 2, or must it be represented as a sample from the population whose variance equals 1? (1.5.2)

We will, by way of developments in subsequent chapters, prove beyond reasonable contest that these two questions, as put forward here, cannot be answered. We will achieve that by showing that in a certain substantive context, the correct answers are variance = 0 and variance = 2, respectively, and in another substantive context the correct answers are variance = 1 and variance = 1, respectively. Hence, just as Miss Rosenthal's mathematical forms are scientifically vacuous when considered without substantive context to show how they are intended to address those bodily experiences (geometrical experiences) referred to in terms of the concept 'the space we live in'. So also, a formal presentation of the mixed sampling problem is scientifically vacuous when considered without substantive context to show how it is intended to address those bodily experiences (statistical experiences) referred to in terms of the concepts 'sampling' and 'long-run frequency'.

In view of the foregoing, the reader must be very careful to avoid reading into our mathematical formalities physical meanings that are not explicitly indicated, and must firmly grasp the physical meanings that will be explicitly indicated. In order to help the reader in this matter, we will from time to time underscore what is meant and what is not meant.

1.6 A few basic statistical ideas

A *sample space* is a set of mutually exclusive and exhaustive descriptions in terms of which we choose to describe the outcome of a conceptual trail. Consider the outcome of twice spinning a spoon. Using the notation of Section 1.4, and the ordering

(outcome of 1st spin, outcome of 2nd spin),

let us choose the sample space to be {(u, u), (u, d), (d, u), (d, d)}. Let the pair of outcomes be modelled as statistically independent, and let μ denote the probability of 'bowl up' ($0 < \mu < 1$). Then we obtain *a class of models* whose members are indexed by different values of μ (Table 1.6.1).

Table 1.6.1: A class of models whose members are indexed by μ for $0 < \mu < 1$

(x_1, x_2)	(u, u)	(u, d)	(d, u)	(d, d)
$\Pr[(X_1, X_2) = (x_1, x_2)]$	μ^2	$\mu(1-\mu)$	$(1-\mu)\mu$	$(1-\mu)^2$

Consider any member of the class of models given in Table 1.6.1, for instance the member indexed by $\mu = 0.3$. Then we obtain *a fully specified model* (Table 1.6.2). We will refer to a fully specified model as *a singleton*.

Table 1.6.2: A fully specified model (i.e. a singleton)

(x_1, x_2)	(u, u)	(u, d)	(d, u)	(d, d)
$\Pr[(X_1, X_2) = (x_1, x_2)]$	0.09	0.21	0.21	0.49

There are infinitely many ways in which any given singleton can be imbedded into a class of models. In Table 1.6.3, for instance, the usual singleton for the outcome of rolling an ordinary die has been imbedded into a class of models. For reasons to be explained in this chapter we often deliberately avoid such imbedding, in which case we will call the singleton involved *an isolated singleton*.

Table 1.6.3: A class of models for the outcome of one roll of a die ($-1/5 < \theta < +1/5$)

x	1	2	3	4	5	6
Pr(X = x)	$(1-5\theta)/6$	$(1-3\theta)/6$	$(1-\theta)/6$	$(1+\theta)/6$	$(1+3\theta)/6$	$(1+5\theta)/6$

The number of times 'bowl up' arises when twice spinning a spoon, say X, is an *observable random variable* (X = 0, 1, 2). Here the term 'observable' distinguishes unobservable variables such as X-2μ for unobservable μ ($0 < \mu < 1$), from observable variables such as X-2μ_0 for specified μ_0. The terms 'observable' and 'unobservable' often refer to 'calculable' and 'not calculable', respectively. An observable random variable is called *a statistic*. A statistic arises from a partitioning of a sample space into mutually exclusive and exhaustive subsets that are differentially and observably labelled. In the present case, for instance, the subsets are

{(u, u)}, labelled '2', {(u, d), (d, u)} labelled '1', and {(d, d)} labelled '0'.

Often, as in the present case, the labels describe how the subsets are formed.

There is a primitive statistical idea that if a given event is rare under presumed circumstances, its occurrence can be held to be indicative of circumstances other than those presumed. However, the development of the idea needs careful consideration. For instance, with just 2n flips of a balanced coin, the probability of equal numbers of outcomes being 'heads' and 'tails', equals 0.5 when n = 1, equals 0.375 when n = 2, equals 0.3125 when n = 3, and so on, eventually becoming exceedingly small. So the idea as it stands would have us consider, nonsensically, that 500 outcomes 'heads' and 500 outcomes 'tails' in just 1 000 flips of a seemingly balanced coin, is indicative of the coin being unbalanced. In fact, it is not the absolute frequency of a given event, but its comparative frequency under alternative circumstances that can lend force to the idea, as will appear in the sequel.

1.7 Measuring the quality of fit of an isolated singleton

We are now ready to provide a heuristic introduction to co-ordination tests. We do so by way of two very simple examples, and we note from the outset that although the examples are simple, they represent problems that are of actual investigative interest.

Example 1.7.1

If each of seven beetles can be expected to settle into one of eight compartments, the animals can occupy the compartments in 8^7 different ways. For instance, the eight compartments might be occupied by 2, 2, 1, 1, 1, 0, 0, and 0 beetles, in some order. Such a pattern is denoted by $2^{[2]}1^{[3]}0^{[3]}$ when $b^{[c]}$ denotes that there are just $\underline{b}$ beetles in each of just $\underline{c}$ different compartments. The 8^7 ways can be sorted into just 15 different occupancy patterns, as shown in Table 1.7.1 overleaf. From this table we learn for instance that the pattern $2^{[2]}1^{[3]}0^{[3]}$ accounts for 705 600 of the 8^7 cases.

Table 1.7.1: Possible occupancy patterns and the corresponding numbers of cases

Pattern	#(cases)	Pattern	#(cases)	Pattern	#(cases)
$7^{[1]}0^{[7]}$	8	$4^{[1]}2^{[1]}1^{[1]}0^{[5]}$	35 280	$3^{[1]}1^{[4]}0^{[3]}$	235 200
$6^{[1]}1^{[1]}0^{[6]}$	392	$4^{[1]}1^{[3]}0^{[6]}$	58 800	$2^{[3]}1^{[1]}0^{[4]}$	176 400
$5^{[1]}2^{[1]}0^{[6]}$	1 176	$3^{[1]}2^{[2]}0^{[5]}$	35 280	$2^{[2]}1^{[3]}0^{[3]}$	705 600
$5^{[1]}1^{[2]}0^{[5]}$	7 056	$3^{[2]}1^{[1]}0^{[5]}$	23 520	$2^{[1]}1^{[5]}0^{[2]}$	423 360
$4^{[1]}3^{[1]}0^{[6]}$	1 960	$3^{[1]}2^{[1]}1^{[2]}0^{[4]}$	352 800	$1^{[7]}0^{[1]}$	40 320

In order to obtain such counts in general, we note that there are A^B different ways in which B beetles can occupy A compartments, and if a{r} then denotes the number of compartments occupied by just r animals, the number of cases accounted for by the occupancy pattern

$$0^{[a\{0\}]}1^{[a\{1\}]}2^{[a\{2\}]}\ldots$$

can be expressed as

$$\frac{A!B!}{a\{0\}!a\{1\}!a\{2\}!\cdots \times 0!^{a\{0\}}1!^{a\{1\}}2!^{a\{2\}}\ldots} \qquad (1.7.1)$$

wherein of course $0!^{a\{0\}}1!^{a\{1\}} = 1$ (Feller, 1970, Section II 5).

Now suppose that the $3^{[1]}1^{[4]}0^{[3]}$ data pattern arises in an actual trial, and that the investigator wishes to test whether a model of random occupancy could account for how the given data came about. Then we need a scale of 'resemblance' to random occupancy, such that the resemblance for data whose pattern frequently occurs with random occupancy will be greater than it is for data whose pattern rarely occurs with random occupancy. Table 1.7.1 shows for instance that $2^{[2]}1^{[3]}0^{[3]}$ describes 100 times more model cases than does $5^{[1]}1^{[2]}0^{[5]}$. So, the resemblance to random occupancy for data with the former pattern would seem to be greater than it is for data with the latter pattern. This principle produces an ordering of model cases ranging from 'most like random occupancy', O_1, to 'least like random occupancy', O_{14}, as in Table 1.7.2. Note that O_8 is a union of two equally frequent patterns. Note also that all of the patterns that have constituents of the form $b^{[c]}$ for $b \geq 5$ are gathered into O_T-like categories such that $T \geq 10$, that is to say, into categories which, according to the present ordering, are relatively unlike random occupancy. So, inasmuch as $b^{[c]}$ for $b \geq 5$ denotes cases where unusually many animals are found in the same compartment, the adopted ordering tests our model of 'random occupancy' against alternatives that can be described as 'aggregative occupancy'.

Table 1.7.2: A partial ordering of possible sample patterns and the respective modelled frequencies of the resulting ordinal classes

Order	Modelled frequency	Order	Modelled frequency
$O_1 = 2^{[2]}1^{[3]}0^{[3]}$	$705\,600/8^{-7}$	$O_8 = 4^{[1]}2^{[1]}1^{[1]}0^{[5]} \cup 3^{[1]}2^{[2]}0^{[5]}$	$2(35\,280)/8^{-7}$
$O_2 = 2^{[1]}1^{[5]}0^{[2]}$	$423\,360/8^{-7}$	$O_9 = 3^{[2]}1^{[1]}0^{[5]}$	$23\,520/8^{-7}$
$O_3 = 3^{[1]}2^{[1]}1^{[2]}0^{[4]}$	$352\,800/8^{-7}$	$O_{10} = 5^{[1]}1^{[2]}0^{[5]}$	$7\,056/8^{-7}$
$O_4 = 3^{[1]}1^{[4]}0^{[3]}$	$235\,200/8^{-7}$	$O_{11} = 4^{[1]}3^{[1]}0^{[6]}$	$1\,960/8^{-7}$
$O_5 = 2^{[3]}1^{[1]}0^{[4]}$	$176\,400/8^{-7}$	$O_{12} = 5^{[1]}2^{[1]}0^{[6]}$	$1\,176/8^{-7}$
$O_6 = 4^{[1]}1^{[3]}0^{[4]}$	$58\,800/8^{-7}$	$O_{13} = 6^{[1]}1^{[1]}0^{[6]}$	$392/8^{-7}$
$O_7 = 1^{[7]}0^{[1]}$	$40\,320/8^{-7}$	$O_{14} = 7^{[1]}0^{[7]}$	$8/8^{-7}$

The resulting test is displayed by the bar diagram in Figure 1.7.1, where the areas of the bars differ in proportion to the different frequencies of the patterns they represent. The given datum is described by $3^{[1]}1^{[4]}0^{[3]}$. In terms of the given ordering, the shaded bar represents model cases whose resemblance to random occupancy equals that of the given datum. The bars to the left of the shaded bar represent model cases whose resemblance to random occupancy is greater than that of the given datum. The bars to the right of the shaded bar represent model cases whose resemblance to random occupancy is lesser than that of the given datum.

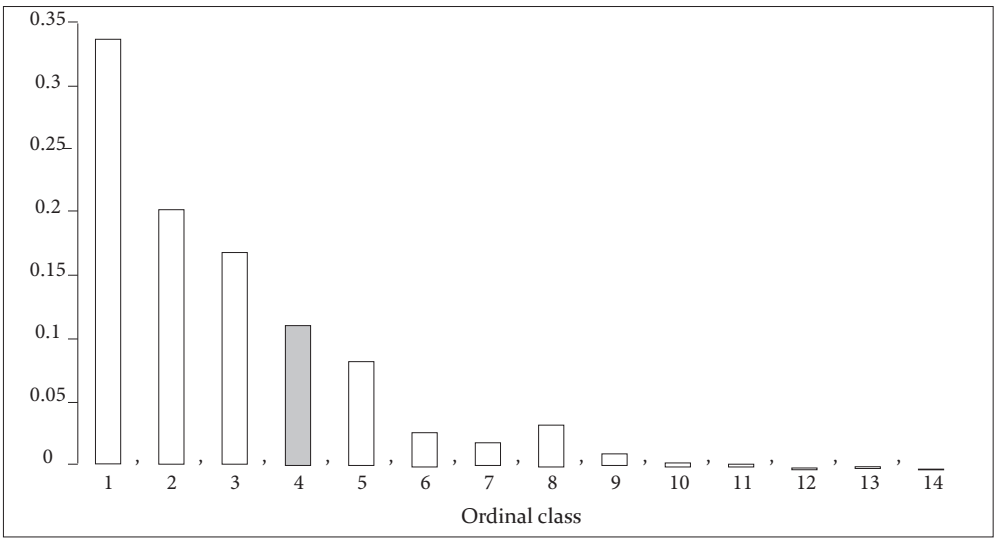


Figure 1.7.1: Testing the quality-of-fit of a model of random occupancy

We digress briefly in order to introduce a general terminology. We will use the notation O_T for T = 1, 2, 3, ..., to denote an ordered array obtained by partitioning the possible sample patterns arising from a given singleton into mutually exclusive and exhaustive subsets and then arranging those subsets in a specific order. We call the resulting array *an ordering*

(or *a partial ordering*) *of sample patterns*, where the term 'partial' (when used) serves to remind us that some of the subsets may comprise more than just one pattern, as exemplified by O_8 in Table 1.7.2. We call the resulting statistic *a test statistic*, and we call its distribution *the test distribution.* Consider Figure 1.7.1. The bars to the left of the shaded bar account for 0.7 of the model cases; we call those cases *the left statistical co-ordinate of the modelled datum*. The bars to the right of the shaded bar account for 0.2 of the model cases; we call those cases *the right statistical co-ordinate of the modelled datum*. The shaded bar accounts for 0.1 of the model cases; we call those cases *the statistical rounding*. The given datum is modelled as a member of the rounding. Thus the modelled datum is *the mental correlate* of the given datum. We will often draw no distinction between a statistical co-ordinate and its measure. Thus, in the present case, we might refer to 0.7 and 0.2 as the left and right statistical co-ordinates of the modelled datum, respectively. Similarly we might refer to 0.1 as the rounding within which the datum is being modelled. When we report the co-ordinates of a modelled datum as a number pair, for instance (0.7, 0.2) as in the present case, the member on the left will be the left co-ordinate.

Now consider what can be learned from Figure 1.7.1 noting that, if needs be, the theoretical frequencies displayed in Figure 1.7.1 could be replaced by simulated frequencies. So the display clearly belongs to the discourse of physical evidence, in which we point at Figure 1.7.1, or at a simulated equivalent, and say 'See for yourself how the members of the rounding, including the modelled datum, are situated snugly within the crowd'. The human body is thus compelled to grasp, as a physically demonstrable fact, that the statistical model under test, by the test performed, fits the given data well. Once this is understood, we can of course dispense with Figure 1.7.1 and instead report simply that for the model under test, and for the test performed, the co-ordinates of the modelled datum are given by (0.7, 0.2).

The reader should carefully note that we have reasoned in terms of a single instance of given occupancies in the real world and an infinite population of random occupancies in the human mind. Our reasoning did not envisage or depend upon the existence, or the future existence, of any population of occupancies in the real world. True, our reasoning was intended to inform opinion about occupancy behaviour in a certain sort of beetle, such as for instance the male beetle of species X. That, however, did not require the existence or the future existence of a population of occupancies in the real world. At the risk of belabouring this point we note that had the seven males in the beetle trial been the last survivors of their species, that would have had no effect on the validity of our reasoning, or upon its ability to inform entomological opinion.

The reader should also take care to note that the result of our co-ordination test is not a result of which the veracity is qualified by probability. The result of our test is a physically perceived fact, which, as such, is forced upon the human body and is thus beyond reasonable contest. We have used the method once used at Pisa by Galileo.

Example 1.7.2

We often require a given string of consecutive non-negative integers to be partitioned into two or more groups using a pseudo-random device. For example, let 1, 2, 4, 5 and 3, 6, 7, 8 be modelled as a random partition of the string 1, 2, 3, ..., 8. How could we test

the quality of fit of the model? The model involves 8-*choose*-4 different sample patterns, that is to say, involves in terms of standard notation

$$\binom{8}{4} = \frac{8!}{4!(8-4)!} = 70 \text{ patterns}$$

These patterns are modelled as all having precisely the same frequency of occurrence. So a test of the quality of fit of the model cannot be based directly on the principle of ordering by size of modelled frequency. We are therefore compelled to replace the 70 patterns by a smaller number of patterns, such that the latter patterns vary in modelled frequency. One way of doing this is based on *runs,* as follows: replace every number in the original string by A if it is in the first group and by B if it is in the second group. The original string is thus replaced by the string AABAABBB, consisting of the four runs AA, B, AA and BBB. The runs are then replaced by the run lengths 2, 1, 2, 3, where the original two groups are no longer distinguished as they complement each other. Depending on the nature of the device used to partition the original string, we might then choose to further reduce the number of patterns by ignoring the order in which the runs occur. We thus obtain a datum described by $1^{[1]}2^{[2]}3^{[1]}$. Now, ordering by size of modelled frequency, we obtain the test distribution given in Table 1.7.3. Note that patterns involving very long runs or many short runs are taken up by O_5 and O_6 where that indicates the nature of the alternatives we have in mind when we choose to order sample patterns by the present method.

Table 1.7.3: A test distribution based on runs

Partial ordering	Modelled frequency
O_1: $1^{[4]}2^{[1]}$	18/70 = 0.26
O_2: $1^{[3]}2^{[1]}3^{[1]}$	12/70 = 0.17
O_3: $1^{[2]}3^{[2]} \cup 1^{[1]}2^{[2]}3^{[1]}$	2(8)/70 = 0.23
O_4: $1^{[2]}2^{[3]} \cup 1^{[6]}2^{[1]}$	2(6)/70 = 0.17
O_5: $1^{[1]}3^{[1]}4^{[1]}$	4/70 = 0.06
O_6: $2^{[2]}4^{[1]} \cup 1^{[8]} \cup 2^{[4]} \cup 4^{[2]}$	4(2)/70 = 0.11

Using Table 1.7.3 the co-ordinates of the samples in O_6 are found to be given by (0.89, Ø) where Ø denotes zero arising from the absence of a right co-ordinate. Note that (0.89, Ø) involves a rounding of measure 0.11, where that is large enough to discourage (0.89, Ø) from being considered descriptive of a poor fit, as the mental correlate of a given datum can be situated *anywhere* in the rounding it belongs to. This is underscored when the present co-ordinates are reported in the explicit form (0.89, 0.11, Ø) rather than in the implicit form (0.89, Ø). These co-ordinates reflect a paucity of data, indicating that any test based on Table 1.7.3 would be nearly vacuous.

The reader should note that our usage of the expression *the given datum* in the foregoing did not refer to the original data set. It referred to a summary datum of which it may be said that it is being *given to be tested*. If preferred, one might call it *the test datum*.

1.8 Measuring the quality of fit of a class characteristic

In this section we consider the case of testing the quality of fit of an isolated singleton that has arisen as characteristic of each and every member of a class of models. As in the previous section we again employ concrete examples to develop the general idea.

Example 1.8.1

Suppose that an investigator counts the numbers of a certain plant species in each of six quadrates and finds 0, 1 and 2 of the plants in 3, 1 and 2 of the quadrates, respectively. The investigator's experience might suggest that the given data might be modelled successfully as six independent counts from a Poisson population with unspecified mean denoted by μ ($0 < \mu < \infty$). This introduces a class of models whose composition was analysed by Fisher (1950). Let B denote the total number of plants, A denote the total number of quadrates, and a{r} denote the number of quadrates with just r plants. Fisher points out that for the proposed class of Poisson models, the probability that a random sample of A counts will have the pattern

$$0^{[a\{0\}]}1^{[a\{1\}]}2^{[a\{2\}]}\ldots$$

can be expressed as a product of two factors, as follows in square brackets:

$$\left[\frac{e^{-A\mu}(A\mu)^{B}}{B!}\right]\times\left[\frac{A!B!\times A^{-B}}{a\{0\}!a\{1\}!a\{2\}!\ldots\times(2!)^{a\{2\}}(3!)^{a\{3\}}(4!)^{a\{4\}}\ldots}\right] \qquad (1.8.1)$$

The first factor at (1.8.1) gives the probability that the sample total equals B, given that the sample comprises A independent counts from a Poisson population with mean equal to μ; for each value of μ it provides a singleton whose quality of fit depends on the total count only. Then, given that the total count equals B, the second factor at (1.8.1) gives the conditional probability of the particular pattern occurring in A individual Poisson counts; as it is independent of μ, it is characteristic only of the Poisson-ness, so to speak, of the class of models. The second factor will be recognised as the singleton that originated at (1.7.1) in the previous section. We must commence by testing the quality of fit of the second factor, as the first factor relies on the sample of counts being Poisson without providing any means whatsoever for judging whether or not that is appropriate. So, using the second factor, we compute the conditional probability of each sample pattern obtainable with A = 6 and B = 5. Table 1.8.1 gives all these possible sample patterns, their respective conditional probabilities, and the ordering that arises from the principle 'a pattern less frequent with random occupancy is a pattern less like those of random occupancy'. Just as in Section 1.7, the O_T-like notations again indicate that 'likeness' to random occupancy decreases as T increases. By inspection of the patterns in O_T for T = 7, 6, 5, ..., it can be seen that the ordering points at aggregative occupancy as an alternative against which our model of random occupancy is being tested.

Table 1.8.1: A test distribution for a Poisson class characteristic

Pattern	Probability	Order	Pattern	Probability	Order
$5^{[1]}0^{[5]}$	$1/6^4$	O_7	$2^{[2]}1^{[1]}0^{[3]}$	$300/6^4$	O_2
$4^{[1]}1^{[1]}0^{[4]}$	$25/6^4$	O_6	$2^{[1]}1^{[3]}0^{[2]}$	$600/6^4$	O_1
$3^{[1]}2^{[1]}0^{[4]}$	$50/6^4$	O_5	$1^{[5]}0^{[1]}$	$120/6^4$	O_4
$3^{[1]}1^{[2]}0^{[3]}$	$200/6^4$	O_3			

As the test datum in the present case is described by $2^{[2]}1^{[1]}0^{[3]}$, the test statistic developed in Table 1.8.1 produces the test displayed in Figure 1.8.1. The shaded bar represents the statistical rounding whose co-ordinates are given by (0.46, 0.31). So, by the test performed, the class characteristic matches the given data well. Should there be any doubt as to what this means, we can point at Figure 1.8.1, or a simulated equivalent, and say, 'See for yourself that the members of the rounding, including the mental correlate of the test datum, are situated well within the bulk of the distribution'. We would thus compel the human body to grasp, as a physically demonstrated fact, that by the test performed, the class characteristic fits the given data well.

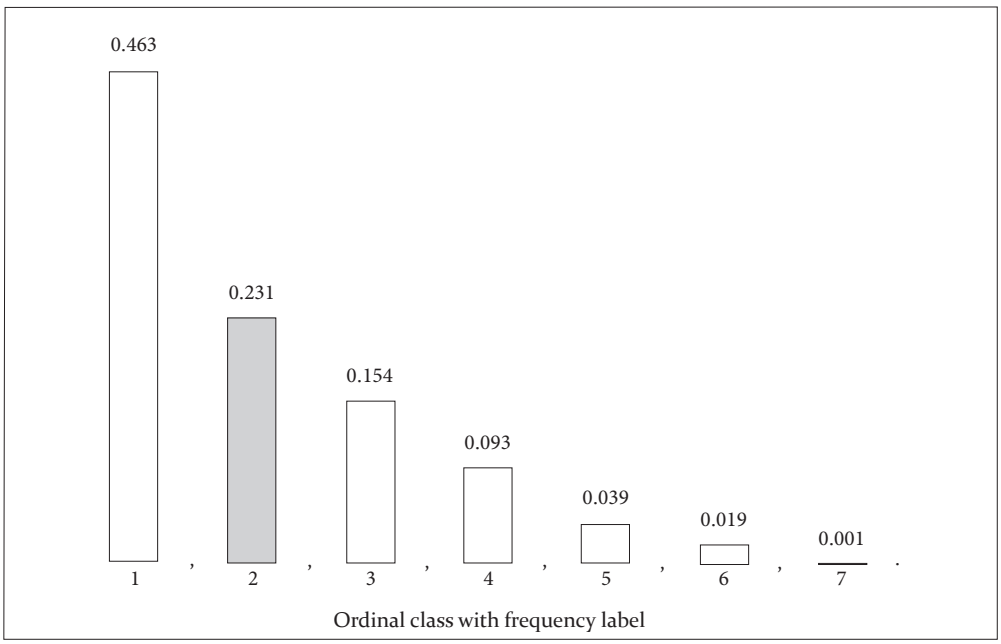


Figure 1.8.1: Testing the quality of fit of a Poisson class characteristic

As in the previous section, so also in the present section, our reasoning neither made reference to, nor relied in any way on the existence, or on the possible future existence, of a population of occupancies in the real world. Our reasoning concerned only a single instance of given occupancies in the real world, corresponding to which it brought into

the human mind an infinite population of random occupancies as a model of how the given occupancies might have come about.

It would be foolish to pretend that our model has provided the only possible explanation of how the given occupancies might have come about, but it would also be foolish to pretend that our model arose from a blind guess. The point here is simply this: informed by facts about the mode of propagation in the particular species of plants involved, and informed by facts about the nature of the particular terrain involved, botanical opinion has chosen the hypothesised class of models, either to be refuted, or to be supported. We, in turn, must then produce appropriate statistical facts of fit (good or bad) that can serve to better inform that botanical opinion.

As in the previous section, our co-ordination test produces a finding of which the veracity is not qualified by probability. The finding is a physically perceived fact of which the veracity is absolute, having been placed beyond reasonable contest. This point must be firmly grasped. So let us state precisely what our finding is, as follows:

A model of random occupancy, by the test performed, fits these data well.

Should we be forced to attach a 'probability of truth' to this finding, we would perforce have to declare that it be unity, as the finding is plainly a fact. We must, however, be extremely reluctant to introduce such a 'probability', as it can serve no positive purpose for an irrelevant concept to be dragged into our development.

Example 1.8.2

Suppose that on a visit to a town named T we spot municipal buses numbered $T.x_i$ for $i = 1, 2, 3, \ldots, n$. A tenable model for these data might be that they are a subset of T.x for $x = 1, 2, 3, \ldots, \theta$. In order to estimate θ, the number of municipal buses in T, we require a probability model for our data. So let $x_{(i)}$ for $i = 1, 2, 3, \ldots, n$, denote the observed numbers ordered from smallest to largest, and let $X_{(i)}$ for $i = 1, 2, 3, \ldots, n$, denote corresponding random variables in the human mind. Let the data be modelled as a random sample drawn without replacement. Then the probability of each of the possible sample patterns is taken to be

$$\binom{\theta}{n}^{-1}.$$

This probability can be expressed as

$$\left[\binom{x_{(n)}-1}{n-1}\binom{\theta}{n}^{-1}\right] \times \left[\binom{x_{(n)}-1}{n-1}^{-1}\right]. \qquad (1.8.2)$$

Here the first factor in square brackets is the probability of obtaining $X_{(n)} = x_{(n)}$ where $x_{(n)} = n, n+1, n+2, \ldots, \theta$; the second factor is the conditional probability of the sample given that $X_{(n)} = x_{(n)}$. The first factor tells us nothing at all about how the sample arose; it represents a class of models indexed by θ, which class is based on the premise that our data can be modelled as having been drawn at random without replacement. The second factor is the class characteristic. Given that $X_{(n)} = x_{(n)}$ it tells us that the pattern of the remaining n-1 sampled numbers is one amongst $(x_{(n)}-1)$-choose-(n-1) different

patterns that are equally frequent when sampling is random without replacement. Let the data be 1, 2, 3, 4, 9 and 10. The largest of these numbers, $x_{(n)}$, equals 10, and the class characteristic models the five smaller numbers, 1, 2, 3, 4, and 9, as arising from a random partition of the first nine positive integers into two sets comprising five integers drawn, and four integers not drawn, respectively. So the class characteristic is a singleton of the type considered in Example 1.7.2. Just as in Example 1.7.2, we must replace a variety of equally frequent sample patterns with a smaller variety of sample patterns that vary in modelled frequency. Consider the use of runs as described in the previous section: when labelling the observed numbers 'A' and the unobserved numbers 'B', the given data set, apart from $x_{(n)} = 10$, yields the following:

1, 2, 3, 4, 5, 6, 7, 8, 9
A A A A B B B B A

The runs are AAAA, BBBB and A. If only the lengths of the runs are recorded, the test datum is $1^{[1]}4^{[2]}$. The corresponding test distribution is given in Table 1.8.2.

Table 1.8.2: A test distribution for the number-of-buses problem

Partial ordering	Modelled frequency
O_1: $1^{[3]}2^{[3]} \cup 1^{[4]}2^{[1]}3^{[1]}$	2(18)/126 = 0.29
O_2: $1^{[5]}2^{[2]} \cup 1^{[2]}2^{[2]}3^{[1]}$	2(15)/126 = 0.24
O_3: $1^{[7]}2^{[1]} \cup 1^{[1]}2^{[1]}3^{[2]} \cup 1^{[2]}3^{[1]}4^{[1]}$	3(8)/126 = 0.19
O_4: $1^{[3]}2^{[1]}4^{[1]} \cup 1^{[3]}3^{[2]}$	2(6)/126 = 0.09
O_5: $1^{[1]}2^{[2]}4^{[1]} \cup 2^{[3]}3^{[1]}$	2(4)/126 = 0.06
O_6: $1^{[1]}2^{[4]} \cup 1^{[6]}3^{[1]}$	2(3)/126 = 0.05
O_7: $1^{[1]}3^{[1]}5^{[1]} \cup 1^{[1]}4^{[2]} \cup 2^{[1]}3^{[1]}4^{[1]} \cup 4^{[1]}5^{[1]}$	4(2)/126 = 0.06
O_8: $1^{[9]} \cup 2^{[2]}5^{[1]}$	2(1)/126 = 0.02

As depicted in Figure 1.8.2, the modelled counterpart of the given datum is situated at (0.92, 0.02) in the test distribution. The characteristic, as tested, does not fit the given data well. Just as we could previously point at the scatter diagram in Figure 1.3.1 and say, 'See for yourself how poorly a straight line through the origin would fit these data', so we can now point at Figure 1.8.2 and say, 'See for yourself how awkwardly the counterpart is placed within the distribution. See how far down it is situated amongst the patterns least typical of runs arising from a random partition.' Note, however, that this test would be utterly vacuous if we cannot produce a more tenable alternative to our hypothesised model, because we cannot doubt the possible occurrence of a data pattern that we ourselves have modelled as being possible with non-zero probability. But perhaps we might recall that buses numbers 9 and 10 were spotted on the outskirts of the town as we were advancing toward the middle of town, where we then spotted the three smaller numbers. It might then occur to us that it could be the *routes* of the buses, rather than the buses themselves, that are numbered. Any new route would then tend to arise on the

outskirts of town and would have a larger number than previously established routes. We thus have a tenable alternative to the hypothesised model.

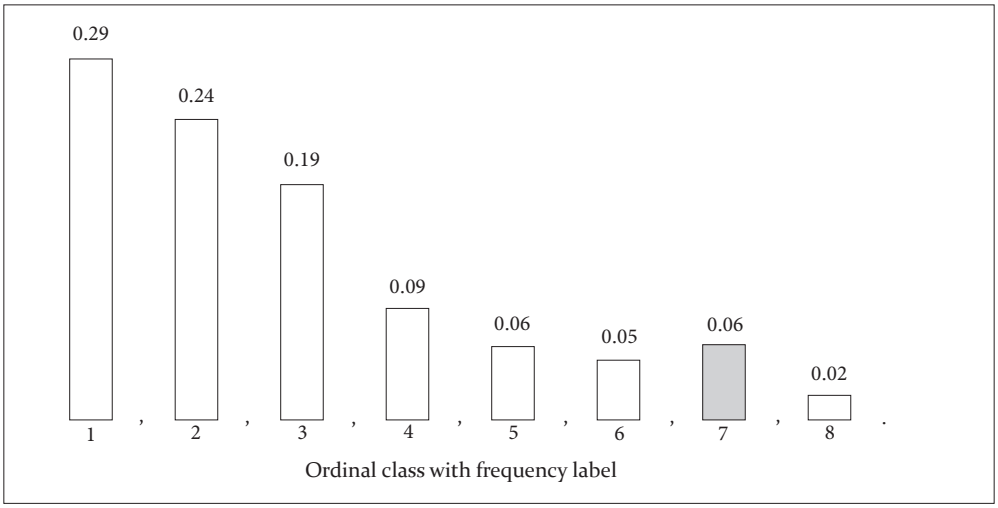


Figure 1.8.2: Testing the quality of fit of the number of buses class characteristic

Example 1.7.2 showed that large rounding discourages extreme co-ordination. The rounding in the present case is much smaller than the rounding in Example 1.7.2, but is nevertheless too large to be fobbed off. In such cases the rounding should be made explicit. So, in the present case, the co-ordination should be reported in the explicit form (0.92, 0.06, 0.02) rather than in the implicit form (0.92, 0.02), as good scientific practice always draws attention to any shortcoming of reported evidence. Henceforth, whenever we use the implicit form it must be tacitly understood that the rounding is much too small, or that both co-ordinates are much too large for the magnitude of the rounding to have forceful bearing on the physical evidence being reported. Sometimes, however, in spite of small rounding or large co-ordinates, we use the explicit form as a reminder that a rounding cannot be zero, as that would in self-contradictory terms try to model a given datum as one that could not have occurred.

1.9 A word of caution

The heuristic method of 'ordering by modelled frequency' cannot be relied upon to provide tests that are 'good' or 'best' in a defendable sense. We will in fact show that the method has an understandable tendency to produce inferior tests of co-ordination. However, for the time being that need not concern us, as we must first grasp in what sense our models are capable of being 'tested', before considering how certain tests might achieve that 'better' than others do. For our immediate purposes, it need only be grasped that different orderings produce different tests.

1.10 The terms 'sampling' and 'sample'

In the foregoing explanations we have been very careful to use language that draws a sharp distinction between the constituents of a particular data set in the real world and the constituents of a corresponding model in the human mind. The distinction we wish to draw is brought forward when we compare the traffic circles in the real world to the mathematical circle in the human mind. The traffic circles differ from the circle in the mind; yet everyday language understands in what sense they are 'like' the circle in the mind, and everyday language is satisfied to ignore their diversity. The very essence of statistics, however, is to not ignore such diversity, but instead to invent a sample space that models that diversity and, going further, to invent a distribution that models its proportional constituency. Nevertheless, the result is still a model only and, as we saw in Section 1.4, statistical modelling commences with a concept called sampling. So we will continue to draw a sharp distinction between the world of real physical experience and the world of conceptual physical experience by not referring to any object in the real world as 'sampling' or 'a sample'. This results in slightly awkward language, but there are at least three good reasons for maintaining such language, these being as follows:

Firstly, given a model that puts forward an explanation of how a given data set might (or might not) have come about, we will find (for instance in Chapter 4) that the language helps us avoid circular reasoning when judging whether (or not) the model is tenable. The point here is that we must carefully distinguish between a data analyst who *asks* whether or not a particular representation is *tenable* (the pursuit of knowledge), and a decision-maker who *assumes* that a particular representation *is* tenable (the use of knowledge).

Secondly, the language will help us come to grips (in Chapter 3) with certain slippery distinctions between hypothesis tests and tests of the kind currently being introduced. The reason for this is that in the current case the population is being brought into the human mind, whereas in the case of hypothesis tests the population is being brought into the real world.

Thirdly, long-run frequency is a conceptual consequence of sampling. So it is indeed a poor epistemology that would have us judge the quality of fit of a given long-run frequency model without also having us judge the quality of fit of the corresponding sampling model, in other words, without having us judge whether for instance a coin is being dropped instead of flipped.

1.11 Alternatives to a hypothesised model

Any test of a hypothesised model for its tenability, as explanation of how given data might have come about, must necessarily involve the idea that 'there could be another explanation'. If no alternative explanation is to be considered, it is utterly impossible to make a non-vacuous ordering of sample patterns. Examples 1.11.1 and 1.11.2 will help to make this clear.

Example 1.11.1

An ornithologist counts the number of non-breeding cape sugarbirds visiting each of 24 different protea bushes and finds 0, 1, 2, 3, 4 and 5 birds at 10, 4, 5, 3, 1 and 1 of the 24 sites, respectively. The ornithologist wants to test the quality of fit of a Poisson class. The present data set is too large to be dealt with readily by the method used in Section 1.7, and in any case our immediate purposes will be better served by a partial ordering of sample patterns according to the magnitude of

(the sample variance)/(the sample mean). (1.11.1)

Under the model to be tested this quantity is approximately distributed as

(χ^2 on 24-1 degrees of freedom)/(24-1).

For the given data

(the data variance)/(the data mean) = 1.541

of which the modelled counterpart in the human mind is then found to be co-ordinated at approximately (0.96, 0.04) in the test distribution.

Example 1.11.2

The ornithologist also counts the number of breeding malachite sunbird males visiting each of 24 different protea bushes and finds 0, 1 and 2 birds at 8, 14 and 2 of the 24 sites, respectively. The ornithologist again wants to test the quality of fit of a Poisson class. In the present case

(the data variance)/(the data mean) = 0.493,

of which the modelled counterpart in the human mind is then found to be co-ordinated at approximately (0.02, 0.98) in the test distribution.

In each of the two cases we have obtained a poor fit that, in each case, prompts the question, 'Might that not be pointing at an explanation other than mere coincidence?' The variance of a Poisson sample is usually close to the mean; so for the sugarbirds the variance seems to be too large, and for the sunbirds the variance seems to be too small. Moreover, ornithology can, in each case, adjoin substantive facts pointing at a substantively conceivable alternative explanation, as follows: on the one hand, outside the breeding season sugarbirds do not display territorial behaviour; so the relatively large variance observed in their case is not unexpected, owing to aggregation at food sources. On the other hand, within the breeding season sunbird males are aggressive, and are often to be seen chasing conspecifics and other sunbirds; so the relatively small variance observed in their case is not unexpected, owing to territorial behaviour.

The two examples involved the same hypothesised model and the same partial ordering of sample patterns. They differed only in the role of the alternatives, where evidence favouring the alternative in the first example would be vacuous as evidence favouring the alternative in the second example, and *vice versa*. Clearly then, in order for numerically extreme co-ordinates arising from the test of a hypothesised model to provide evidence

against the tenability of that model, a substantively conceivable alternative source of the observed data pattern must be brought forward.

The foregoing development shows that, given an ordering of sample patterns arising from a particular hypothesised model, a certain alternative might be indicated by small values of the left co-ordinate, and a different alternative might be indicated by small values of the right co-ordinate. This is not unusual. Let a data set of the form y_x (x = 1, 2, 3, ···, n) be modelled as a sample of $\underline{n}$ independent realisations of $Y_x = -1$ or $Y_x = +1$ with equal frequency. Consider an ordering based on the magnitude of the covariance of y_x and x. A negative covariance (a small left co-ordinate) might point at an increasing frequency of $Y_x = -1$ as the alternative. A positive covariance (a small right co-ordinate) might point at an increasing frequency of $Y_x = +1$ as the alternative. It might also be that only one of the two alternatives is substantively conceivable. Consider, for example, investigating the efficacy of sulphur applications for the control of stem rust in wheat. Consider ten pairs of pseudo-randomised plots dusted with different, and not excessively high, levels of sulphur, as follows:

{(x units of sulphur), (x+1 units of sulphur)} where x = 0, 1, 2,..., 9.

Let $y_x = +1$ when the plot receiving the higher level of sulphur is less affected by stem rust, and let $y_x = -1$ otherwise. It is entirely realistic that the investigator might not be prepared to give credence to the possibility that any of the sulphur applications could *increase* the level of rust infection. So, on the one hand, if the mental correlate of the co-variance of y_x and $\underline{x}$ is co-ordinated at (0.9, 0.1), we might consider that indicative, albeit slightly, of stem rust having been controlled by the sulphur applications. On the other hand, if the mental correlate of the co-variance of y_x and $\underline{x}$ is co-ordinated at (0.1, 0.9), we would regard that as a good fit of the hypothesised model 'no effect', and giving no indication that stem rust was controlled by any of the sulphur applications. In order to avoid any misunderstanding in such cases, we now introduce a scaffolding symbol we refer to as *the pointer*. When we use the symbol to label one member of a co-ordinate pair, that member is identified as the co-ordinate that might be pointing (by way of smallness) or that might not be pointing (by way of largeness) at a specified alternative under consideration. For example:

In the case of the sugarbirds: (0.96, 0.04*).
In the case of the sunbirds: (*0.02, 0.98).

In the 1st case of the sulphur applications (0.9, 0.1*).
In the 2nd case of the sulphur applications (0.1, 0.9*).

Again, (0.08, 0.92*) would indicate an unusually good fit, whereas (*0.08, 0.92) would indicate a moderately poor one.

The pointer defines *the pointing co-ordinate*.

It might be thought that introduction of a pointer risks redundancy. We do not dispute that and note instead that R.A. Fisher left certain statisticians under the impression (or perhaps the faulty impression) that a significance test does not rely on any alternative to the hypothesised model (Jeffreys 1961, p. 377; Edwards 1972, pp. 177, 178, 180). So we wish to underscore the following. Let any co-ordination test of a hypothesised model produce a poor fit to a given data pattern. If we are unable to provide a substantively

credible alternative explanation of how that pattern might have come about, we can but ask for further data, perhaps to see if that kind of pattern will recur. The following three examples will help to make this clear.

Example 1.11.3

This writer was once asked to examine a thesis on *inter alia* results obtained in five different 2×2 factorial trials with sheep. The thesis reported five independent values of Snedecor's *F* ratio when testing for treatment interactions, none of which, so it was claimed, provided any evidence of interaction. The five computed *F* ratios co-ordinated as follows in Snedecor's test distribution:

(0.07, 0.93), (0.02, 0.98), (0.04, 0.96), (0.00, 1.00), (0.11, 0.89).

Having closely questioned the candidate about the conduct of the trials, this writer in effect held that the co-ordinations should be viewed as follows, pointing at improper randomisation:

(*0.07, 0.93), (*0.02, 0.98), (*0.04, 0.96), (*0,00, 1.00), (*0.11, 0.89).

Example 1.11.4

An investigator wishes to test, for each of three different species of fynbos, whether or not plants from seeds gathered at four different sites differ in growth potential under arid conditions. The investigator uses three separate and properly randomised designs and, testing for site mean differences (treatment differences), obtains values of Snedecor's *F* that co-ordinate as follows in Snedecor's test distribution:

(0.09, 0.91*), (0.07, 0.93*), and (0.11, 0.89*).

The investigator may conclude, reasonably, that these co-ordinations do not provide any evidence of site differences.

Example 1.11.5

A geologist wishes to test whether or not sandstone pebbles gathered at ten different sites differ in constituent grain size. Testing for site mean differences, the geologist obtains a value of Snedecor's *F* that co-ordinates at (0.03, 0.97*) in Snedecor's test distribution. The geologist might wonder whether something went wrong; might re-check the data record; might re-examine graphical representations; might repeat the various calculations testing for normality, homogeneity of variance, and so on. But, if nothing untoward is found, the geologist, if unwilling to give credence to the unusual data pattern as just such by mere co-incidence, can but resort to gathering further data, perhaps to see if a similarly unusual pattern will recur.

An inescapable fact of investigative science

It is inescapable that, depending on prior experience, one investigator might notice a data pattern of alternative significance, of which another investigator might be utterly oblivious. For a remarkable real-life example the reader is referred to an exchange of papers between Berkson (1942) and Fisher (1943).

1.12 Data analysis, science and sufficiency

The development of a sound theory of data analysis must begin with Immanuel Kant's recognition that we cannot know anything about a thing-in-itself (*ding-an-sich*). If, for instance, we say Thembi is 'a woman', 'small', 'black', we find that we are using universal terms, rather than terms limited to Thembi's proprietary. If we try very hard to overcome this by, for instance, saying: ' ... but Thembi has a tiny little mole on her left cheek', we again use universal terms. How else could the reader possibly grasp what is being said? We touched upon this in a previous section when explaining our usage of the terms 'sampling' and 'sample' as referring to universal concepts that cannot be made part of the proprietary of particular data sets. In this section we wish to come to grips with certain fundamental principles of statistical data analysis, where a data set is then always a 'thing' whose proprietary cannot be conveyed in any language. So it will be found that, willy-nilly, any principle for statistical 'data' analysis always seems to turn into one for statistical 'sample' analysis so to speak. This is important only inasmuch as it needs to be clearly understood that a 'thing' can be invested with meaning, only by way of making universal concepts address it. So we may anticipate that principles for scientific data analysis will turn out to be principles for analysis of the scientific concepts that might address those data. In our case (the statistical case), the first step in that direction is to analyse any class of models giving the probability of the sample into two main subsidiary models: on the one hand, we must obtain the subsidiary that represents *the characteristics of the class as a whole* and, on the other hand, we must obtain the subsidiary that represents *all the different members of the class*. We have in fact already met examples of such an analysis at (1.8.1) and (1.8.2), respectively. We now wish to develop the principles of such analysis. So let a data set given by $\underline{n}$ numbers, $\{x_1, x_2, x_3, \ldots x_n\}$, be modelled as a random sample, of which the long-run frequency in terms of corresponding random variables, $\{X_1, X_2, X_3, \ldots X_n\}$, and a vector of parameters, $\Theta = (\theta_1, \theta_2, \theta_3, \ldots, \theta_m)$, is given by

$$\Pr(x_1, x_2, x_3, \ldots, x_n; \Theta) \text{ in an obviously simplified notation.}$$

We envisage Θ as taking different values. We thus envisage a class of models indexed by the different values of Θ. Now bear in mind that the very essence of any scientific model is the universe of all predictions that can be obtained from it. In our case every member of our class of models predicts the long-run frequency of each of the possible sample patterns envisaged for that class, and from that further predictions can then be derived. We might, for instance, derive the frequency with which the mean of $X_1, X_2, X_3, \ldots, X_n$ is predicted to exceed a given value. Or we might wish to predict what the average value, over repetitions, of the variance of $X_1, X_2, X_3, \ldots, X_n$ would be. This prompts a concern that replacement of the sample with a set of summary statistics can inadvertently result in diminished predictive capacity; in other words, the summary might be *insufficient for all of the predictions of the class*. Consider a set of statistics $\{T_1, T_2, T_3, \ldots T_n\}$, defined by a one-to-one transformation:

$$t_i = t_i(x_1, x_2, x_3, \ldots, x_n) \text{ for } i = 1, 2, 3, \ldots, n.$$

As the transformation is one-to-one, the original numbers can be recovered by inverse transformation. So, any prediction that can be derived from $\{X_1, X_2, X_3, \ldots, X_n\}$ can also be derived from $\{T_1, T_2, T_3, \ldots, T_n\}$. We will express this by saying that the set of statistics $\{T_1, T_2,$

T_3, ..., T_n} is *sufficient for the predictions of the class*. A familiar example of this is the transformation from a number of 'yields' to the same number of independent degrees of freedom in the analysis of variance, when the grand total of the 'yields' is viewed as one of the degrees of freedom (See for instance Fisher 1970, p. 120). This motivates Definition 1.12.1.

Definition 1.12.1:

Let a data set $S_x = \{x_1, x_2, x_3, ..., x_n\}$ be modelled as a random sample $S_X = \{X_1, X_2, X_3, ..., X_n\}$, which arises from a class of models indexed by $\Theta = (\theta_1, \theta_2, \theta_3, ..., \theta_m)$. Let $\underline{k}$ transformations

$t_i = t_i(x_1, x_2, x_3, ..., x_n)$ for i = 1, 2, 3, ..., k

define a set of statistics $S_T = \{T_1, T_2, T_3, ..., T_k\}$. If and only if all the predictions that can be derived from S_X can also be derived from S_T, we say that S_T is *sufficient for the predictions of the class*.

Let $\{T_1, T_2, T_3, ..., T_k\}$ be sufficient for the predictions of a class, and let T_{k+1} be any function of $\{T_1, T_2, T_3, ..., T_k\}$. Then $\{T_1, T_2, T_3, ..., T_k, T_{k+1}\}$ will also be sufficient for the predictions of the class, but *redundantly* so. This brings us to the concept of a necessary and sufficient set of statistics, or (in more customary language) *a minimally sufficient set of statistics*, where this motivates Definition 1.12.2.

Definition 1.12.2:

Let a data set $S_x = \{x_1, x_2, x_3, ..., x_n\}$ be modelled as a random sample $S_X = \{X_1, X_2, X_3, ..., X_n\}$, which arises from a class of models indexed by $\Theta = (\theta_1, \theta_2, \theta_3, ..., \theta_m)$. Let $\underline{k}$ transformations

$t_i = t_i(x_1, x_2, x_3, ..., x_n)$ for i = 1, 2, 3, ..., k

define a set of statistics $S_T = \{T_1, T_2, T_3, ..., T_k\}$. Then S_T is *minimally* sufficient for the predictions of the class if and only if it is both sufficient for the predictions of the class and can be derived from any other set of statistics that is sufficient for the predictions of the class.

Trivially, S_X is itself minimally sufficient for the predictions of the class. We wish to analyse the predictions of the class into all those that are characteristic of the class as a whole, and all those that distinguish the different members of the class from one another. So, we introduce Definition 1.12.3.

Definition 1.12.3:

Let a data set $S_x = \{x_1, x_2, x_3, ..., x_n\}$ be modelled as a random sample $S_X = \{X_1, X_2, X_3, ..., X_n\}$, which arises from a class of models indexed by $\Theta = (\theta_1, \theta_2, \theta_3, ..., \theta_m)$. Let $\underline{k}$ transformations

$t_i = t_i(x_1, x_2, x_3, ..., x_n)$ for i = 1, 2, 3, ..., k

define a set of statistics $S_T = \{T_1, T_2, T_3, ..., T_k\}$. Then we say:

S_T is *sufficient for the class* if and only if it can supply all the predictions that can be made with the given class of models.

S_T is *sufficient for the index* if and only if it can supply all the indexed predictions that can be made with the given class of models. (An *indexed prediction* is one whose physical meaning depends on the value of the index.)

S_T is *sufficient for the characteristic* if and only if it can supply all the predictions that are characteristic of the given class of models as a whole.

The members of the class are distinguished from each other by the index of the class. So the predictions that can differentiate between the members of the class are those that

depend on the index. As the sample itself, S_X, is sufficient for the class, it is also sufficient for the index. And, as S_X comprises $\underline{n}$ statistics, we might ask whether it is possible to summarise 'whatever S_X can "tell" us about the index' in terms of fewer than $\underline{n}$ statistics. Here the phrase 'whatever S_X can "tell" us about the index' is short for 'whatever S_X can provide by way of predictions that depend on the index'. In fact we have already met this idea at (1.8.1) and (1.8.2), where we saw how a given data set might be modelled as a sample from the one or the other, of various members of a class of models indexed by a parameter, and where we saw how the probability of the sample could then be factored such that a single scalar statistic sufficient for the index is given by one of the factors. So, let us re-label the T-like statistics as:

$$\{T_1, T_2, T_3, ..., T_n\} = \{R, C_1, C_2, C_3, ..., C_{n-1}\}$$

and let us consider how to define R as being sufficient for the index. (For the moment we are thinking of R as a scalar random variable.) In order to allow for certain trivial cases, a factorisation of the form

$$\Pr(r; \Theta) \times \Pr(c_1, c_2, c_3, ..., c_{n-1} \mid r)$$

might have to be interpreted as $\Pr(r; \Theta)\times(1)$, in which case $c_1, c_2, c_3, ..., c_{n-1}$ are any arbitrary constants. This being understood, we introduce Definition 1.12.4.

Definition 1.12.4:

Let $\{R, C_1, C_2, C_3, ..., C_{n-1}\}$ be a set of statistics derived from a random sample of size $\underline{n}$ arising from a class of models indexed by a vector $\Theta = (\theta_1, \theta_2, \theta_3, ..., \theta_m)$. Let the probability of the sample be given by

$$\Pr(r; \Theta) \times \Pr(c_1, c_2, c_3, ..., c_{n-1} \mid r)$$

where the first factor is the unconditional probability that R = r, and the second factor is the conditional probability that

$$(C_1, C_2, C_3, ..., C_{n-1}) = (c_1, c_2, c_3, ..., c_{n-1}) \text{ given that } R = r.$$

R is *sufficient for the index* if and only if the second factor is independent of Θ.

It should be noted that the phrase 'independent of', as used in Definition 1.12.4, refers not only to frequency, but also to range. For instance, if $\theta \neq 0$ and C ranges over the values $\theta, 2\theta, 3\theta, ..., 10\theta$ with equal frequency, that frequency, 0.1, is independent of θ, but the distribution of C is *not* independent of θ. Definition 1.12.4 hardly requires any explication; $\Pr(c_1, c_2, c_3, ..., c_{n-1} \mid r)$ stands for a predicted long-run frequency from which further predictions can be derived, and as the initial prediction is independent of Θ, the further predictions would also be independent of Θ. So the predicted long-run frequency that $\Pr(r; \Theta)$ stands for can provide whatever indexed predictions the class of models can provide.

Sometimes no scalar statistic sufficient for the index exists. However, as the sample itself is always sufficient for the index, there always exists a set of statistics *jointly sufficient* for the index. Consider for instance a sample of $\underline{n}$ integers drawn at random without replacement from the set of all the integers straddled by two integers, θ and 2θ, where $\theta > 3$ and $2 \leq n \leq \theta\text{-}1$. The probability of the sample is given by:

$$\binom{\theta\text{-}1}{n}^{-1}.$$

Given the values of the smallest number drawn, $x_{(1)}$, and of the largest number drawn, $x_{(n)}$, the conditional probability of the other n-2 sample values is given by

$$\begin{pmatrix} x_{(n)}-x_{(1)}-1 \\ n-2 \end{pmatrix}^{-1} .$$

So the probability of the sample factors as follows:

$$\left[\begin{pmatrix} x_{(n)}-x_{(1)}-1 \\ n-2 \end{pmatrix}\begin{pmatrix} \theta-1 \\ n \end{pmatrix}^{-1}\right] \times \left[\begin{pmatrix} x_{(n)}-x_{(1)}-1 \\ n-2 \end{pmatrix}^{-1}\right] . \qquad (1.12.1)$$

Here the first factor in square brackets gives the joint distribution of $X_{(1)}$ and $X_{(n)}$, and its range depends on θ, as $\theta+1 \leq X_{(1)} < X_{(n)} \leq 2\theta-1$. The second factor in square brackets gives the joint distribution of the other n-2 sampled integers. Its range is independent of θ, being bounded by $x_{(1)}$ and $x_{(n)}$. So we have that $X_{(1)}$ and $X_{(n)}$ are jointly sufficient for θ. Moreover, as

$$\theta < X_{(1)} < X_{(n)} < 2\theta,$$

each one of the two X-like statistics tells us something about θ that the other one cannot tell us. For instance, $X_{(1)} = 20$ would tell us that θ must be < 20, but not by how much, and $X_{(n)} = 30$ would tell us that θ must be > 15, but not by how much. So we need *both* statistics for the sample to tell us *all* that it can tell us about θ. We express this by saying:

$X_{(1)}$ and $X_{(n)}$ jointly, are *minimally sufficient* for θ.

It can also be that a single scalar statistic is sufficient for a vector of two independent scalar parameters. For instance, let X be the sum of two independent random numbers; the first a random one of the integers 1, 2, 3, ..., ι ($\iota < \infty$) and the second a random one of the fractions 0.1, 0.2, 0.3, ..., ϕ ($\phi < 1$). If X = 7.3, say, its integral part tells us $\iota \geq 7$, but tells us nothing at all about ϕ, whereas its fractional part tells us $\phi \geq 0.3$, but tells us nothing at all about ι. Therefore X is minimally sufficient for (ι, ϕ). This leads to Definition 1.12.5.

Definition 1.12.5:

Let $\{R_1, R_2, R_3, \ldots R_k, C_1, C_2, C_3, \ldots, C_{n-k}\}$ be a set of statistics derived from a random sample of size n arising from a class of models indexed by a vector $\Theta = (\theta_1, \theta_2, \theta_3, \ldots, \theta_m)$. Let the probability of the sample be given by

$$\Pr(r_1, r_2, r_3, \ldots, r_k; \Theta) \times \Pr(c_1, c_2, c_3, \ldots, c_{n-k} \mid r_1, r_2, r_3, \ldots, r_k)$$

where the first factor is the unconditional probability that

$$(R_1, R_2, R_3, \ldots, R_k) = (r_1, r_2, r_3, \ldots, r_k)$$

and the second factor is the conditional probability that

$$(C_1, C_2, C_3, \ldots, C_{n-k}) = (c_1, c_2, c_3, \ldots, c_{n-k})$$

given that $(R_1, R_2, R_3, \ldots, R_k) = (r_1, r_2, r_3, \ldots, r_k)$.

Then the set of statistics $\{R_1, R_2, R_3, \ldots, R_k\}$ is *sufficient for the index* if and only if the second factor is independent of Θ.

Definition 1.12.6 then tells us that no set of statistics could tell us anything about the index that a set minimally sufficient for the index cannot tell us.

Definition 1.12.6:

A set of statistics is *minimally sufficient for the index* if and only if it is both sufficient for the index and can be derived from any set of statistics that is sufficient for the index.

The concept of a set of statistics minimally sufficient for the index now enables us to define a counterpart concept of a set of statistics minimally sufficient for the characteristic. This is accomplished by Definitions 1.12.7, 1.12.8 and 1.12.9.

Definition 1.12.7:

Let R_1, R_2, R_3, ..., R_k, C_1, C_2, C_3, ..., C_{n-k} be $\underline{n}$ statistics derived from a random sample of size $\underline{n}$ arising from a class of models indexed by a vector $\Theta = (\theta_1, \theta_2, \theta_3, ..., \theta_m)$. If ($R_1$, R_2, R_3, ..., R_k, C_1, C_2, C_3, ..., C_{n-k}) is sufficient for the class, then (C_1, C_2, C_3, ..., C_{n-k}) is *sufficient for the characteristic* if and only if (R_1, R_2, R_3, ..., R_k) is minimally sufficient for the index.

Definition 1.12.8:

A set of statistics is *minimally sufficient for the characteristic* of a class of statistical models if and only if it is both sufficient for the characteristic, and can be derived from any other set of statistics that is sufficient for the characteristic. Such a set of statistics may meaningfully be referred to as *the* class characteristic.

Definition 1.12.9:

Consider a set of statistics derived from a set of statistics that is minimally sufficient for the characteristic of a class of statistical models, without being itself minimally sufficient for the characteristic of the class. Such a set of statistics may be meaningfully referred to as *a* class characteristic.

We note that in Example 1.8.1, *the* class characteristic, which is displayed in Table 1.8.1, leads directly to a suitable test of fit. In the case of Example 1.8.2, however, we obtained a test of fit by replacing *the* class characteristic by *a* class characteristic.

We note in passing that a derivation of the normal class characteristic is given by Kempthorne and Folks (1971, pp. 260-261). We also note that in the examples we have given, it is easy to see how the probability of the sample must be factored so as to provide an analysis in terms of the two main subsidiary models, but that is not always the case. So certain special methods for achieving such factoring have been developed by, for instance, Lehmann and Scheffé (1950, 1955). However, that is a technical matter and so, for the purposes of the present book, need not concern us.

In Section 1.15 we return to the foregoing. First, however, there are two matters that must briefly be remarked upon.

1.13 Remark on ancillary frames of reference

In order to avoid a possible misunderstanding of the concept 'sufficiency', we revisit the mixed sampling problem, as exemplified in Section 1.5. Recall that μ, an unknown quantity, would, in terms of our model, be measured precisely by Instrument A, and be measured with equally frequent errors, -2, -1, 0, +1, +2 by Instrument B. We then flip a coin of balanced sort to pick an instrument to make a measurement. This introduced a class of models indexed by μ ($-\infty < \mu < +\infty$), and to express this mathematically, we now define a random variable, Y, as follows:

Y = 1 when Instrument A is used to make the measurement. Pr(Y = 1) = 0.5.
Y = 2 when Instrument B is used to make the measurement. Pr(Y = 2) = 0.5.

If a random variable X then denotes the modelled value of the measurement, X and Y are jointly distributed as in Table 1.13.1. As {X, Y} is the sample, and as the sample is

Table 1.13.1: A class of models involving an ancillary partition

	X = μ-2	X = μ-1	X = μ	X = μ+1	X = μ+2	Total
Y = 1	0	0	0.5	0	0	0.5
Y = 2	0.1	0.1	0.1	0.1	0.1	0.5
Total	0.1	0.1	0.6	0.1	0.1	1

always sufficient for the index, {X, Y} is sufficient for the index. As any prediction that can be derived from Y alone is independent of the index, it may be thought that X alone is sufficient for the index. However, that is not the case because, for instance, the following pair of indexed predictions cannot be made without involving Y:

$$\Pr(X > \mu \mid Y = 1) = 0 \text{ and } \Pr(X > \mu \mid Y = 2) = 0.4. \qquad (1.13.1)$$

So {X, Y} is minimally sufficient for the index. Now if the whole of our model is to be considered as the frame of reference within which we wish to predict the frequency of the event 'X > μ', we obtain the prediction

$$\Pr(X > \mu) = 0.2. \qquad (1.13.2)$$

Comparison of predictions (1.13.1) and (1.13.2) shows that the statistic Y is an indicator of two different subsidiary frames of reference, corresponding to which substantive science has in this case provided Instruments A and B, respectively. Such statistical indicators of possible subsidiary frames of reference are known as *ancillary statistics*. In Section 1.5 we saw that they lead to statistical questions, exemplified at (1.5.1) and (1.5.2), which mathematical statistics cannot possibly hope to answer, as they are questions exemplifying Gödel's incompleteness principle. The principle asserts that in any rigidly logical mathematical system there arise questions that cannot be answered on the basis of the axioms within that system. They also cannot be resolved by imbedding the given logical system in a more extended logical system, as Gödel's incompleteness principle would also apply to the more extended logical system. They might, however, be resolved by extra-logical

means. For instance, in the case of Miss Evelyn Rosenthal's finite geometry, two different physical meanings are displayed in Figures 1.5.1(a) and 1.5.1(b), where physical science might select just one of the two as the *intended* physical meaning. So we must take the position that as mathematical statistics is just a servant of substantive physical science, it is in substantive physical science that we will find correct answers to the seemingly unanswerable questions raised by ancillary statistics. Stated otherwise, if there were several possible frames of reference, a minimal sufficient statistic would cater for *all* the predictions arising in those several possible frames of reference, and substantive science must then select the appropriate frame of reference. At this point it would be premature to try to develop this matter any further. We raised the matter here only to prove that if a statistic is minimally sufficient for the indexed predictions of a class, it does not follow that all the predictions that can be made with that statistic will depend on the index of the class.

1.14 Remark on continuous sample spaces

One cannot measure a continuous variable exactly. Any data set therefore involves the representation of physical experiences on a discrete and essentially finite grid of class marks. This means that, in principle, the fundamental theory of statistical data analysis can be exhaustively formulated in terms of discrete sample spaces only. That does not imply that we have to avoid any 'continuous' sample spaces, whether for convenient approximation when appropriate, or for taking into account an underlying theoretical source of variation. It implies, however, that our development is not open to criticism on the grounds of formulating the fundamentals of statistical data analysis in terms of discrete sample spaces only. Given the purposes of the present book, there would be no point in becoming involved in unnecessary technicalities. So we take the position that the problem of extending fundamental theory in order to accommodate continuous sample spaces is a purely technical matter, and does not involve any intrinsically different principles of reasoning. We involve continuous sample spaces only for the sake of exhibiting the consequences of our development in familiar contexts, or when that achieves simplification without loss of principle. This approach also makes for a development that is accessible to a wider audience than would otherwise be the case.

1.15 Definition of a commencement test

The development in Section 1.12 has far-reaching consequences, just one of which is embodied in Definition 1.15.1.

Definition 1.15.1:

Let a class of statistical models for the representation of given data be analysed into the two *main subsidiary models*, these being the subsidiary model giving the set of statistics minimally sufficient for the characteristic, and the subsidiary model giving the set of statistics minimally sufficient for the index. A test of fit that involves the former subsidiary only is a *commencement test*. A test of fit that involves the latter subsidiary only is an *elimination test*. (For reasons to be explained in subsequent development, we take the position that a test of fit that tries to involve both subsidiaries is an ill-conceived one.)

This chapter concerns commencement tests, of which simple examples were given in Section 1.8. However, the relation between sections 1.7 and 1.8 prompts Definition 1.15.2 as being more inclusive than Definition 1.15.1.

Definition 1.15.2:

A *commencement test* is a test of the quality of fit of an isolated singleton.

In terms of Definition 1.15.2, the tests presented *both* in Section 1.7 and in Section 1.8 are commencement tests. In each of the two examples in Section 1.8 the test statistic was explicitly derived from the class characteristic. Another approach is to derive a commencement test by developing a test statistic that is invariant with respect to the class index, as is done in the following three examples.

Example 1.15.1

In a study of induced repression, Lowenfeld (1955) had fifteen subjects learn ten nonsense syllables. He then attempted to associate a negative response to a random five of those syllables by giving the subjects an electrical shock whenever one of those five syllables was shown tachistoscopically. (Presumably the syllables for the shock treatment were chosen randomly per subject.) After a lapse of 48 hours each subject tried to recall as many of the syllables as possible. The subjects and their responses are listed in Table 1.15.1 in non-decreasing order of the total numbers of syllables recalled. This listing

Table 1.15.1: Numbers of shock and non-shock syllables recalled by 15 subjects

Subject number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Control syllables recalled	1	1	2	3	4	2	2	4	4	5	3	3	4	5	5
Shock syllables recalled	0	2	1	0	1	3	3	2	2	2	4	4	3	2	3
Total number recalled	1	3	3	3	5	5	5	6	6	7	7	7	7	7	8

displays the advantage of each subject providing an own control; subjects 1 through 4, for instance, recalled only half as many or fewer syllables as did subjects 8 through 15. An appropriate model must therefore account for different subject effects. So, let the probability of any shock-syllable being recalled by the j^{th} subject be modelled as:

$$\theta_j = \frac{\exp(\alpha_j+\beta)}{1+\exp(\alpha_j+\beta)}$$, where α_j is the subject effect, and β is the shock effect,

thus giving the linear logistic model $\ln\left\{\frac{\theta_j}{1-\theta_j}\right\} = \alpha_j+\beta$ for j = 1, 2, 3, ..., 15.

The usual analysis would model the data as 15 pairs of binomial samples (Cox 1970). But here the defence of such a model on theoretical grounds is awkward, unless we can appeal to a historical record. By way of contrast, a binomial model for the number

of outcomes 'bowl up' from 35 spins of a spoon is easily defended on the theoretical grounds that the spinning of spoons is akin to the flipping of coins, the rolling of dice, and the shuffling of cards. But, as we cannot reasonably hold that subjects somehow 'flip', 'roll' or 'shuffle' memories in the mind, we would need an empirical defence for our model by way of a satisfactory result from a commencement test for binomial sampling. Such a test is obtained by using the standard analysis of variance for paired comparisons to compute the error sum of squares on the arc sine scale (Snedecor and Cochran 1989, p. 289). In the case of binomial sampling, this is modelled by chi-square times the theoretical variance on arc sine scale. With the angles expressed in degrees, the computed sum of squares for the present data equals 3 014.4 on 15-1 df, and the theoretical variance is equal to 820.7÷5. So the value of the chi-square test datum is

$$3014.4 \div (820.7 \div 5) = 18.365 \text{ on } 14 \text{ df},$$

whose mental correlate is situated at (0.80, 0.20) in the test distribution. And in order to test for an additional source of error variation, the pointing co-ordinate in this case is on the right. The class characteristic, as tested, fits the given data well.

Example 1.15.2

The volcanic eruption of Vesuvius in AD 79 preceded ten eruptions over the next 1 552 years (Table 1.15.2). This brings to mind a Poisson process, in which case the waiting

Table 1.15.2: Waiting times for the first ten eruptions of Vesuvius after that of AD 79, given in years and in order of occurrence

124	269	40	275	81	23	8	8	29	595

times are modelled as a random sample from an exponential population, as described by the cumulative distribution function $1\text{-}\exp\{-x \div \theta\}$ for $0 < \theta < \infty$. A commencement test for this class of models, using the Cramer-Von Mises statistic, is obtained as follows: if X is a random value from a population with continuous cumulative distribution function F(x), then A = F(X) is a random value from a population that is uniformly distributed on the unit interval [a U(0, 1) population]. So if $X_{(1)}$, $X_{(2)}$, $X_{(3)}$, ... , is a random sample from a population of exponential waiting times, ordered in non-decreasing order of magnitude, then $A_{(1)}$, $A_{(2)}$, $A_{(3)}$, ... , for

$$A_{(1)} = 1\text{-}\exp\{-X_{(1)} \div \theta\},\ A_{(2)} = 1\text{-}\exp\{-X_{(2)} \div \theta\},\ A_{(3)} = 1\text{-}\exp\{-X_{(3)} \div \theta\},\ \ldots,$$

is a random sample from a U(0, 1) population, ordered in non-decreasing order of magnitude. For the given data, the mean waiting time is 155.2 years, which is an estimate of θ. So the A-like quantities can be estimated from the given data as

$$\hat{A}_{(1)} = 1\text{-}\exp\{-8 \div 155.2\},\ \hat{A}_{(2)} = 1\text{-}\exp\{-8 \div 155.2\},\ \hat{A}_{(3)} = 1\text{-}\exp\{-23 \div 155.2\},\ \ldots,$$

provided that *an exponential population* of waiting times is appropriate, where that is the basis of the test, as the A-like quantities can for *any population* be estimated as

$$\tilde{A}_{(1)} = (1\text{-}0.5) \div n,\ \tilde{A}_{(2)} = (2\text{-}0.5) \div n,\ \tilde{A}_{(3)} = (3\text{-}0.5) \div n,\ \ldots, \text{ for a sample of size } \underline{n}.$$

The Cramer-Von Mises test compares the Â-like estimates to the Ã-like estimates by means of the following test statistic, whose distribution is invariant with respect to θ:

$$\text{CvM} = \sum_{i=1}^{n} \left[\hat{A}_{(i)} - \tilde{A}_{(i)}\right]^2 + (12n)^{-1}.$$

The term $(12n)^{-1}$ merely serves to improve an approximation for the test distribution, which is more or less independent of sample size (Bain and Engelhardt 1989, p. 423) As a large value of the test statistic points at a poor fit, the pointing co-ordinate is on the right. In the present case CvM = 0.131, whose mental correlate co-ordinates at (0.85, 0.15*) in the test distribution. So, by the test performed, a model of exponentially distributed waiting times fits the given data well.

It is a substantive fact that the character of the volcano changed after AD 1631. Activity became continuous, alternating between so-called quiescent and eruptive stages. And it is a statistical fact that the waiting times for the next 20 eruptions tended to be much shorter than they were for the ten earlier eruptions (Table 1.15.3). The mean waiting time for these later eruptions is a mere 15.65 years, compared to 155.2 years for the earlier eruptions.

Table 1.15.3: Waiting times for the first 20 eruptions of Vesuvius after that of AD 1631, given in years and in order of occurrence

29	22	12	4	9	30	23	7	12	15	28	12	5	11	5	6	7	4	34	38

As judged by the Cramer-Von-Mises test, the later waiting times can also be satisfactorily modelled as a sample of exponential waiting times: CvM = 0.150, whose mental correlate co-ordinates at (0.87, 0.13*) in the test distribution. Furthermore, for all the eruptions taken together, a model of exponential waiting times, as judged by the Cramer-Von-Mises test, fits the data very poorly: CvM = 2.13, whose mental correlate co-ordinates at (1.00, 0.00*) in the test distribution. Thus substantive and statistical evidence taken together encourages geological opinion to hold that the volcanic action of Vesuvius from AD 79 up to AD 1944 falls into two distinctly and understandably different eras.

Example 1.15.3

A commencement test for the fixed-model analysis of variance was introduced in a brilliant paper by Tukey (1949a), as follows: consider a data set of responses arising from an experiment in randomised blocks, and make the notations:

Y_{ij} = the response arising from the ith treatment in the jth block.
$\overline{Y}_{\cdot j}$ = the mean response arising from the jth block.
$\overline{Y}_{i\cdot}$ = the mean response arising from the ith treatment.
$\overline{Y}_{\cdot\cdot}$ = the mean response arising from the trial as a whole.

The standard class of models for such data is based on the identity

$$Y_{ij} = Y_{\cdot\cdot} + (Y_{\cdot j} - Y_{\cdot\cdot}) + (Y_{i\cdot} - Y_{\cdot\cdot}) + d_{ij}, \qquad (1.15.1)$$

where d_{ij} denotes the following measure of observed non-additivity:

$$Y_{ij}-[\overline{Y}_{..}+(\overline{Y}_{.j}-\overline{Y}_{..})+(\overline{Y}_{i.}-\overline{Y}_{..})]. \tag{1.15.2}$$

In the absence of any systematic non-additivity, d_{ij} measures the unsystematic 'error' subject to which the first three terms on the right at (1.15.1) model the systematic part of response on the left. This suggests a class of additive models of the form

$$Y_{ij} = E(\overline{Y}_{..})+E(\overline{Y}_{.j}-\overline{Y}_{..})+E(\overline{Y}_{i.}-\overline{Y}_{..})+\epsilon_{ij}, \tag{1.15.3}$$

where different members of the class are indexed by different values of

$E(\overline{Y}_{..})$ = the general mean,
$E(\overline{Y}_{.j}-\overline{Y}_{..})$ = a deviation attributable to the additive effect of the jth block, and
$E(\overline{Y}_{i.}-\overline{Y}_{..})$ = a deviation attributable to the additive effect of the ith treatment,

where the response differs from the sum of these constants by way of the random variable:

ϵ_{ij} = the error of the response arising from the ith treatment in the jth block.

Let the errors be modelled as uncorrelated, homoscedastic, normal random variables, with expectations that are equal to zero. In order to test the quality of fit of additively systematic effects, as modelled at (1.15.3), we imbed the additive model into a wider class of models, which is broadly envisaged by taking the expected values at (1.15.3) to be the constant term and the first-degree terms of a Taylor expansion. The second-degree terms of the expansion would then be given by terms

$$\text{in } [E(\overline{Y}_{.j}-\overline{Y}_{..})]^2, \text{ in } [E(\overline{Y}_{i.}-\overline{Y}_{..})]^2 \text{ and in } [E(\overline{Y}_{.j}-\overline{Y}_{..})]\,[E(\overline{Y}_{i.}-\overline{Y}_{..})]. \tag{1.15.4}$$

The first two terms at (1.15.4) are absorbed by the additive effects, as they vary with *j* only and with *i* only, respectively. So the third term at (1.15.4) is the lowest-order term that could account for systematic non-additivity. Its estimated value,

$$(\overline{Y}_{.j}-\overline{Y}_{..})\,(\overline{Y}_{i}-\overline{Y}_{..}), \tag{1.15.5}$$

provides a predictor variable whereupon we can regress the observed measure of non-additivity given at (1.15.2). Following Snedecor and Cochran (1989, p. 284), let B denote the resulting linear regression coefficient. B = N/D when D denotes the sum of squares of the quantities arising at (1.15.5) and N denotes the sum of products of these quantities with those arising at (1.15.2). For non-additivity of the type to be tested for, we expect a non-zero regression coefficient. So we have introduced a class of models having E(B) = 0 owing to additivity, and we wish to test for non-additive alternatives broadly specified as having E(B) ≠ 0, owing to some or other non-additivity. Consider the distribution of B, conditional on the values of the predictor variable. As B is a contrast among the residuals, it is an error contrast in the additive case. The standardised form of the contrast is given by N/√D. By subtracting $(N/\sqrt{D})^2$ from the usual sum of squares for error, we obtain a residual sum of squares for error, with one degree of freedom less than the usual sum of squares for error. For a numerical example consider Table 1.15.4, which gives a data set reproduced by Snedecor and Cochran (1989, p. 256) concerning speci-

mens of three species of citrus, whose ratio of leaf area to dry weight was determined under three different conditions of shading.

Table 1.15.4: Ratio of leaf area to dry weight in three species of citrus under three conditions of shading

	Shamouti orange	Marsh grapefruit	Clementine mandarin
Sun	112	90	123
Half shade	86	73	89
Shade	80	62	81

For these data we find

Sum of squares for error = 87.11 on 4 degrees of freedom
$(N/\sqrt{D})^2$ = 57.50 on 1 degree of freedom
Residual sum of squares = 29.61 on 3 degrees of freedom

The standard error of B is estimated by

$\sqrt{s^2 \div D}$, where s^2 denotes the residual mean square
$= \sqrt{(29.61 \div 3) \div (178175)}$
= 0.007443

We find B = +0.017965. Our hypothesised model has E(B) = 0, and in order to test that against E(B) > 0, as indicated by the B value, and using Student's *t*, we find

$t = B \div \sqrt{s^2 \div D} = +2.414$ on 3 degrees of freedom,
whose mental correlate is found at (0.952, 0.048) in the test distribution. (1.15.7)

In order to attach a pointer we note that for each one citrus species involved, different leaves would tend to be of similar shape. So, if l is any given linear dimension, for instance $\sqrt{(length)(breadth)}$, the surface areas of the different leaves will tend to be proportional to l^2. Different leaves will also tend to be of similar, specific weight. So dry weight would tend to be proportional to fresh volume, which would in turn tend to be proportional to l^3. Thus, for different leaves the ratio of leaf area to dry weight would tend to be proportional to l^{-1}. One would expect additivity on the l scale, rather than on the l^{-1} scale and therefore the inverse transformation is indicated. Testing for non-additivity on the inverse scale we find

$t = B \div \sqrt{s^2 \div D} = +0.997$ on 3 degrees of freedom
whose mental correlate is found at (0.607, 0.393) in the test distribution. (1.15.8)

This provides an alternative to the hypothesised model, owing to which we can attach a pointer to the right statistical co-ordinate given at (1.15.7).

Note that the tests at (1.15.7) and (1.15.8) do not correspond to the conventional procedure. A co-ordination test that corresponds to the conventional procedure would have

us use Snedecor's F instead of Student's t, and would for instance at (1.15.7) have us calculate

$$\left[B \div \sqrt{s^2 \div D}\right]^2 = 5.827 \text{ on 1 and 3 degrees of freedom, whose mental correlate co-ordinates at (0.904, 0.096*) in Snedecor's test distribution.} \qquad (1.15.9)$$

This co-ordination test procedure is wrong, as it tries to test $E(B) = 0$ against $E(B) > 0$ and $E(B) < 0$ simultaneously. We return to this point in a subsequent section.

1.16 Remark on the generality of Definitions 1.15.1 and 1.15.2

In the event of small data sets, the precision that typifies co-ordination tests enables us to recognise apparently appreciable, but nevertheless not unusual, departures from the data patterns predicted by a hypothesised model. At the same time, that precision can often, despite meagre data, uncover patterns that depart strongly from those predicted by the model. Note, however, that neither of the two definitions in the previous section specifies a commencement test as necessarily being a co-ordination test. That is so because we must recognise that, especially with larger data sets, commencement tests also take other valuable forms, such as plotting data values against the expected values of the corresponding order statistics for sampling from a hypothesised singleton, and then judging whether or not the data points seem to scatter round a straight line. So we must avoid being prescriptive as to how investigators may test the quality of fit of a hypothesised singleton. We must rather try to establish the relative merits of different methods of commencement testing, bearing in mind only that in the usage of scientific data analysis, 'tests of fit' are tests that lead to physical facts of fit. That is to say, they are tests that lead to facts of fit that the human body can be compelled to grasp.

1.17 The specification of alternatives for a commencement test

Consider a scientific investigator who, at some stage, is able to say: 'In my opinion the possibilities can now be limited to those listed here.' In the statistical case, that would take the form of an investigator being able to say: 'In my opinion the possibilities can now be limited to *the members* of this class of models.' In neither case, however, can we view that as the situation at commencement. Thus, for instance, Tukey's test for non-additivity is a commencement test, as it uses a ruled surface as an approximation for an essentially boundless variety of non-additive alternatives. It would be silly to think we could gather all that variety into a tightly specified class of models.

Again, consider the problem of testing for whether or not the hypothesised model called 'random numbers' fits the output of a particular pseudo-random number generator. It appears at once that we cannot hope to imbed the hypothesised model into a class of models such that the alternatives embrace every possible form of non-random numbers. As we must then be leery of inadvertently precluding possibly important alternatives, we must, by way of a variety of tests, specify broad varieties of alternatives, where such specifications rely on our insight into the substantive subject matter involved. If, for instance, the numbers are generated by rolling a special die we will not expect to

find a periodicity in the output, but we might well test for digits that occur with unequal frequency. For a different kind of generator, however, we would test for periodicity of various kinds. The point here is this: if we test a hypothesised member of a class of models against other members of that class, *we can (and this is crucial) arrange for no member to be overlooked,* but in commencement testing *there can be no such assurance*. To give an example, we note that an early non-mechanical pseudo-random number generator involved a non-random pattern that escaped the notice of those who devised it, R.A. Fisher and F. Yates. In order to grasp the slippery nature of choosing appropriate alternatives in the case of commencement testing, we need only to grasp that it could also happen to us.

In order to help us grasp why the members of a class of statistical models cannot satisfactorily serve as the alternatives for a commencement test, let us try to devise such a test. Let the points of the sample space on which the models are to be defined be denoted by $x_1, x_2, x_3, \ldots$. Let the probabilities of these points be denoted by $P_{01}, P_{02}, P_{03}, \ldots$, for the hypothesised model, and by $P_{11}, P_{12}, P_{13}, \ldots$, for a fully specified statistical alternative. Then we have introduced the class of models

$$\Pr(X = x_j) = (1-\delta)P_{0j}+\delta P_{1j} \text{ for } j = 1, 2, 3, \ldots, \text{ where } \delta \in \{0, 1\}. \qquad (1.17.1)$$

The class arising at (1.17.1) consists of just two models indexed by the parameter δ. So we are asserting the appropriateness of a certain class of models without having established whether or not the characteristics of *the class as a whole* fit the given data satisfactorily. For instance, if the two members are characteristic of a Poisson class, and that of a negative binomial class, *neither* model might fit the data well. In commencement testing we must try to avoid such difficulties by specifying a broadly embracive variety of alternatives via partial orderings based on broadly descriptive concepts such as 'aggregated', 'over-dispersed', 'skewed', and so on. We thus specify certain data patterns (rather than certain models) as being patterns that are 'atypical of the hypothesised model' in the broad sense of being 'more frequent' under various broadly envisaged circumstances. The alternatives thus broadly envisaged are appropriately described by Definition 1.17.1 as being '*incipient*'.

Definition 1.17.1:

An *incipient* statistical model is a statistical model that is only broadly envisaged, that is to say, without rigid mathematical specification.

A typical commencement test thus involves 'an incipient *class* of models' whose members comprise one fully specified singleton, as the hypothesised model, and one or more incipient alternatives.

1.18 Remark on terminology

Cox and Hinkley (1974) use the term 'pure significance test' in a sense that is close to our term 'commencement test' and they then refer to some 'pure significance tests' as 'goodness of fit tests'. Kempthorne and Folks (1971) refer to all commencement tests as 'goodness of fit tests'. However, though Pearson's goodness of fit statistic is often used

for a commencement test, such as when testing for normality, it is also used for other kinds of tests, such as when, presupposing a binomial class of models, Bi(n, μ), to be applicable, it is used to test $\mu = 0.5$ against $\mu \neq 0.05$. So we will reserve the term 'goodness of fit' as the name of Pearson's statistic and of tests based on his statistic since that is after all the original usage. We will develop the term 'quality of fit' in a much broader sense. In fact, we will agree with Anscombe (1963) inasmuch as he holds that in the discourse of data analysis, the only tests of interest are tests of 'fit'.

1.19 Definition of co-ordination tests and significance tests

We must deliberately emphasise the distinction between hypothesis tests and the very different kind of tests required by a data analyst, as the statistical literature has largely failed to do so. We therefore follow Kempthorne and Folks (1971) by relinquishing the terms *null hypothesis* and *alternative hypothesis* and the corresponding notations H_0 and H_1, respectively, to the literature of hypothesis tests. We follow them also by using, for the purposes of the kind of tests being considered, the terms *hypothesised model* and *alternative model*, denoted by M_0 and M_1, respectively. These distinctions are not simply semantic or notational, as developments in subsequent chapters will show. We now introduce Definitions 1.19.1 and 1.19.2 in order to establish a formal relationship between significance tests and co-ordination tests.

Definition 1.19.1:

Let O_T for T = 1, 2, 3, ..., denote a partial ordering of all the sample patterns that arise from a singleton M_0 being put forward as a probability model for how given data might have come about. Let the given data be modelled as if a sample in O_t, where t denotes a particular value of T. A *significance test* of this model attaches to the given data the calculable number:

$$SL(t; 0) = \Pr(\text{a sample pattern} \in O_T \text{ for } T \geq t \mid M_0),$$

which is called the *significance level* for the given data with regard to the model M_0 and for the partial ordering chosen.

Definition 1.19.2:

Let O_T for T = 1, 2, 3, ..., denote a partial ordering of all the sample patterns that arise from a singleton M_0 being put forward as a probability model for how given data might have come about. Let the data be modelled as if a sample in O_t, where t denotes a particular value of T. A *co-ordination test* of this model attaches to the given data the calculable ordered number triplet:

$$C(t; 0) = [U(t; 0), \varepsilon(t; 0), V(t; 0)]$$

given by:

$$U(t; 0) = \Pr(\text{a sample pattern} \in O_T \text{ for } T < t \mid M_0),$$

$$\varepsilon(t; 0) = \Pr(\text{a sample pattern} \in O_T \text{ for } T = t \mid M_0), \text{ and}$$

$$V(t; 0) = \Pr(\text{a sample pattern} \in O_T \text{ for } T > t \mid M_0),$$

whose members are called the *statistical co-ordinates* and the *statistical rounding* for the given data with regard to the model M_0 and for the partial ordering chosen.

Definition 1.19.1 is essentially that of Kempthorne and Folks (1971, p. 222). Both of the definitions envisage the sample space as discrete and effectively finite. This must be so, as the recording of any data can only be by using a finite set of descriptions to describe a finite set of items. This is underscored by Theorem 1.19.1.

Theorem 1.19.1:

Let C = (U, ε, V) denote the statistical co-ordinates arising from a co-ordination test of the quality of fit of a given singleton M_0 being put forward as a probability model for how given data might have come about. Then U+V < 1, as ε cannot be equal to zero.

Hence, for any given data set and any partial ordering of the samples arising from any given singleton put forward as a model of how those data might have come about, the co-ordinates defined by Definition 1.19.2 cannot be recovered from the corresponding significance level defined by Definition 1.19.1. True, in practice a statistical rounding is usually very small, so that the co-ordinates can often be recovered almost precisely. However, that is beside the point here. The point is that the different definitions intentionally convey two very different meanings. So it would be foolish to brush the formal distinction aside without having a very clear understanding of each of the two different meanings being conveyed, and without having a very clear understanding of the consequences of reasoning in terms of one or the other of those two meanings. The reader is cautioned not to be impetuous on this point.

1.20 Derivation of co-ordination tests from given significance tests

Many optimal co-ordination tests can be derived from existing significance tests. In fact, from every significance test, a corresponding co-ordination test can be derived, though not always an optimal one. In order to show how such derived co-ordination tests are obtained, we need Definition 1.20.1 and Theorem 1.20.1.

Definition 1.20.1:

Let O_T for T = 1, 2, 3, ..., k denote any partial ordering of the samples arising from a given singleton. Then we refer to O_S for S = 1+k-T for T = 1, 2, 3, ..., k as the *inverted ordering*.

Theorem 1.20.1:

Let O_T for T = 1, 2, 3, ..., k denote any partial ordering of the samples arising from a given singleton, and let (A, B, C) give the co-ordinates of t (a particular value of T) with respect to that partial ordering. Then (C, B, A) gives the co-ordinates of t with respect to the inverted ordering.

The term 'significance test' is widely misused in the statistical literature as a name for various kinds of 'tests' that are demonstrably not significance tests. Our usage of the term is restricted to Definition 1.19.1. That being understood, Theorem 1.20.2 follows directly from Definitions 1.19.1, 1.19.2 and 1.20.1, as well as from Theorem 1.20.1.

Theorem 1.20.2:

Any significance test yields a *formal* co-ordination test because:
The complement of the left co-ordinate = the significance level for the chosen ordering
The complement of the right co-ordinate = the significance level for the inverted ordering
The statistical rounding = [(the sum of the two complements)-1]÷2.

Theorem 1.20.2 enables co-ordination tests to draw on the mathematics of significance tests. The theorem emphasises the term *'formal'*, because the distinction between the two kinds of tests cannot be mathematically (*formally*) understood; it can be understood only by grasping the distinction between the different scientific *meanings* conveyed. Once again, the reader is cautioned to be patient on this point.

1.21 A NORMATIVE PRESCRIPTION WITH DESTRUCTIVE CONSEQUENCES

The theory of hypothesis tests has exerted an exceedingly pervasive influence on the practices of statistical data analysis, particularly by way of a normative prescription that may for our immediate purposes be formulated in terms of the following rules:

(1) Let SL denote the significance level attached to given data by a significance test of a hypothesised model, M_0, against an alternative, M_1.

(2) Specify a test size α, which is a small fraction selected *without reference to the data*. A much favoured value is $\alpha = 0.05$.

(3) If $SL \leq \alpha$, reject M_0 and accept M_1. If $SL > \alpha$, accept M_0 and reject M_1.

In Chapter 4 we show how this prescription originates in a profoundly mistaken view. Here we merely prove that use of the prescription destroys scientific evidence, both of a statistical nature and of a substantive nature. We present three examples to prove the destruction of statistical evidence, and three further examples to prove the destruction of substantive evidence; in each of these two cases we present the three examples and then discuss them jointly. The reader will find that in each case the proof consists of presenting two of a kind and one that is different, and then showing that the use of the normative prescription results in a misclassification.

Example 1.21.1

The term 'learning' has a diversity of meanings, such as learning to judge distances, learning to ride a bicycle, and learning to find your way home. Some animals, sharks for instance, seem to have no learning ability. Other animals display various kinds of learning ability. It is even possible for different memory systems to occur in the same animal; in *Octopus* for instance visual and tactile memories are stored in anatomically distinct parts (Young, 1965). So, consider a sequence of just 13 trials in which an octopus either succeeds (S) or fails (F) to perform a task to be learned, as follows:

F, F, S, F, S, F, F, S, S, S, F, S, S. (1.21.1)

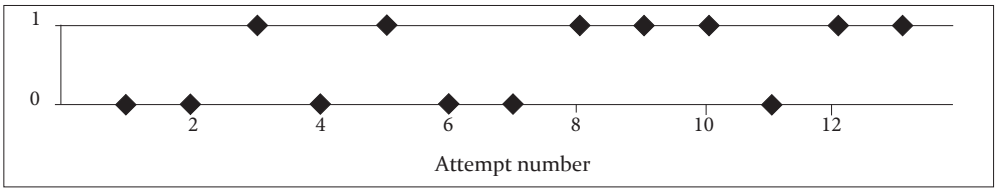


Figure 1.21.1: Success (1) or failure (0) in 13 consecutive attempts in a learning trial

A trend toward an increased frequency of success is made obvious by plotting the data as in Figure 1.21.1. In order to assess this more incisively, we consider the possibility, to which we need not grant credence, that the animal is unable to learn the task. That leads us to a hypothesised class of models in which the sequence is represented as the outcome of just 13 independent Bernoulli trials, with the constant probability of success denoted by μ ($0 < \mu < 1$). Let $y_x = 0$ or 1, depending on whether the outcome in the ordinal position x is F or S, respectively (x = 1, 2, 3, ... 13). Let $y_.$ denote the total number of successes. The data are then modelled by a sample of which the probability is given by:

$$\prod_{x=1}^{13} \mu^{y_x}(1-\mu)^{1-y_x} = \mu^{y_.}(1-\mu)^{13-y_.}$$

where this probability can be expressed as:

$$\left[\binom{13}{y_.}\mu^{y_.}(1-\mu)^{13-y_.}\right] \times \left[\binom{13}{y_.}^{-1}\right]. \qquad (1.21.2)$$

The first factor in square brackets at (1.21.2) tells us that under the hypothesised class, the total number of successes, $y_.$, is being represented as a realisation of a binomial random variable. It tells us nothing about the appropriateness of the class; the second factor is independent of μ and thus is the class characteristic; it tells us that, given the value of $y_.$, there are 13-choose-$y_.$ equally frequent sample patterns. We now wish to test the class characteristic against alternatives where Figure 1.21.1 reflects a trend towards an increased frequency of success, owing to learning ability. So we order all the possible sample patterns according to the size of the product moment correlation coefficient of y_x and x. For the data at (1.21.1) we find r = +0.453, which is a class mark for the interval from +0.433 to +0.473. The hypothesised class characteristic models this as having an approximately normal distribution given in an obvious notation by

$$N(\mu, \sigma^2) = N[0, (n-1)^{-1}],$$

where n = 13 in our case (Cox and Hinkley 1974, pp. 185-186). The mental correlate of the correlation is thus found to be co-ordinated at approximately (0.93, 0.02, 0.05*) in the test distribution.

Example 1.21.2

Suppose the sequence at (1.1.1) represents the results of another learning trial with an octopus, this time to see if the animal can be induced by rewards to learn a task that

is different to the one considered in the previous example. A trend toward an increased frequency of success, albeit slight, is clearly perceived when the data are plotted as in Figure 1.1.1. This trend is brought forward strongly by the kind of test performed in Example 1.21.1, as follows: the product moment correlation coefficient of y_x and x in this case proves to be r = +0.606, which is a class mark for the interval from +0.520 to +0.693. Using the approximation given in Example 1.21.1, the mental correlate of the present correlation is found to be co-ordinated at approximately (0.92, 0.05, 0.03*) in the test distribution.

Example 1.21.3

Suppose that for yet another different task to learn, an octopus performs as follows in just 15 successive attempts:

F, F, S, S, F, F, F, S, S, F, S, S, S, F, F.

Performing the same kind of test used in Examples 1.21.1 and 1.21.2 we find that the product moment correlation coefficient of y_x and x is given by r = +0.124, which is a class mark for the interval from +0.093 to +0.155. Using the approximation given in Example 1.21.1, the mental correlate of the present correlation coefficient is found to be co-ordinated at approximately (0.56, 0.08, 0.36*) in the test distribution.

Discussion of Examples 1.21.1, 1.21.2 and 1.21.3

Consider the following summary of tests performed in the three trials with *Octopus*:

Task	Observed correlation	Number of attempts	Level of co-ordination
Task 1	+0.453	$n_1 = 13$	(0.93, 0.02, 0.05*)
Task 2	+0.606	$n_2 = 9$	(0.92, 0.05, 0.03*)
Task 3	+0.124	$n_3 = 15$	(0.56, 0.08, 0.36*)

The investigator could hardly conclude otherwise than to consider that some evidence of learning ability, albeit slender, has been gathered in Tasks 1 and 2, whereas there is utterly no indication of learning ability in Task 3. Yet, if the test size, *which must be selected without reference to the data,* is to be $\alpha = 0.075$, the normative prescription destroys statistical evidence by mismatching Tasks 2 and 3, as follows:

Task	Significance level	Compared to α	Conclusion
Task 1	0.02 + 0.05 = 0.07	0.07 < 0.075	*Octopus* can learn Task 1
Task 2	0.05 + 0.03 = 0.08	0.08 > 0.075	*Octopus* cannot learn Task 2
Task 3	0.08 + 0.36 = 0.44	0.44 > 0.075	*Octopus* cannot learn Task 3

Apart from such destruction of statistical evidence, the normative rules also destroy substantive evidence. This is because the rules spring from the view that, concerning the matter being investigated, the investigator is ignorant to the extent of being utterly dependent on the numerical data. The rules do not provide for such knowledge as, for instance, knowing that *Octopus* can learn to respond to certain visual clues. Thus, if Task 2 required learning to respond to a previously untested visual clue, the rules will disallow an investigator to consider, along with the trial results, prior experience with visual clues. Conversely, there is no provision for the investigator to be surprised by a result. Consider,

for instance, that although *Octopus* adjust correctly for the weight of objects they pick up, they cannot (and this is the surprise) learn to distinguish between objects differing only in weight (Wells 1961). So, had Task 3 required learning to distinguish certain objects by weight only, the investigator would wish to report a surprising result and so would surely consider SL > 0.075 to be an obscuring description of that result. In fact, one can hardly imagine an investigator who would not consider SL > 0.075 to be an obscuring description of the results for both Tasks 2 and 3. Even for Task 1, the normative prescription obstructs scientific debate, as substantive science might with good reason be sceptical of the conclusion that the rules would in that case foist upon it. This point is strongly brought forward by the following three examples.

Example 1.21.4

A test of fit introduced by R. J. Strutt (later Lord Raleigh) is known as Raleigh's test. Let x_j, for $0° \leq x_j < 360°$ and j = 1, 2, 3, ..., n, denote a sample of independent angles from a population of angles uniformly distributed on the interval [0°, 360°). We wish to test the quality of fit of this, as a hypothesised model, against alternatives with angles tending to cluster around a 'preferred' angle. So the variety of equally frequent sample patterns must be replaced by a lesser variety of patterns such that an ordering of patterns by the magnitude of their frequencies would tend to detect any tendency to cluster round a preferred angle. In Raleigh's test we do so by vector addition of the unit vectors $(\sin x_j, \cos x_j)$ for j = 1, 2, 3, ..., n, and then ordering the possible samples according to the amplitude of the resultant vector, i.e. according to the magnitude of

$$q^2 = \left(\sum_{j=1}^{n} \sin x_j \right)^2 + \left(\sum_{j=1}^{n} \cos x_j \right)^2 . \qquad (1.21.3)$$

A large value of q^2 points at a preferred angle. Under the hypothesised model, the distribution of Q^2, the random variable corresponding to q^2, is approximately such that

$$\Pr(Q^2 > q^2) = \exp(-q^2/n) \qquad (1.21.4)$$

(Raleigh, 1880). When appropriate, a preferred angle is estimated by the *vector mean*, defined by Krumbein (1939) as the angle corresponding to the resultant vector. Figure 1.21.2 uses an equiareal rose diagram (a circular histogram) to display a data set given by Krumbein. It comprises 18 petals, so to speak, whose positions and relative areas make two opposing super-petals seem to appear. Glacial till pebbles were collected from a road cut through a late Wisconsin drumlin. The drumlin trends S 74°W, and is one among a field of drumlins whose average trend is S 82°W. The ice presumably travelled east to west from the Lake Michigan basin on the east. The directions of dip of the long axes of 100 pebbles were measured. Krumbein's interpretation of the data is that the pebbles tend to present their minimum cross-sectional areas opposed to the direction of flow, so that statistically the maximum cross-sectional areas, and thus the long axes, tend to lie parallel to the direction of movement. The dip of a pebble's long axis would then be independent of its direction of dip. So, a bimodal distribution would arise, as the dip of such a pebble contributes to one or the other of two opposite modes. So, adopting Krumbein's model as alternative

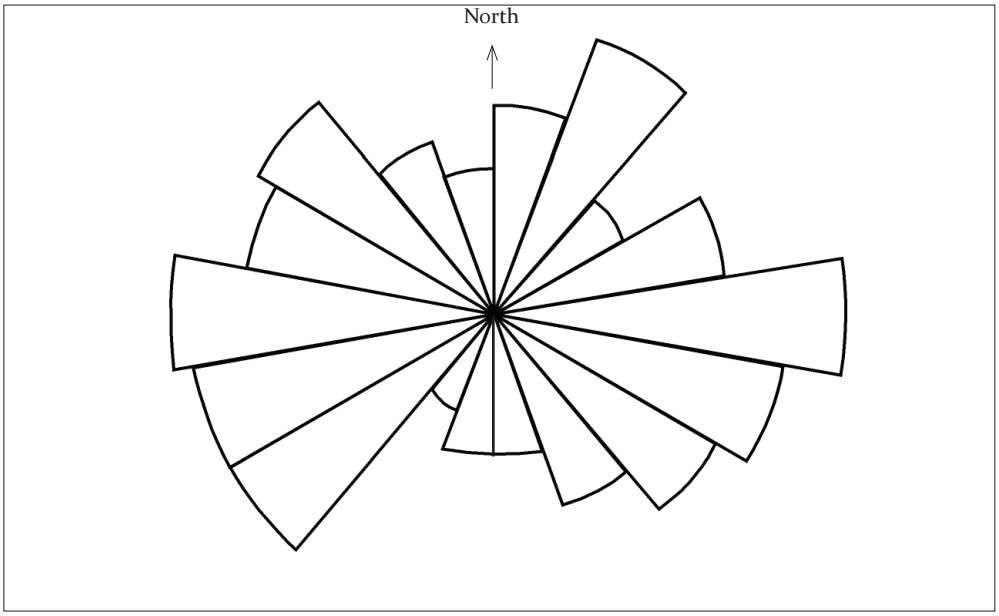


Figure 1.21.2: Equiareal rose diagram displaying the directions of the dip of the long axes of 100 pebbles collected in a road cut through a drumlin that trended S 74°W. The numbers of pebbles counted in the different class intervals, clockwise starting from due north, are: 4 8 2 5 12 8 6 4 2 2 1 9 9 10 6 7 3 2

to the hypothesised model, i.e. to Raleigh's model envisaging random directions of dip, we double the observed angles and apply Raleigh's test to the doubled angles. By doubling the angles, the 18 original class intervals are replaced by nine new intervals such that the data received in each new interval originates from two original intervals separated by 180°. So the two opposite modes in the original data would thereby be replaced by a single mode in the derived data. It turns out that

$$q^2 = 877.09, \text{ for } n = 100, \text{ where } \exp(-q^2/n) = 0.0002. \qquad (1.21.5)$$

As the rounding is close to zero, the mental correlate of the q^2 value given at (1.21.5) is situated at approximately (0.9998, 0.0002*) in Raleigh's test distribution. This favours Krumbein's model over the hypothesised model. Moreover, the vector mean of the doubled angles, 177.3°, transforms to S 89°W on the original scale, compared to the drumlin trend, S 74°W, and the average drumlin trend, S 82°W.

Example 1.21.5

Table 1.21.1 gives the numbers of babies with harelip born month by month during 1951 in Birmingham, England. Raleigh's test can be applied by transforming dates to angles as shown in the table. The datum to be modelled as a value of Q^2 is $q^2 = 782$. As the rounding is close to zero, the co-ordinates of the modelled datum are given approximately by (0.999, 0.001*) where a tentative pointer indicates an increase in the number of babies born with harelip during February onward through May, owing to which the hypothesised model fits the data very poorly. The vector mean, 67.1°, corresponds to about 68

Table 1.21.1: Numbers of babies born with harelip in Birmingham, England, 1951

Month	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
Number	8	19	11	12	16	8	7	5	8	3	8	8
Angle	0°	30°	60°	90°	120°	150°	180°	210°	240°	270°	300°	330°

days after mid-January, i.e. to round about the 23rd of March. The data used in this example are given by Edwards (1961) who reports that the total numbers of babies born in the different months were more or less constant, showing no relation to the numbers with harelip. So the finding is not explained by an increase in the total numbers born during February onward through May.

Example 1.21.6

The early-morning orientation of 11 locusts with respect to the direction of the rising sun causes an entomologist to recall that under such circumstances certain species of poikilotherms sun themselves. Taking the direction of the sun to be 0°, the entomologist finds the facing directions of the animals to be

51°, 85°, 90°, 94°, 111°, 214°, 222°, 260°, 281°, 302° and 315°.

As some animals would be sunning their left sides and others would be sunning their right sides, a bimodal grouping would arise. So the entomologist applies Raleigh's test to the doubled angles. The datum to be modelled as a value of Q^2 is $q^2 = 33.2$. This q^2 value is co-ordinated at approximately (0.951, 0.001, 0.048*) in Raleigh's test distribution. The vector mean of the doubled angles, 179°, transforms to 89.5° on the original scale, where 90° would be perpendicular to the direction of the sun's rays.

Discussion of Examples 1.21.4, 1.21.5 and 1.21.6

Mathematical statistics can find very little difference between the first two examples. Substantive science, however, will judge them very differently. In Krumbein's example the co-ordination test has provided a fact of poor fit, which adds strongly to the force of an understandable train of explanatory reasoning, so much so that his interpretation of the data is essentially placed beyond reasonable contest. In Edwards' example, a similar test provides a similar fact of poor fit, but where is the train of explanatory reasoning to which it might add, or from which it might subtract? The data pattern is not explained by fluctuations in the total numbers of babies born. It is known that the frequency of babies born with harelip increases with the age of the mother, and also that it varies between families, but as neither of these two facts would seem to explain Edwards' data, we can but ask for further investigation. For instance, did the pattern recur in subsequent years? As for the example with locusts, substantive science would judge it to be much closer to Krumbein's example than to Edwards' example. In order to grasp this we need only to note that substantive science has contributed drumlin trends for comparison to the vector mean of the pebble directions. It has also contributed the perpendicular to the direction of the sun's rays for comparison to the vector mean of the locusts' directions. But what has it contributed for comparison to the vector mean of the babies' days of birth? Such understanding is utterly

destroyed by the normative rules. Consider, for instance, the specification $\alpha = 0.01$, often favoured as a conventional means of avoiding reference to the given data. The normative rules would destroy substantive understanding by mismatching Examples 1.21.4 and 1.21.5 as very similar cases, and viewing Example 1.21.6 as dissimilar to the other two.

1.22 Combining independent levels of co-ordination

Definition 1.19.2 requires an ordering that will lead to co-ordinates that are *observable* under the hypothesised model. Thus $C(t; 0) = [U(t; 0), \varepsilon(t; 0), V(t; 0)]$ denotes one of the values in the range of a *statistic* denoted by $C(T; 0) = [U(T; 0), \varepsilon(T; 0), V(T; 0)]$. Similarly, Kempthorne and Folks (1971, p. 223) want $SL(t: 0)$ to be *observable* for any hypothesised model. They therefore note that, for instance, the Behrens-Fisher test is not a significance test. Hence, our Theorem 1.20.2 cannot be used to derive a co-ordination test from the Behrens-Fisher test. We refer to the range of $C(T: 0)$ as *the attainable co-ordinations* for the test it represents. This being understood, Theorem 1.22.1 now states a fundamental property of co-ordination tests.

Theorem 1.22.1:

Consider the statistic $C(T; 0) = [U(T; 0), \varepsilon(T; 0), V(T; 0)]$ arising from any test of co-ordination and let $C(t; 0) = [U(t; 0), \varepsilon(t; 0), V(t; 0)]$ denote any one of the attainable co-ordinations that comprise the range of $C(T; 0)$, where $t = 1, 2, 3, \ldots$ Then:

$$\Pr[U(T; 0) < U(t; 0) \mid M_0] = U(t; 0),$$
$$\Pr[\varepsilon(T; 0) = \varepsilon(t; 0) \mid M_0] = \varepsilon(t; 0), \text{ and}$$
$$\Pr[V(T; 0) < V(t; 0) \mid M_0] = V(t; 0).$$

Proof of Theorem 1.22.1

It follows directly from Definition 1.19.2 that:

$$\Pr[U(T; 0) < U(t; 0) \mid M_0] = \Pr(\text{a sample pattern} \in O_T \text{ for } T < t \mid M_0) = U(t; 0).$$
$$\Pr[\varepsilon(T; 0) = \varepsilon(t; 0) \mid M_0] = \Pr(\text{a sample pattern} \in O_T \text{ for } T = t \mid M_0) = \varepsilon(t; 0).$$
$$\Pr[V(T; 0) < V(t; 0) \mid M_0] = \Pr(\text{a sample pattern} \in O_T \text{ for } T > t \mid M_0) = V(t; 0).$$

Q.e.d.

The theorem yields a method for combining the results of independent co-ordination tests, perhaps performed on very different kinds of data, but in which the hypothesised models are tested against alternatives that represent a common substantive source of possibly poor fit. The method, whose basic idea is due to R. A. Fisher, uses the fact that chi-square based on just two degrees of freedom, χ^2_2, has the property

$$-2\ln[1-\Pr(\chi^2_2 < \text{chi}^2_2)] = \text{chi}^2_2, \text{ or (equivalently) } -2\ln\Pr(\chi^2_2 > \text{chi}^2_2) = \text{chi}^2_2.$$

So, given $[U(t; 0), \varepsilon(t; 0), V(t; 0)] = (U, \varepsilon, V)$, we compute the pair of values given by

$$-2\ln[1-U] \text{ and } -2\ln[1-(U+\varepsilon)], \text{ or (equivalently) } -2\ln[\varepsilon+V] \text{ and } -2\ln[V],$$

and we model the pair as the boundaries of a class interval whose midpoint is a value of χ^2_2, as measured approximately on a rounding grid. For example:

(U, ε, V)	Lower boundary	Upper boundary	Midpoint
(0.80, 0.06, 0.14)	3.22	3.93	chi^2_2= 3.575
(0.92, 0.02, 0.06)	5.05	5.63	chi^2_2= 5.340
(0.88, 0.03, 0.09)	4.41	4.82	chi^2_2= 4.615
			Total: chi^2_6= 13.530

We interpret chi^2_6 as the value of a chi-square random variable based on six degrees of freedom, χ^2_6, as measured on a rounding grid, as follows: let the rounding errors of the three chi^2_2 components be modelled as independent and uniformly distributed. Then the variance of the rounding error of their sum is modelled as being given by

$$[(3.93-3.22)^2/12]+[(5.63-5.05)^2/12]+[(4.82-4.41)^2/12] = (2 \times 0.50)^2/12.$$

So, modelling the rounding error for the chi^2_6 value as also uniformly distributed, the observed value, 13.530, is modelled as the midpoint of a class interval bounded by 13.53-0.50 = 13.03 from below, and 13.53+0.50 = 14.03 from above. Hence, under the hypothesised model, the given chi^2_6 value is co-ordinated at (0.96, 0.01, 0.03) in the χ^2_6 distribution.

The foregoing theorem and application have not involved the pointer; we have formulated a mathematical theorem and noted a mathematical consequence thereof. In order to apply the resulting recipe to a practical instance of substantive investigation, we must ensure that the pointers involved have all been aligned as pointing to the left, or as pointing to the right, possibly by inverting one or more of the orderings. This could only be done by understanding what alternatives would arise from the common underlying source of possibly poor fit being tested for. It is after all precisely this kind of consideration that motivated us to introduce the pointer in the first place.

The problem solved by the foregoing method for combining several statistical co-ordinates when testing for a common underlying source of poor fit is of practical interest. However, the reason for presenting it here is to underscore the following: the co-ordinates produced by a co-ordination test *direct* us to where, in the particular test distribution involved, the mental correlate of a given test datum is to be found. They are *directions* in much the same sense as those given to us by a friend who telephones us to arrange a meeting 'in the coffee bar on the corner of 3rd Avenue and 7th Street'. Of course, the coffee bar is in the real world, whereas the rounding is in the human mind. Nevertheless, directions to these places have this in common:

They are not interpretable as probabilities.
They are not subject to normative prescriptions.
They employ meanings that need to be grasped by the human body.

1.23 The sensitivity of a given test to different alternatives

Consider the distributions displayed in Figures 1.7.1 and 1.8.1. The first is bimodal and the second is unimodal. However, that does not contribute to any co-ordination tests that might be based upon them. The entire contribution made by such distributions to such tests is taken up by statistics of the type denoted by C(T; 0) in the previous section, as any shape that the test distribution might otherwise have then falls away. For a given test of co-ordination, the distribution of C(T; 0), i.e. of the random co-ordinate arising under the hypothesised model, is in effect given by Theorem 1.22.1. In the case of the test distribution given in Table 1.7.3, for instance, the distribution of C(T; 0) is given in Table 1.23.1.

Table 1.23.1: A distribution of hypothesised co-ordinates

Attainable co-ordinates	Modelled frequency
(Ø, 0.26, 0.74)	18/70 = 0.26
(0.26, 0.17, 0.57)	12/70 = 0.17
(0.43, 0.23, 0.34)	2(8)/70 = 0.23
(0.66, 0.17, 0.17)	2(6)/70 = 0.17
(0.83, 0.06, 0.11)	4/70 = 0.06
(0.89, 0.11, Ø)	4(2)/70 = 0.11

In such a table the right-hand column is redundant, as each frequency given in the right-hand column is also the corresponding rounding given in the left-hand column. In this section we consider the distribution of the random co-ordinates arising under possible alternatives to the hypothesised model.

Let us recall that although commencement tests do not involve fully specified alternatives, they nevertheless do involve alternatives with certain broadly specifiable statistical properties. In the case of the sunbirds of Example 1.11.2, for instance, their territorial behaviour does not readily lend itself to statistical modelling. Nevertheless, their behaviour would generate an over-dispersed distribution. Again, in the case of Ronald Fisher's analysis of Gregor Mendel's data, it would be silly to try to model the enormously complicated alternatives that Fisher had in mind. Nevertheless, we can describe those alternatives as the consequences of prejudiced data collection, and we can understand how that leads to chi-square values that are unduly small. Once again, when testing a pseudo-random number generator for excessively many runs, we need not be able to produce fully specified alternatives in order to be able to specify the kind of data patterns that, if unduly frequent, would serve to specify such runs. So we take the position that statistical co-ordinates calculated under a hypothesised model also have a distribution under an alternative, albeit only broadly envisaged. However, in order to come to grips with the matter mathematically, we must devise an example with a fully specified alternative. Consider four animals occupying five compartments. Let the hypothesised model, M_0, be that of random occupancy, and let an alternative model, M_1, be that of an aggregative occupancy, whereby each animal would

twice avoid an unoccupied compartment, and then randomly occupy one of the other compartments. The distributions of the possible occupancy patterns are given in Table 1.23.2.

Table 1.23.2: Alternative distributions of occupancy patterns

Pattern	Pr(pattern | M_0)	Pr(pattern | M_1)
$2^{[1]}1^{[2]}0^{[2]}$	0.576	0.100
$1^{[4]}0^{[1]}$	0.192	Ø
$3^{[1]}1^{[1]}0^{[3]}$	0.128	0.414
$2^{[2]}0^{[3]}$	0.096	0.270
$4^{[1]}0^{[4]}$	0.008	0.216

As we have already mentioned, a co-ordination test involves such distributions only via the statistical co-ordinates that arise from them. So, consider the co-ordinate distributions arising in the present case. They are given in Table 1.23.3, where C(T; 0) and C(T; 1) denote the random co-ordinates that arise under M_0 and M_1, respectively.

Table 1.23.3: Distributions of alternative random co-ordinates

Data	Ordering	C(T: 0)	C(T; 1)
$2^{[1]}1^{[2]}0^{[2]}$	O_1	(Ø, 0.57, 0.43)	(Ø, 0.10, 0.90)
$1^{[4]}0^{[1]}$	O_2	(0.57, 0.19, 0.24)	(0.10, Ø, 0.90)
$3^{[1]}1^{[1]}0^{[3]}$	O_3	(0.76, 0.13, 0.11)	(0.10, 0.41, 0.49)
$2^{[2]}0^{[3]}$	O_4	(0.89, 0.10, 0.01)	(0.51, 0.27, 0.22)
$4^{[1]}0^{[4]}$	O_5	(0.99, 0.01, Ø)	(0.78, 0.22, Ø)

In this table the ordering is based on the hypothesised model only, by using the principle that 'the more frequent under the model, the more "like" the model'. Note again, that the frequency of a co-ordinate is given by its rounding; so, a list of all of the co-ordinates attainable under any given model fully specifies the distribution of the random co-ordinate under that model.

Now consider how the information given in Tables 1.23.2 and 1.23.3 might be used to test the quality of fit of M_0 vs. M_1, as alternative explanations of how a given datum of occupancy, say $2^{[2]}0^{[3]}$, might have come about. The distributions in Table 1.23.2 are displayed by bar diagrams in Figure 1.23.1, where the shaded bars represent the roundings within which the mental correlate of the given datum is placed by M_0 and by M_1, respectively.

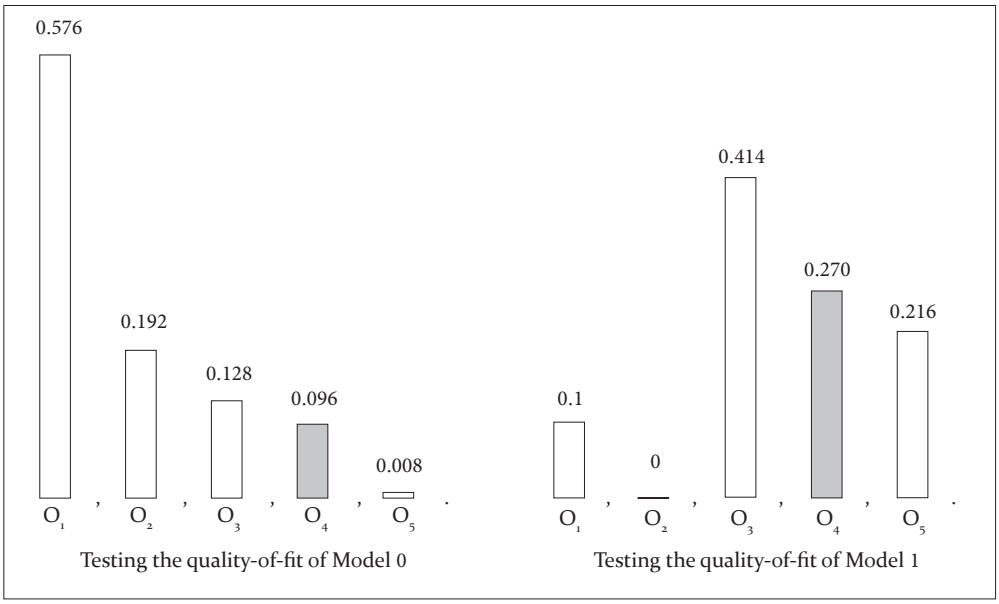


Figure 1.23.1: Testing the quality-of-fit of two alternative models of occupancy behaviour in respect of a given data set

As the meanings conveyed by the alternative displays in Figure 1.23.1 could by simulation be forced upon the human body, those meanings belong to the discourse of physical evidence, in which we point and say:

> 'See for yourself how the mental correlate of the given datum is not so snugly placed within the crowd that M_0 brings to mind.'
>
> 'See for yourself how the mental correlate of the given datum is more snugly placed within the crowd that M_1 brings to mind.' (1.23.1)

The human body is thus forced to grasp that by the test performed the explanation offered by M_1 fits the data better than does the explanation offered by M_0.

For the test we have just performed, Figure 1.23.1 conveys irrelevant detail, which is stripped away when, using the information given in Table 1.23.3, we display the test as in Figure 1.23.2, showing only the co-ordinates for $2^{[2]}0^{[3]}$, given by M_0 and by M_1, respectively.

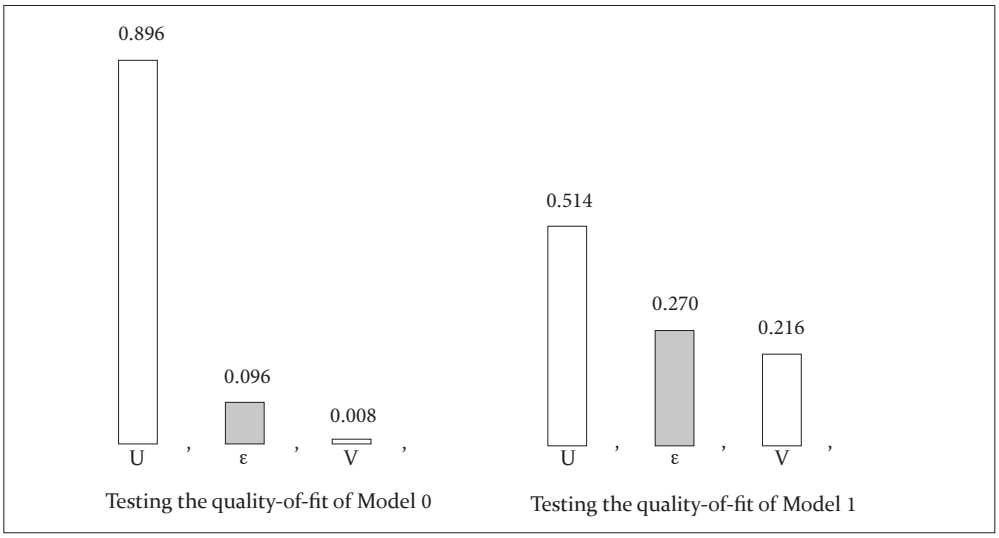


Figure 1.23.2: Statistical co-ordinates displaying the quality-of-fit of two alternative models of occupancy behaviour in respect of a given data set

In Figure 1.23.2 the co-ordinates are now represented by rectangles whose areas differ in proportion to the frequencies they represent. This representation fully serves the purpose of the test, as we can point at Figure 1.23.2 and say *with equal meaning* precisely what we said before when we were pointing at Figure 1.23.1.

Three aspects of the foregoing must now be firmly grasped.

Firstly, we have reasoned in terms of a single instance of occupancies in the real world, and in terms of two alternative populations of occupancies in the human mind. Our reasoning did not envisage, or in any way depend upon, the actual existence, or future existence, of any population of occupancies in the real world.

Secondly, the result of our co-ordination test is not one whose veracity is qualified by probability. The result of our test is a physically perceived fact, as follows:

> As an explanation of how the given data might have come about, the one offered by M_0 does not, by the test performed, fit the data as well as the one offered by M_1.

This fact has been forced upon the human body, and as such is beyond any reasonable contest.

Thirdly, by putting forward an explanation of how the given data might have come about, we have not in any way prevented other explanations from also being put forward.

Having firmly fixed these three ideas in mind, we now wish to consider how we might describe the extent to which a given co-ordination test might enable us to discriminate between alternative models for data in hand. We begin by noting that the ideas introduced by way of Table 1.23.3 motivate Definition 1.23.1.

Definition 1.23.1:

Let O_T for T = 1, 2, 3, ..., k denote a partial ordering of all the sample patterns that arise from one, or the other, or both of a pair of singletons M_0 and M_1, being considered as alternative probability models of how given data might have come about. Let the given data be modelled as if a sample in O_t, where t denotes a particular value of T. A co-ordination test of M_0 versus M_1 attaches to the given data a pair of calculable ordered number triplets

$$C(t, j) = [U(t, j), \varepsilon(t, j), V(t, j)], \text{ for } j = 0, 1,$$

given by:

$$U(t, j) = \Pr(\text{a sample pattern} \in O_T \text{ for } T < t \mid M_j),$$

$$\varepsilon(t, j) = \Pr(\text{a sample pattern} \in O_T \text{ for } T = t \mid M_j), \text{ and}$$

$$V(t, j) = \Pr(\text{a sample pattern} \in O_T \text{ for } T > t \mid M_j),$$

whose members are called the statistical co-ordinates for the given data with regard to the models M_0 and M_1, respectively, and for the partial ordering chosen. We call C(t, 0) and C(t, 1) the *hypothesised co-ordination* and the *alternative co-ordination*, respectively, and we call C(T, 0) and C(T, 1) the *hypothesised co-ordinator* and the *alternative co-ordinator*, respectively.

The reader should note that nothing in this definition prevents one from re-labelling (M_0, M_1) as (M_1, M_0). So, the last sentence of the definition deliberately introduces a terminological asymmetry. This provides for seamless extensions of our development to cases in which such asymmetry is required. In the case of the sunbirds of Example 1.10.2, for instance, the hypothesised co-ordination is hypothesised, not because it is the more credible, but because it is calculable, whereas the alternative, though capable of being broadly envisaged, is not calculable. We have already seen that such is the usual situation at commencement. Theorem 1.23.1 is mathematically useful; it arises directly from Definition 1.23.1.

Theorem 1.23.1:

Let C(T, j) = [U(T, j), ε(T, j), V(T, j)] for j = 0 as opposed to j = 1, denote the co-ordinator pair arising from a co-ordination test of a hypothesised singleton M_0, as opposed to an alternative singleton M_1.

Let T = 1, 2, 3, ..., k be the full range of T. Then the following recurrence relations are satisfied:

$$\varepsilon(t, j) + V(t, j) = V(t - 1, j), \text{ for } j = 0, 1, \text{ and for all of } t = 2, 3, 4, ..., k$$

$$U(t, j) + \varepsilon(t, j) = U(t + 1, j), \text{ for } j = 0, 1, \text{ and for all of } t = 1, 2, 3, ..., k - 1.$$

We digress briefly in order to develop a general terminology. The term *test distribution* names the distribution of a *test statistic*. Let the statistic be denoted by T, where O_T for T = 1, 2, 3, ..., denotes an ordering of the sample patterns arising from a hypothesised singleton M_0 being considered as a probability model of how given data might have come about. The terms *test distribution* and *central test distribution* are synonymous. The term *non-central test distribution* names the distribution of T under an alternative singleton M_1. In place of the *number labels* T = 1, 2, 3, ... , we often use equivalent *value labels* T = t_1, t_2, t_3, ... , indicating how the ordering of the sample patterns was achieved. In that case we also use terms such as *central* T and *non-central* T when it is clear what value T refers to. Examples such as central χ^2 and non-central χ^2 come to mind. We note that the use of the pointer can be extended to any alternative co-ordination. For instance, referring back to Table 1.23.3, and again supposing the datum in hand to be $2^{[2]}0^{[3]}$, the facts stated at (1.23.1) are described by

$$C(T, 0) = (0.89, 0.10, 0.01^*) \text{ vs. } C(T, 1) = (0.51, 0.27, 0.22^*) \qquad (1.23.2)$$

When the pointer is attached to either a left or a right co-ordinate, we refer to that co-ordinate as *the pointing co-ordinate*. At (1.23.2) for instance, the right statistical co-ordinates are *the pointing co-ordinates*. We note that in this particular example the hypothesised co-ordinates place the mental correlate of the given datum on the *right-hand* outskirts of the central test distribution, whereas the alternative co-ordinates place the correlate well within the bulk of the non-central test distribution. We express this by saying that the test is *right sensitive* to the alternative in question.

In trying to develop the foregoing ideas into a mathematically tractable theory, we encounter several fundamental difficulties, as follows.

A first fundamental difficulty is that there is a multiplicity of ways in which a co-ordination test might be made to discriminate between a hypothesised model and an alternative. Recall for instance how the following five co-ordinations arising from the use of Snedecor's *F*, as circumstantially described in Example 1.11.3, were found to be pointing at inflated error estimates:

(*0.07, 0.93), (*0.02, 0.98), (*0.04, 0.96), (*0.00, 1.00), (*0.11, 0.89).

Each of the five *F* ratios can be expressed in terms of Student's *t* as $F = t^2$, as each *F* arose from a test for 2 × 2 factorial interaction. So we might instead have considered the co-ordinates of the corresponding *t* values in Student's distribution, as follows, where the pointers are pointing at persistent mediocrity of calculated *t* values:

(0.54*, *0.46), (0.51*, *0.49), (0.48*, *0.52), (0.50*, *0.50), (0.56*, *0.44).

As there are many other ways for alternatively indicative patterns to be contrived, we are compelled to agree on Definition 1.23.2 as the definition of sensitivity in general.

Definition 1.23.2:

A co-ordination test of a hypothesised model is *sensitive* to a specific alternative if and only if we can point out data patterns that would be more frequent under the alternative circumstances than under the hypothesised circumstances.

Definition 1.23.2 merely serves to underscore the contrived nature of co-ordination tests. The most satisfactory tests usually turn out to be tests whose orderings place sample patterns typical of an alternative of interest, into a specific tail of the test distribution. Unless stated otherwise, it must be tacitly understood that we will be considering tests that are thus contrived.

A second fundamental difficulty arises because for any mutually exclusive and exhaustive ordering of sample patterns, $O_1, O_2, O_3, \ldots, O_k$, the initial member, O_1, has no left co-ordinate, and the final member, O_k, has no right co-ordinate. Therefore, in the notation of Definition 1.23.1, we must necessarily have

$U(1, j) = \varnothing$ and $V(k, j) = \varnothing$ regardless of whether $j = 0$ or $j = 1$.

This is exemplified in Table 1.23.3, where $k = 5$. Now consider, for instance, just $\underline{n}$ independent attempts to discriminate by taste between two items. Let the hypothesised model be that the taster is simply guessing with equal chances of success or failure at each

attempt. An appropriate ordering for achieving sensitivity to the alternative that the taster has some ability to so discriminate, is O_j for j = 1, 2, 3, ... , where j-1 = the number of successful attempts. The co-ordinates under the hypothesised model for n successes in just n attempts are then given by $(1\text{-}2^{-n}, 2^{-n}, \varnothing^*)$ as, for instance,

$$(0.5, 0.5, \varnothing^*) \text{ if } n = 1, \text{ and } (0.9375, 0.0625, \varnothing^*) \text{ if } n = 4. \qquad (1.23.3)$$

The magnitude of the right co-ordinate is evidentially vacuous in such cases, and they are not so rare in practice that they may be fobbed off. So we must devise a definition of sensitivity that accounts for them.

A third fundamental difficulty arises owing to the variable magnitude of statistical roundings, as exemplified by the following sets of possible co-ordination:

Set 1: (0.90, 0.04, 0.06*), (0.90, 0.05, 0.05*), (0.90, 0.06, 0.04*).
Set 2: (0.89, 0.06, 0.05*), (0.90, 0.05, 0.05*), (0.91, 0.04, 0.05*).

A co-ordination test places the mental correlate of the given datum *anywhere* within the rounding. Hence, by placing the mental correlate at the right-hand 'edge' of each rounding in Set 1, we bisect the test distribution at 0.94:0.06, 0.95:0.05, 0.96:0.04, respectively, which bisections are progressively more extreme from left to right. This progression is reflected by the right co-ordinate values 0.06, 0.05, 0.04, respectively, but not by the left co-ordinate values, 0.90 in each case. Next, by placing the mental correlate at the left-hand 'edge' of each rounding in Set 2, we bisect the distribution at 0.89:0.11, 0.90:0.10, 0.91:0.09, respectively, which bisections also are progressively more extreme from left to right. But this progression is not reflected by the right co-ordinate values, 0.05 in each case; it is reflected by the left co-ordinate values, 0.89, 0.90, 0.91, respectively. The source of this difficulty was noted at the very outset of our development by way of Theorem 1.19.1, where it was underscored that any co-ordination test requires, for its precise expression, the values of two variables (not just one). However, that being understood, we note that the rounding produced by such a test is often small to the extent of having no forceful bearing on the outcome of the test. So, unless stated otherwise, it must be tacitly understood that the main thrust of our development concerns tests such that, in the notation of Definition 1.23.1,

$$U(t, j)+V(t, j) \approx 1 \text{ both for } j = 0 \text{ and for } j = 1.$$

Either the values of U(t, j) for j = 0, 1, or (equivalently) those of V(t, j) for j = 0, 1, can then be used to describe the properties of such a test.

We now revisit our analysis of Krumbein's angular data (Example 1.21.4). We applied Raleigh's test after doubling the observed angles. The rose diagram displayed in Figure 1.23.3 describes the doubled angles. The result of the test is given by:

$$C(t; 0) = (0.9998, \varepsilon, 0.0002^*), \qquad (1.23.4)$$

where here and henceforth the symbol ε in an otherwise numerically expressed co-ordination denotes a near-to-zero rounding.

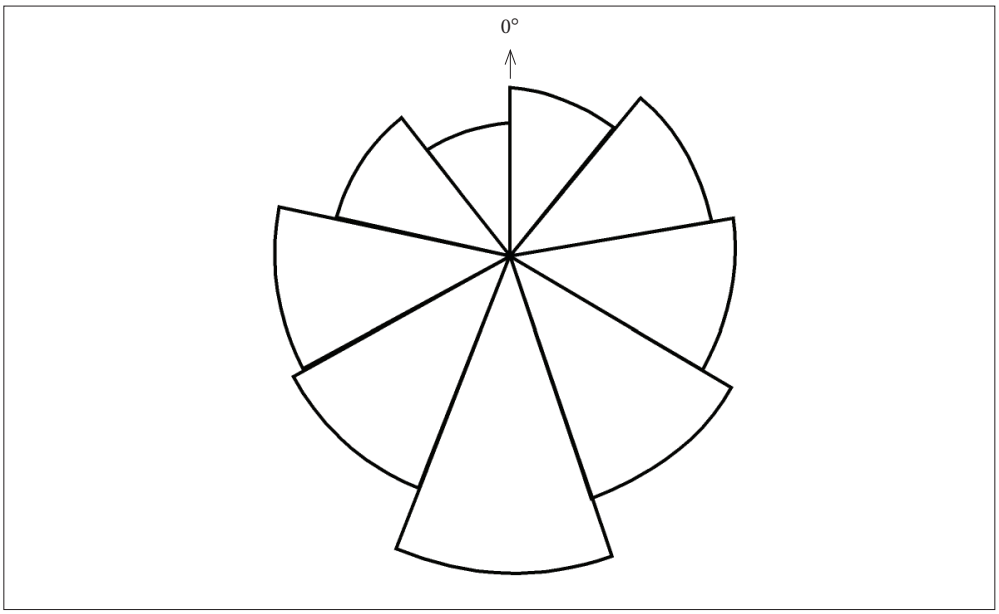


Figure 1.23.3: Equiareal rose diagram of directions obtained by doubling those of the dip of the long axes of 100 pebbles collected in a road cut through a late Wisconsin drumlin. In this diagram the doubled angle that corresponds to the drumlin trend is 182°

For the interpretation of the co-ordinates at (1.23.4) we envisage the hypothesised model, M_0, as an isotropic distribution, and we envisage the alternative, M_1, incipiently as a unimodal distribution whose shape is broadly indicated by Figure 1.23.3. The amplitude of the resultant vector, q^2 as defined at (1.21.3), will in such a case tend to be larger for samples from M_1 than for samples from M_0. For near-to-zero rounding this is described by

$$U(t, 0) > U(t, 1), \text{ or (equivalently) by } V(t, 0) < V(t, 1). \qquad (1.23.5)$$

We will say that such a test is *right sensitive* to M_1. *An ideal right-sensitive test* would be such that the values of

$$[U(t, 0), \varepsilon(t, 0), V(t, 0)] \text{ and } [U(t, 1), \varepsilon(t, 1), V(t, 1)]$$

would be given by

$$(1\text{-}\varepsilon, \varepsilon, \varnothing) \text{ and } (\varnothing, \varepsilon', 1\text{-}\varepsilon'),$$

respectively. Apart from shared sample patterns in near-to-zero roundings denoted by ε and ε', respectively, the non-central test distribution would then be situated entirely to the right of the central test distribution.

Consider now that instead of applying Raleigh's test to the doubled angles, we apply the test directly to the raw angles displayed in Figure 1.21.2. Contributions from class intervals separated by 180° would then tend to cancel, as:

$$\sin x + \sin (x+180°) = 0, \text{ and } \cos x + \cos (x+180°) = 0$$

So, small rather than large values of q^2 in such a case point at Krumbein's alternative. Raleigh's test would then be *left sensitive* rather than *right sensitive* to Krumbein's alternative. For near-to-zero rounding this is described by:

$$U(t, 0) < U(t, 1) \text{ or (equivalently) } V(t, 0) > V(t, 1). \qquad (1.23.6)$$

An ideal left-sensitive test would be such that the values of

$$[U(t, 0), \varepsilon(t, 0), V(t, 0)] \text{ and } [U(t, 1), \varepsilon(t, 1), V(t, 1)]$$

would be given by

$$(\varnothing, \varepsilon, 1\text{-}\varepsilon) \text{ and } (1\text{-}\varepsilon', \varepsilon', \varnothing),$$

respectively. Apart from shared sample patterns in near-to-zero roundings denoted by ε and ε', respectively, the non-central test distribution would then be situated entirely to the left of the central test distribution.

We note in passing that in order to compare right sensitivity achieved by one test to left sensitivity achieved by another test, one of the orderings has to be inverted. For instance, when applying Raleigh's test directly to the raw angles shown in Figure 1.21.2, we obtain the co-ordinates (*0.1371, ε, 0.8629). If we invert the ordering used in this test for comparison to the co-ordinates obtained when applying Raleigh's test to the doubled angles shown in Figure 1.23.3, the co-ordinations obtained are

$$(0.9998, \varepsilon, 0.0002^*) \text{ for the doubled angles, and}$$
$$(0.8629, \varepsilon, 0.1371^*) \text{ for the raw angles.} \qquad (1.23.7)$$

The test based on the raw angles is evidently exceedingly insensitive compared to the test based on the doubled angles. This was to be expected, as a realistic model for the original data would have to involve nine independent binomial partitions of pebbles received in each of the nine pairs of opposing class intervals, respectively. This source of irrelevant variation is removed when the angles are doubled.

Proceeding from ideas introduced at (1.23.5) and (1.23.6) we take account of cases such as those exhibited at (1.23.3) by introducing a definition of sensitivity such that when T = 1, 2, 3, ..., k, gives the full range of T, sensitivity at T = k requires

$$\varepsilon(k, 1) > \varepsilon(k, 0) \text{ because } V(k, 1) \text{ and } V(k, 0) \text{ are both given by } \varnothing,$$

and sensitivity at T = 1 requires

$$\varepsilon(1, 1) > \varepsilon(1, 0) \text{ because } U(1, 1) \text{ and } U(1, 0) \text{ are both given by } \varnothing.$$

This is accomplished by way of Definition 1.23.3 overleaf.

Definition 1.23.3:

Let C(T, j) = [U(T, j), ε(T, j), V(T, j)] for j = 0 as opposed to j = 1, denote the co-ordinator pair arising from a co-ordination test of a hypothesised singleton M_0, as opposed to alternative singleton M_1. Let

T = 1, 2, 3, ..., k give the full range of T.

If ε(t, 1) ≥ ε(t, 0) and V(t, 1) ≥ V(t, 0) with at least one inequality sharp, we say the test is *right sensitive* to M_1 at the C(t, 0) level of co-ordination, and if this is the case for every one of t = 1, 2, 3, ..., k, we say the test is *invariably* right sensitive to M_1.

If ε(t, 1) ≥ ε(t, 0) and U(t, 1) ≥ U(t, 0) with at least one inequality sharp, we say the test is *left sensitive* to M_1 at the C(t, 0) level of co-ordination, and if this is the case for every one of t = 1, 2, 3, ..., k, we say the test is *invariably* left sensitive to M_1.

In many cases of practical interest, the rounding is too small to have a forceful bearing on the interpretation of any realised co-ordinates. In such cases the simpler Definition 1.23.4 may be used.

Definition 1.23.4:

Let C(T, j) = [U(T, j), ε(T, j), V(T, j)] for j = 0 as opposed to j = 1, denote the co-ordinator pair arising from a co-ordination test of a hypothesised singleton M_0, as opposed to an alternative singleton M_1. Let

T = 1, 2, 3, ..., k give the full range of T.

If V(t, 1) > V(t, 0), we say the test is *right sensitive* to M_1 at the C(t, 0) level of co-ordination, and if this is the case for every one of t = 1, 2, 3, ..., k – 1, we say the test is *invariably* right sensitive to M_1.

If U(t, 1) > U(t, 0), we say the test is *left sensitive* to M_1 at the C(t, 0) level of co-ordination, and if this is the case for every one of t = 2, 3, 4, ..., k, we say the test is *invariably* left sensitive to M_1.

Note that the explanations leading to Definitions 1.23.3 and 1.22.4 make it clear that the definitions are limited to the extent of trying to pick off various important cases rather than trying to take account of every mathematical possibility. Consider Definition 1.23.5.

Definition 1.23.5:

Let C(T, j) = [U(T, j), ε(T, j), V(T, j)] for j = 0 as opposed to j = 1, denote the co-ordinator pair arising from a co-ordination test of a hypothesised singleton M_0, as opposed to an alternative singleton M_1. Let

T = 1, 2, 3, ..., k give the full range of T. We refer to the mapping:

V(t, 0) → V(t, 1) for t = 1, 2, 3, ..., k-1

as *the right sensitivity function of the test*, and we refer to the mapping:

U(t, 0) → U(t, 1) for t = 2, 3, 4, ..., k

as *the left sensitivity function of the test*.

The test displayed in Table 1.23. 3 is right sensitive. So, the appropriate function for describing its sensitivity is V(t, 0) → V(t, 1) for t = k-1, k-2, k-3, ..., 1, which (as k = 5) is given by:

(0.01 → 0.22), (0.11 → 0.49), (0.24 → 0.90), (0.43 → 0.90).

This mapping is depicted in Figure 1.23.4.

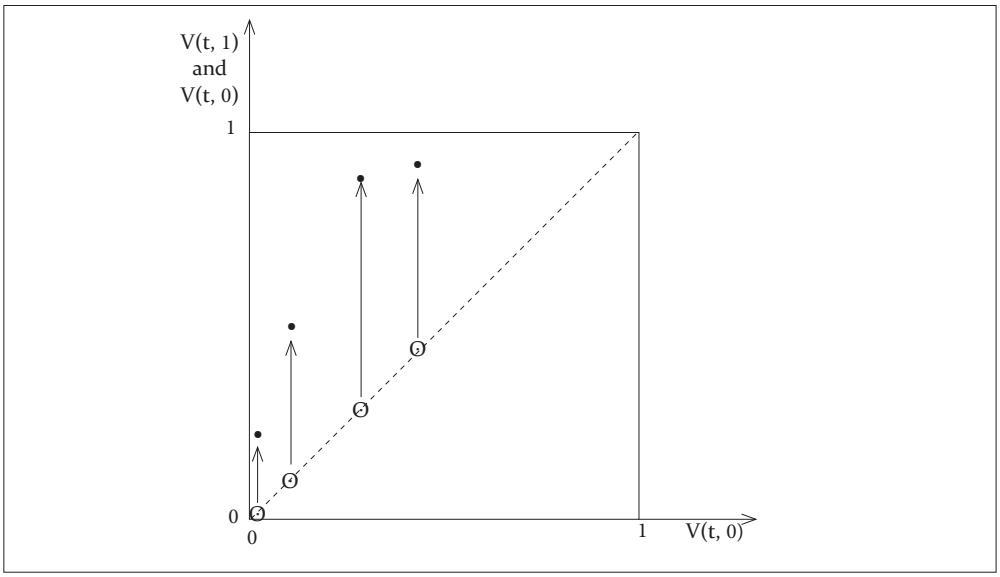


Figure 1.23.4: The sensitivity function of a right-sensitive test. The arrows in the diagram depict the mapping of hypothesised right co-ordinates onto alternative right-co-ordinates

In the sense of Definition 1.23.4 the present test is invariably right sensitive to the alternative involved. If, for such a test, all the roundings throughout the entire ranges of C(T, 0) and C(T, 1) were to be near to zero, a mapping of the type depicted in Figure 1.23.5 would arise.

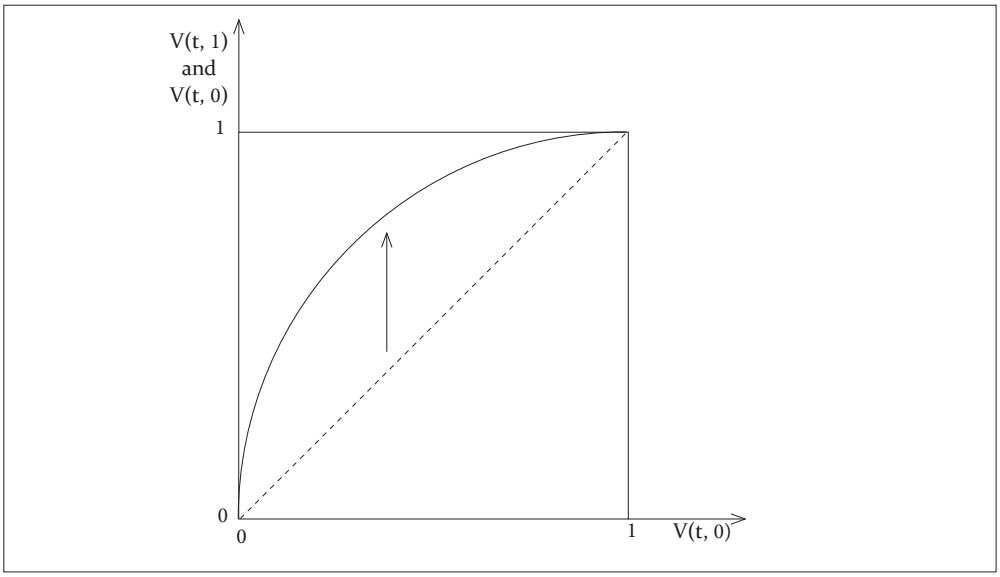


Figure 1.23.5: The sensitivity function of a right-sensitive test in a near-continuous case

We now give two concrete examples that display the practical value of the concepts introduced in this section.

Example 1.23.1

In horse racing around an eight-horse circular track, each horse is assigned to one of the eight possible post positions in the starting line-up. Position 1 is closest to the inside rail of the track, followed by Positions 2, 3, 4, and so on up to Position 8 on the outside rail, furthest from the inside rail. It is widely thought that of any two horses, the one whose post position is closer to the inside rail has an advantage. Table 1.23.4 gives a data set adapted for the purposes of this example from an actual data set. We wish to test the

Table 1.23.4: Numbers of winning horses in 50 races from each of the eight possible starting post positions at a particular eight-horse circular track

Post position	1	2	3	4	5	6	7	8
Number of winners	9	7	6	8	7	4	5	4

hypothesised model, M_0, that a winning horse's post position is just a random number from 1, 2, 3, ..., 8. We also wish the test to be sensitive to the alternative, M_1: a winning horse's post position tends to be smaller than just a random number from 1, 2, 3, ..., 8. Now if X denotes a randomly sampled one of the numbers 1, 2, 3, ..., 8, its expected value and variance are given by

$$E(X) = 4.5 \text{ and } V(X) = 5.25,$$

respectively. Consider a partial ordering of sample patterns based on the magnitude of the sample mean, $\overline{X}$. Let $\underline{n}$ denote the sample size. Owing to the central limit theorem the distribution of the following quantity is approximately standard normal:

$$\frac{\overline{X}\text{-}E(\overline{X}).}{\sqrt{V(X)\div n}}$$

Our test datum is the mean of the n = 50 winning post positions given in Table 1.23.4, which is given by $\bar{x} = 3.98$. We find that

$$\frac{3.98\text{-}4.5}{\sqrt{5.25\div 50}} = -1.605,$$

which is a class mark for the interval between -1.543 and -1.667. The mental correlate of the test datum is thus found to be situated at approximately (0.05, 0.01, 0.94) in the test distribution. In order to test

M_0: 'A post position closer to the inside rail entails no advantage'
vs.
M_1: 'A post position closer to the inside rail entails an advantage',

we must attach the pointer to the left co-ordinate, thus obtaining (*0.05, 0.01, 0.94), which points quite strongly away from M_0 at M_1. In order to make the nature of our test

entirely clear, let us consider that under M_1, a random post position might have expectation E(X) = 4.5-0.75, with variance more or less the same as under M_0. We then find that

$$\frac{3.98-(4.5-0.75)}{\sqrt{5.25\div 50}} = -0.710,$$

which is a class mark for the interval between -0.648 and -0.772. So, again using the normal approximation, we find that the mental correlate of our test datum is situated approximately at the (*0.22, 0.04, 0.74) level of co-ordination in the non-central test distribution. Should the meanings conveyed by the co-ordinates here given as

(*0.05, 0.01, 0.94) and (*0.22, 0.04, 0.74) (1.23.8)

respectively, be questioned, those meanings can be forced upon the human body by the use of diagrams of the type displayed in Figure 1.23.1, or (better) Figure 1.23.2. We note in passing that this example nicely exemplifies the purpose of data analysis, which is to enlarge the investigator's supply of relevant facts. The example proceeds from the fact that the distance between the starting gate and finishing line is shorter for a post position closer to the inside rail. So the investigator is of the prior opinion that a post position closer to the inside rail entails an advantage, i.e. that E(X) < 4.5. Our statistical analysis now produces, at for instance (1.23.8), further relevant facts, which (in the present case) strengthen the grounds for the investigator's opinion.

Example 1.23.2

As a further example of the practical value of the concepts introduced in this section, Table 1.23.5 draws on certain results Shapiro and Wilk (1965) obtained by simulating the performance of several tests for normality against several alternatives. The symbol ε

Table 1.23.5: Alternative co-ordinations achieved by four commencement tests of normality in respect of four different alternatives to the hypothesised model for samples of size n = 20 when the given datum co-ordinates at (0.95, ε, 0.05) in the test distribution. Right-sensitive ordering has been employed throughout

	Shapiro-Wilk W	Pearson χ^2	Pearson $\sqrt{b_1}$	Pearson b_2
M_0: Normal	(0.95, ε_{10}, 0.05)	(0.95, ε_{20}, 0.05)	(0.95, ε_{30}, 0.05)	(0.95, ε_{40}, 0.05)
M_1: Logistic	(0.92, ε_{11}, 0.08)	(0.94, ε_{21}, 0.06)	(0.88, ε_{31}, 0.12)	(0.94, ε_{41}, 0.06)
M_2: Uniform	(0.78, ε_{12}, 0.23)	(0.89, ε_{22}, 0.11)	(1.00, ε_{32}, 0.00)	(0.71, ε_{42}, 0.29)
M_3: Cauchy	(0.12, ε_{13}, 0.88)	(0.59, ε_{23}, 0.41)	(0.23, ε_{33}, 0.77)	(0.19, ε_{43}, 0.81)
M_4: Log-normal	(0.07, ε_{14}, 0.93)	(0.05, ε_{24}, 0.95)	(0.11, ε_{34}, 0.89)	(0.42, ε_{44}, 0.58)

serves as a reminder of non-zero rounding. In order to construct the table all the orderings were made to be right sensitive. For instance, the natural ordering of the W statistic was inverted, as small values of W point at the alternatives. The second row in Table 1.23.5 shows that of the four tests considered, none achieves useful sensitivity to the logistic. That was to be expected, as the logistic is so close to normal that it is often used as a more convenient stand-in for the normal. Pearson's $\sqrt{b_1}$ is insensitive to M_2, which was

to be expected, as $\sqrt{b_1}$ tests for skewness. Somewhat unexpectedly, $\sqrt{b_1}$ turns out to be sensitive to M_3. In retrospect, however, the result is understandable because, as Shapiro and Wilk point out, the Cauchy distribution, although symmetric, has extremely high tails. So, with sample size as small as 20, it often happens that a small number of extreme values contribute unequally to the tails of the sample, and the $\sqrt{b_1}$ test, as used by Shapiro and Wilk, does not discriminate between skewness to the right and skewness to the left. We note that these remarks have, in effect, introduced Definition 1.23.6, involving the same sort of simplification that leads from Definition 1.23.3 to Definition 1.23.4.

Definition 1.23.6:

Let $C(T, j) = [U(T, j), \varepsilon(T, j), V(T, j)]$ for $j = 0, 1, 2$ denote the co-ordinator triplets arising from a particular ordering for a co-ordination test of a hypothesised singleton, denoted by M_j for $j = 0$, against one or the other of two alternative singletons, denoted by M_j for $j = 1, 2$, respectively. At the $C(t, 0)$ level of co-ordination, the test is *more right sensitive* to M_1 than it is to M_2 if:

$$V(t, 1) > V(t, 2),$$

and is *more left sensitive* to M_1, than it is to M_2, if:

$$U(t, 1) > U(t, 2).$$

Definition 1.23.6 can obviously be extended to embrace further possibilities, such as the test being *invariably* more sensitive to one alternative than to another.

In making use of Table 1.23.5 we may compare entries in the same column or in the same row. The two types of comparison are fundamentally different, as follows: When we compare entries in the same column, we consider a particular data pattern whose mental correlate recurs in every rounding in that column, as the same ordering of sample patterns is repeatedly involved. For entries in different columns, that is not the case. So, apart from the passing remark on the comparative sensitivity of the two Raleigh tests applied to Krumbein's data, this section avoided comparisons involving different tests, that is to say, avoided comparisons involving different orderings of sample patterns.

1.24 The sensitivity of different tests to a given alternative

In the previous section certain difficulties had to be overcome in trying to develop a mathematically tractable way of describing the sensitivities of a given test to various alternatives. In the present section we wish to extend the previous development, so as to be able to describe the sensitivities of various tests to a given alternative. A further difficulty must then be overcome because, as pointed out by Kempthorne and Folks (1971, p. 317), it is fundamentally difficult to make meaningful comparisons between tests that do not have the same set of attainable levels. Consider, for instance, the derivation of Table 1.23.3 from Table 1.23.2. Instead of ordering the sample patterns according to their frequencies under M_0, we can order the patterns according to their frequencies

under M_0 *relative* to their frequencies under M_1. This so-called likelihood ratio ordering is displayed in Table 1.24.1.

Table 1.24.1: A co-ordination test based on a likelihood ratio ordering

Data pattern	O_T	C(T; 0)	C(T: 1)
$1^{[4]}0^{[1]}$	O_1	(Ø, 0.19, 0.81)	(Ø, Ø, 1.00)
$2^{[1]}1^{[2]}0^{[2]}$	O_2	(0.19, 0.57, 0.24)	(Ø, 0.10, 0.90)
$2^{[2]}0^{[3]}$	O_3	(0.76, 0.10, 0.14)	(0.10, 0.27, 0.63)
$3^{[1]}1^{[1]}0^{[3]}$	O_4	(0.86, 0.13, 0.01)	(0.37, 0.41, 0.22)
$4^{[1]}0^{[4]}$	O_5	(0.99, 0.01, Ø)	(0.78, 0.22, Ø)

Clearly, the orderings in Tables 1.23.3 and 1.24.1 are both right sensitive to M_1. Equally clearly, we cannot *readily* decide which test is the more sensitive to M_1. When comparing the two Raleigh tests at (1.23.7) we could ignore this difficulty, as the distribution of 100 pebbles between nine or eighteen class intervals generates many attainable levels of hypothesised co-ordination. In general, however, Definition 1.24.1 may not be ignored.

Definition 1.24.1:

The sensitivities of two different co-ordination tests are *strictly comparable* only at those levels of hypothesised co-ordination that are attainable by both tests.

For example, when pooling two consecutive members of a given ordering, we obtain an ordering that is strictly comparable to the original only at levels not involved by the pooling. We note, however, that Definition 1.24.1 does not forbid us to consider certain tests to be *approximately comparable*, provided we can show that the approximation is not misleading.

Let (U, ε, V) denote a level of co-ordination *numerically* attainable under a hypothesised singleton, M_0, by each of two different tests, T and T′, both contrived to test M_0 against an alternative singleton, M_1. Here we emphasise the term 'numerical', as two different orderings are involved. In other words, we now consider two different data patterns that in two different orderings, respectively, are co-ordinated at the same numerical level under M_0, or at least approximately so, as is the case for any pair in the first row of Table 1.23.5. Let the two different data patterns under M_1 then be co-ordinated at $(U_1, \varepsilon_1, V_1)$ and at $(U'_1, \varepsilon'_1, V'_1)$ by T and T′, respectively. If ε_1 and ε'_1 are both near to zero, we would consider test T to be the more right sensitive to M_1 if $V_1 > V'_1$, and we would consider test T to be the more left sensitive to M_1 if $U_1 > U'_1$. Consider, for instance, a comparison based on Table 1.23.5 of the sensitivity to M_3, of the χ^2 test and the $\sqrt{b_1}$ test. The table tells us that if any given data pattern is by the χ^2 ordering placed in ε_{20}, the same data pattern is also by the χ^2 ordering placed in ε_{23}. Similarly, if any given data pattern is by the $\sqrt{b_1}$ ordering placed in ε_{30}, the same data pattern is also by the $\sqrt{b_1}$ ordering placed in ε_{33}. However, the two patterns will almost surely differ. Therefore under M_0, two different sample patterns have by two different orderings, respectively, both been placed at the (0.95, ε, 0.05) numerical level of co-ordination. At the same time, under M_3, the same two patterns have

by the same two orderings been placed at the (0.59, ε, 0.41) and (0.23, ε, 0.77) numerical levels of co-ordination, respectively. So, inasmuch as (0.23, ε, 0.77) is more toward the right than (0.59, ε, 0.41), we consider the $\sqrt{b_1}$ ordering to be more right sensitive to M_3 than the χ^2 ordering, at the (0.95, ε, 0.05) level of hypothesised co-ordination. We note that these remarks have, in effect, introduced Definition 1.24.2, which once again involves the kind of simplification used to obtain Definitions 1.22.4 and 1.22.5. Definition 1.24.2 can obviously be extended to embrace further possibilities, such as the one test being *invariably* more right sensitive or more left sensitive to M_1 than the other.

Definition 1.24.2:

Let (U, ε, V) denote a level of co-ordination that is *numerically* attainable under a hypothesised singleton M_0, by each of two different tests of M_0 against an alternative singleton M_1. Let the co-ordinator triplet for test k (k = 1, 2) under singleton M_j (j = 0, 1) be denoted by $[U_k(T_k, j), \varepsilon_k(T_k, j), V_k(T_k, j)]$. Let:

$[U_k(T_k, 0), \varepsilon_k(T_k, 0), V_k(T_k, 0)] = (U, \varepsilon, V)$ when $T_k = t_k$ (k = 1, 2).

Then, at the (U, ε, V) level of hypothesised co-ordination:

Test 1 is *the more right sensitive* to M_1 if $V_1(t_1, 1) > V_2(t_2, 1)$, and

Test 1 is *the more left sensitive* to M_1 if $U_1(t_1, 1) > U_2(t_2, 1)$.

1.25 The separating characteristics of a co-ordination test

We must distinguish between the *separating characteristics* of co-ordination tests and the *operating characteristics* of hypothesis tests. In the case of hypothesis testing, a decision-maker employs the test as an operational rule whose purpose it is to place into the real world a population of decisions whose constituency in terms of correct and incorrect decisions satisfies certain specifications. Typically the decisions have to do with certain items that might correctly or incorrectly be classified as 'good', or might correctly or incorrectly be classified as 'bad'. The two types of misclassification, i.e. of erroneous decision, are called the Type I and Type II errors, respectively, whose frequencies are specified in some or other way. The specifications cannot (and this is crucial) tell us whether a solitary decision is correct or incorrect; the specifications apply only to the host of decisions as a whole. It is therefore entirely appropriate to speak of those specifications as 'the operating characteristics' of the decision rule, because they are meaningful in an operational sense only. Co-ordination tests concern problems of an entirely different kind. To begin with, any co-ordination test is designed to deal with a particular data set, i.e. with one solitary item in the real world. The problem to be addressed is to try to provide *tenable answers* to the question: 'How *might* (or *might not*) these data have come about?' Here the word 'might' and the plurality of possible answers underscore the *investigative* rather than the *decision-making* nature of the problem. An *investigator* (not a *decision-maker*) brings to mind this, that and the other possible explanation of how the given data *might*, or *might not*, have come about, and co-ordination tests then try to *separate* the more tenable explanations from the less tenable ones. So it is entirely appropriate to say that the previous two sections concern the *separating characteristics* of co-ordination tests. Indeed we need only consider the statements made at (1.23.1) to understand this terminology.

1.26 A symmetric representation

Many test distributions can precisely or with satisfactory approximation be expressed as originating from a partial ordering of the form:

$$\ldots, O_{-3}, O_{-2}, O_{-1}, O_0, O_1, O_2, O_3, \ldots,$$
$$\text{such that } U(-t, 0) = V(t, 0) \text{ for } t = 1, 2, 3, \ldots,$$
$$\text{and where } O_0 \text{ might be an empty set.} \qquad (1.26.1)$$

We refer to this as *a symmetric representation* of the ordering. Consider, for example, the distribution of Raleigh's test statistic, which is approximated by that of a random variable Q^2, whose distribution is defined by:

$$\Pr(Q^2 > q^2) = \exp(-q^2/n) \text{ for } 0 < q^2 < \infty.$$

Suppose $n = 100$. By solving for q^2 from the equations

$$\exp(-q^2/100) = \ldots 0.475, 0.485, 0.495, 0.505, 0.515, 0.525, \ldots, \qquad (1.26.2)$$

and then labelling the solutions ... -3, -2, -1, +1, +2, +3, ... , respectively, we obtain an approximately symmetric representation of the ordering for Raleigh's test, with (in this case) O_0 as the empty set. The grid at (1.26.2) can be replaced by a finer grid, or by a coarser one. As the distribution of Raleigh's test statistic is asymmetric, this example shows that an approximately symmetric representation does not require an initially symmetric distribution. Often, however, initial ordering needs to be near to continuous so as to be capable of a satisfactorily precise symmetric representation.

1.27 Simultaneous statistical inference

We are now ready to take the first steps toward coming to grips with a primitive idea that leads to one of the largest sub-literatures of the statistical literature – an idea known as *simultaneous statistical inference*. It may as well be said at once that this book will prove to be extremely critical of the primitive idea, let alone the procedures inspired by it. However, at this early stage of our development it would be premature to expect the reader to take a firm stance on the matter. Here we can only try to provide a modicum of advance insight into an argument whose complete development will be possible only after we have come to grips with a deeply hidden flaw in R. A. Fisher's theory of significance testing.

Consider the problem of testing a hypothesised singleton, M_0, against one or the other of two alternative singletons, denoted by M_1 and M_2, respectively. Let the ordering for a suitable test statistic have a symmetric representation given by

$$\ldots, O_{-3}, O_{-2}, O_{-1}, O_0, O_1, O_2, O_3, \ldots. \qquad (1.27.1)$$

Suppose the test is invariably left sensitive to M_1, and invariably right sensitive to M_2, and that both alternatives are of interest. Then simultaneous statistical inference would typically have us test for both the alternatives simultaneously, by using the so-called '*two-tailed test*' that arises from the ordering

$$O_0,\ O_{-1} \cup O_1,\ O_{-2} \cup O_2,\ O_{-3} \cup O_3,\ \ldots\ . \qquad (1.27.2)$$

Given the supposed properties of the original ordering at (1.27.1) the new ordering at (1.27.2) would tend to be right sensitive to both M_1 and M_2. However, in most, if not all, cases of practical interest the test based on the two-tailed ordering, as compared to the test based on the original ordering, has inferior separating characteristics. In order to make this clear, we will present two concrete examples, followed by a discussion. Before proceeding, a word of caution: In much of the statistical literature certain well-known probability arguments are used to motivate two-tailed tests and other, more general, recipes for simultaneous statistical inference. *Those arguments have no place at all, in the theory of co-ordination tests.* So, any intrusion of those arguments into the present development can only sow confusion. We will meet and analyse the arguments in subsequent chapters; here they must be kept at bay. However, as the arguments have promoted deeply entrenched habits of thought, they are very difficult to fend off. So we must meticulously recognise the presence of any statistical rounding, as that will serve to remind us that statistical co-ordinates are intended to provide *directions* to certain places brought into the human mind. Such co-ordinates are not at all intended to convey any *probabilities* and certainly not any probabilities of the kind associated with simultaneous statistical inference.

Example 1.27.1

Snedecor and Cochran (1989, p. 87) reproduce a data set giving the mean numbers of florets produced per plot by seven pairs of plots of gladiolus, one plot from each pair planted with high (first-year) corms, the other with low (second-year or older) corms – this presumably in pseudo-randomised pairs. (A corm is an underground propagating stem.) Calculating the differences 'mean for high corm' minus 'mean for low corm' for each of the seven pairs of plots, we obtain

$$-0.1,\ -0.5,\ +0.7,\ -1.0,\ -1.9,\ -2.2,\ -3.4. \qquad (1.27.3)$$

Do these data indicate a difference in the yielding capacity of the two kinds of corm? Consider, as hypothesised model, M_0 such that the whole of the sampling variation is attributable to randomisation only. The expected value of the sample mean difference is then given by $E(\overline{D}) = 0$. We wish to test M_0 against two alternatives, M_1 such that $E(\overline{D}) < 0$, and M_2 such that $E(\overline{D}) > 0$. An appropriate ordering of sample patterns is then according to the value of the sample mean difference. Such an ordering is left sensitive to M_1 and right sensitive to M_2. The mean difference for the data is given by $d = -1.2$. Under M_0 the absolute values of the sample differences are fixed, and their signs are attributable to randomisation only and so altogether 2^7 equally frequent sample patterns arise. The samples that comprise the rounding have the pattern corresponding to the data, i.e. the pattern

0.1	0.5	0.7	1.0	1.9	2.2	3.4
-	-	+	-	-	-	-

(1.27.4)

Sample mean differences less than the data mean difference arise from just four of the 2^7 different sample patterns, these four patterns being as follows:

0.1	0.5	0.7	1.0	1.9	2.2	3.4
+	+	-	-	-	-	-
-	+	-	-	-	-	-
+	-	-	-	-	-	-
-	-	-	-	-	-	-

(1.27.5)

Hence the level of co-ordination at which the mental correlate of the given datum of mean difference is situated in the test distribution, is given by

$$\left\langle \frac{4}{2^7}, \frac{1}{2^7}, 1-\frac{4+1}{2^7} \right\rangle \approx (0.031, 0.008, 0.961). \qquad (1.27.6)$$

Here the left co-ordinate and rounding are small enough to place the mental correlate well out in the left-hand outskirts of the distribution. We can offer a substantively credible alternative as causative explanation. We introduce that alternative by way of the pointer, stating that (*0.031, 0.008, 0.961) points strongly at the older corms being the more productive. This result is a physical fact, as it can be forced upon the human body. We note in passing that one might imagine the investigator commenting on this by saying: 'The result does not surprise me, as a number of anatomical characteristics of the younger corms are known to be indicators of immaturity in certain gladiolus species'. We are thus again reminded that the sole purpose of statistical data analysis is to augment the investigator's supply of relevant physical facts.

Now let us consider how the two-tailed test arises. The ideas of simultaneous statistical inference would typically have us proceed as follows:

If $E(\overline{D}) \leq 0$ is known for sure use a one-tailed test for $E(\overline{D}) = 0$ vs. $E(\overline{D}) < 0$.
If $E(\overline{D}) \geq 0$ is known for sure use a one-tailed test for $E(\overline{D}) = 0$ vs. $E(\overline{D}) > 0$.
If neither of the two certitudes use a two-tailed test for $E(\overline{D}) = 0$ vs. $E(\overline{D}) \neq 0$.

As a substantive investigator is typically reluctant to embrace a certitude, a two-tailed test is often (even routinely) introduced. In the present case that implies that, instead of ordering the sample patterns according to the magnitude of $\overline{D}$, we have to order them according to the magnitude of $|\overline{D}|$. The rounding doubles, as $|\overline{D}| = 1.2$ arises not only with the pattern of signs listed at (1.27.4), but also with the converse pattern, i.e.:

0.1	0.5	0.7	1.0	1.9	2.2	3.4
+	+	-	+	+	+	+

(1.27.7)

The pointing co-ordinate also doubles, as it receives, not only the four sample patterns listed at (1.27.5) but also the four converse patterns, i.e.:

0.1	0.5	0.7	1.0	1.9	2.2	3.4
-	-	+	+	+	+	+
+	-	+	+	+	+	+
-	+	+	+	+	+	+
+	+	+	+	+	+	+

(1.27.8)

For the two-tailed ordering, the pointing co-ordinate is on the right. Thus, the level of co-ordination at which the mental correlate of the datum of absolute mean difference is situated in the derived test distribution, is given by

$$\left\langle 1-\frac{2\times(1+4)}{2^7}, \frac{2\times 1}{2^7}, \frac{2\times 4}{2^7} \right\rangle \approx (0.922, 0.016, 0.062^*). \qquad (1.27.9)$$

But look at what we have done! Out of altogether $2^7 = 128$ different sample patterns, those listed in (1.27.4) and (1.27.5) are the five sample patterns most typical of M_1 and least typical of M_2. Conversely, the five sample patterns listed at (1.27.7) and (1.27.8) are those most typical of M_2 and least typical of M_1. Moreover, we cannot conceive of an alternative such that both $E(\overline{D}) < 0$ and $E(\overline{D}) > 0$ can simultaneously be the case. So, at (1.27.9) we are trying to test for one or the other of two mutually exclusive alternatives, and for each alternative we use an ordering that deliberately confounds the sample patterns most indicative of that alternative, with the sample patterns least indicative of that same alternative. We will take the position that common sense and science must surely hold that for an investigator to thus confound evidential patterns with counter-evidential patterns is wrong.

In subsequent discussion it is convenient to consider the use of Student's *t* for examples of the present kind. So we note that the foregoing results can satisfactorily be approximated as follows: the usual formula for Student's *t* under M_0 is given in obvious notation by

$$t = (\overline{D}-0) \div s_{\overline{D}}.$$

The values of the sample total advance in steps of 0.2, from -8.6 for the first pattern at (1.27.5), to -8.4 for the pattern at (1.27.4), to -8.2 for the pattern that gives the next smallest total. The raw sum of squares of the differences equals 21.76 throughout. So, using Yates's principle to adjust for continuity, the rounding is found as if the sample total advances from -8.5 to -8.3 in the rounding. This gives approximate co-ordinates for the datum *t* under M_0 as (*0.031, 0.005, 0.964) in place of those at (1.27.6). For a two-tailed test the ordering of sample patterns must be according to the value of $|t|$. We note in passing that the term 'two-tailed' refers to the composition of the pointing co-ordinate and rounding, whose constituents are the samples that contribute the tails of the original distribution. In the present case, the samples contributing one tail are those whose patterns are listed at (1.27.4) and (1.27.5), and the samples contributing the other tail are those whose patterns are listed at (1.27.7) and (1.27.8). We also note that the term 'two-tailed' might well be confusing, as the original ordering is sensitive in each of its two tails, whereas the derived ordering is sensitive in the right tail only.

Example 1.27.2

We now revisit Example 1.15.3 in which we considered how to apply Tukey's test for non-additivity to a data set given in Table 1.15.4. We first considered a co-ordination test using Student's *t* as indicated at (1.15.7). Next we considered a co-ordination test corresponding to the customary procedure using Snedecor's *F* as indicated at (1.15.9). We stated that the latter test procedure is wrong. Our reason for that statement is that the values of Snedecor's *F* as used at (1.15.9) label a partial ordering of sample patterns that are equivalently labelled by the corresponding values of Student's $|t|$. Thus the ordering used at (1.15.9) replaces the one used at (1.15.7) with a two-tailed ordering, where that confounds any evidential patterns being tested for with counter-evidential patterns.

Discussion of Examples 1.27.1 and 1.27.2

In order to compare the two different tests performed on the same data set in Example 1.27.1 we must reverse the original ordering. Hence, in Example 1.27.1 the two-tailed ordering replaces

(0.961, 0.008, 0.031*) with (0.922, 0.016, 0.062*).

Similarly, from the results arising at (1.15.7) and (1.15.9) we find that in Example 1.27.2 the two-tailed ordering replaces

(0.952, ε, 0.048*) with (0.904, ε, 0.096*).

In both examples, the original ordering leads to a symmetric central test statistic and, because of that, the values of the pointing co-ordinate and rounding for the two-tailed ordering are exactly twice the values of the pointing co-ordinate and rounding for the original ordering. Tests using Student's $|t|$ are ubiquitous in present-day statistical practice, especially so inasmuch as Snedecor's F is often just a convenient way of using Student's $|t|$ via the relationship $F = |t|^2$. We have argued that it is wrong for a data analyst to use a two-tailed test; in subsequent chapters we show that such tests belong to decision-making under risk. The reader may wish to reserve judgement. It needs to be understood, however, that the argument has very serious implications in terms of costs to substantive science, as follows.

Consider planning the number of replicates required for contrasting the mean yields of two treatments in randomised pairs. Let us specify that should the future trial place the mental correlate of the mean difference at say (0.95, ε, 0.05) in Student's central t distribution, the same correlate must be placed at say (0.20, ε, 0.80) in the non-central t distribution when the observed differences are modelled as independent $N(\delta, \sigma^2)$ random variables for a specified positive value of $\delta \div \sigma$. On second thoughts, however, suppose we are (wrongly) persuaded to use the derived two-tailed co-ordination test based on Student's $|t|$. By how much must the requisite number of replicates then be increased to meet the selfsame specifications? A table of Snedecor and Cochran (1989, p. 104) shows that the requisite proportional increase is given approximately by $7.9 \div 6.2 = 1.27$. So, should it indeed be wrong to use the two-tailed ordering, we would be wasting $0.27 \div 1.27 = 21\%$ of the input of substantive science by way of its contribution in material, salaries, time, plant, etc. This matter cannot be shrugged off as small beer.

The reader might for the time being reserve judgement on the matter discussed in this section, but must bear in mind that when a friend telephones us to arrange a meeting 'in the coffee bar on the corner of 3rd Avenue and 7th Street', the possibility that the coffee bar could have been situated on the corner of 7th Avenue and 3rd Street must not persuade us to proceed to the corner of 10th Avenue and 10th Street. In short:

Directions are not intended to convey probabilities.
Directions do not 'add up'.

1.28 Ordering by modelled frequency – a defective principle

We return to the caution in Section 1.9. The test distribution displayed in Table 1.8.2 is obtained when the ordering of the sample patterns is by their frequencies under the hypothesised model. In Table 1.28.1 a very different test distribution is obtained when the ordering of the sample patterns is by their number of runs under the hypothesised model.

Table 1.28.1: Another test distribution for the number-of-buses problem

Partial ordering	Modelled frequency
O_1: $4^{[1]}5^{[1]}$	2/126 = 0.02
O_2: $1^{[1]}3^{[1]}5^{[1]} \cup 1^{[1]}4^{[2]} \cup 2^{[1]}3^{[1]}4^{[1]} \cup 2^{[2]}5^{[1]}$	7/126 = 0.06
O_3: $1^{[1]}2^{[1]}3^{[2]} \cup 1^{[1]}2^{[2]}4^{[1]} \cup 1^{[2]}3^{[1]}4^{[1]} \cup 2^{[3]}3^{[1]}$	24/126 = 0.19
O_4: $1^{[1]}2^{[4]} \cup 1^{[2]}2^{[2]}3^{[1]} \cup 1^{[3]}2^{[1]}4^{[1]} \cup 1^{[3]}3^{[2]}$	30/126 = 0.24
O_5: $1^{[3]}2^{[3]} \cup 1^{[4]}2^{[1]}3^{[1]}$	36/126 = 0.29
O_6: $1^{[5]}2^{[2]} \cup 1^{[6]}3^{[1]}$	18/126 = 0.14
O_7: $1^{[7]}2^{[1]}$	8/126 = 0.06
O_8: $1^{[9]}$	1/126 = 0.01

In this new table we find that patterns involving an unusually large number of runs are gathered together in one tail of the distribution, while patterns involving an unusually small number of runs are gathered together in the other tail of the distribution. That is as it should be when we can only conceive of the possible physical causes of the one kind of pattern as being very different to the possible physical causes of the other kind of pattern. In that case we would avoid confounding the two different kinds of patterns with each other by way of a two-tailed test. In contrast, the method of ordering by modelled frequency simply gathers together in one of the tails of the test distribution all those patterns that are infrequent under the hypothesised model, regardless of what alternative circumstances might have caused them. Consider, for instance, the co-ordinates of the sample patterns $2^{[2]}5^{[1]}$ and $1^{[9]}$ in the two different test distributions. In Table 1.8.2 these patterns are gathered into the same ordinal class at (0.98, 0.02, ∅) in the test distribution, whereas in Table 1.28.1 they are put into the opposite extremes of the ordering, respectively at (0.02, 0.06, 0.92) and (0.99, 0.01, ∅) in the test distribution. Clearly then, sound principles of ordering must take account of different possible alternatives. And so, for different alternatives that tend to produce different sample patterns, those patterns must be gathered into different tails of the test distribution, or else into one tail each of different test distributions. Thus, if three different alternatives are distinguished by different sample patterns, sound principles of ordering would lead us to tail areas in at least two different test distributions.

1.29 Some formal relations between co-ordination tests and significance tests

Using Definitions 1.19.1 and 1.19.2, we obtain a pair of formal significance levels, as follows:

$$SL_U(T, 0) = U(T, 0)+\varepsilon(T, 0).$$
$$SL_V(T, 0) = \varepsilon(T, 0)+V(T, 0).$$

Kempthorne and Folks (1971, p. 224) note that any significance level has the property:

$$Pr[SL(T, 0) \leq SL(t, 0) \mid M_0] = SL(t, 0). \qquad (1.29.1)$$

For instance, $SL_U(T, 0)$ arises by inverting the ordering used for $SL_V(T, 0)$. So, using Definition 1.19.1 for an inverted ordering, we find that

$$Pr[SL_U(T, 0) \leq SL_U(t, 0) \mid M_0] = Pr(\text{a sample} \in O_T \text{ for } T \leq t \mid M_0)$$
$$= SL_U(t, 0).$$

We can, without loss of generality, restrict further development to the V-like forms, as any results obtained can also be applied to the U-like forms by inverting the ordering. As a reminder of this, we will use the notation SL_V, instead of SL, throughout the rest of this section. The result at (1.29.1) prompts Definitions 1.29.1 and 1.29.2. Note that $SL_V(1, j) \equiv 1$ identically in j, where that explains an aspect of Definition 1.29.2 and of the present development.

Definition 1.29.1:

Let O_T for T = 1, 2, 3, ..., k denote a partial ordering of all the sample patterns that arise from one or the other or both of a pair of singletons M_0 and M_1, being considered as alternative probability models of how given data might have come about. Let the data be modelled as if a sample in O_t, where t denotes a particular value of T. A significance test of M_0 versus M_1 attaches to the given data a pair of calculable numbers, $SL_V(t, 0)$ and $SL_V(t, 1)$, given by:

$$SL_V(t, j) = Pr(\text{a sample pattern} \in O_T \text{ for } T \geq t \mid M_j), \text{ for } j = 0, 1.$$

We call $SL_V(t, 0)$ *the significance level* for the given data with regard to the model M_0 and for the partial ordering chosen, and we call $SL_V(t, 1)$ *the sensitivity level* for the given data with regard to the model M_1 and for the partial ordering chosen.

Definition 1.29.2:

Let $SL_V(T, 0)$ and $SL_V(T, 1)$ denote the significance level and the sensitivity level arising from a significance test of a hypothesised singleton M_0 against an alternative singleton M_1. Let T = 1, 2, 3, ..., k give the full range of T. If $SL_V(t, 1) > SL_V(t, 0)$, we say *the test is sensitive to* M_1 at the $SL_V(t, 0)$ level of significance, and if this is the case for all of t = 2, 3, 4, ..., k, we say *the test is invariably sensitive to* M_1.

The ideas brought forward by Definitions 1.29.1 and 1.29.2 are clearly implicit in the usage of the term 'sensitivity' by Kempthorne and Folks (1971, e.g. p. 236 and p. 317). The same is true of the usage of the term 'sensitiveness' by Fisher (1966, Sections 11 and 12). We draw on that by way of Theorem 1.29.1.

Theorem 1.29.1:

Let O_T for T = 1, 2, 3, ..., k denote a partial ordering of all the sample patterns that arise from one or the other or both of a pair of singletons, M_0 and M_1, being considered as alternative probability models of how given data might have come about. If the ordering produces a significance test of M_0 that is invariably sensitive to M_1, then the same ordering also gives a co-ordination test of M_0 that is invariably sensitive to M_1.

Proof of Theorem 1.29.1

Let $SL_V(t, 0)$ for t = 1, 2, 3, ..., k denote the attainable significance levels, and $SL_V(t, 1)$ for t = 1, 2, 3, ..., k denote the corresponding sensitivity levels. According to Definition 1.29.2, the premise of the theorem then implies that

$$SL_V(t, 1) > SL_V(t, 0) \text{ for all of } t = 2, 3, 4, ..., k. \qquad (1.29.3)$$

(The enumeration starts at t = 2, because $SL_V(1, j) \equiv 1$ identically in j.) It then follows from Definitions 1.23.1 and 1.29.1 that the inequalities at (1.29.3) can be expressed as

$$\varepsilon(t, 1)+V(t, 1) > \varepsilon(t, 0)+V(t, 0) \text{ for all of } t = 2, 3, 4, ..., k.$$

So it follows from the recurrence relations given in Theorem 1.23.1 that

$$V(t\text{-}1, 1) > V(t\text{-}1, 0) \text{ for all } t = 2, 3, 4, ..., k,$$

that is to say, that

$$V(t, 1) > V(t, 0) \text{ for all } t = 1, 2, 3, ..., k - 1.$$

(This last enumeration stops at t = k-1, because $V(k, j) \equiv \varnothing$ identically in j.) *Q.e.d.*

It is not the aim of the present chapter to develop significance tests for comparison to co-ordination tests. We will do so in a subsequent chapter. Presently we merely point out certain *formal* relationships for later use. The reader is yet again cautioned not to jump to premature conclusions about the two kinds of tests.

1.30 Predictions, predications and forecasts

We must come to grips with a precise usage of the terms *prediction*, *predication* and *forecast*. For example, the usual rainfalls of the Western Cape of South Africa are brought on by cold fronts that originate in the South Atlantic Ocean. So, a weather forecast often takes the following form: a photograph taken from a weather satellite is displayed and, pointing at the photograph, the forecaster says: 'As can be seen here, a cold front is approaching. We expect it to bring rain in Cape Town tomorrow.' This relies on a *prediction:* 'cold fronts bring rain', proceeds to a *predication:* 'this is a cold front', and so arrives at a *forecast*: 'this cold front will bring rain in Cape Town tomorrow'. Here it is useful to distinguish between universal *concepts* and particular *individuals*. For instance, 'a woman' *is a universal concept seated in the human mind*, whereas 'Ethel, Jane and Sally' are corresponding *particular individuals seated in the real world*. By drawing these distinctions, the following appears:

> 'Cold fronts bring rain' makes one *concept* (cold front) address another *concept* (rain), it does not refer to any *particular* individual. This conveys *a proposition in theoretical knowledge*.
>
> 'This is a cold front' makes a *concept* (cold front) address a *particular individual* (this one here). This conveys *a proposition in empirical knowledge (experienced knowledge)*.
>
> 'This cold front will bring rain in Cape Town tomorrow' makes a *present* particular individual (this one here) address a *future* particular individual (the one that will be in Cape Town tomorrow). This *conveys a proposition in applied knowledge*.

The concepts are not individuals; they are the *sorts* of corresponding individuals. Also, 'cold front' and 'rain' are *scientific* concepts, as they refer to *bodily* experience. Any concepts are always seated in the human mind. It is only the corresponding particular individuals that are seated in the real world, or that are represented in the data record (the historical record) as having *been* seated in the real world. As noted at the end of Section 1.3, there is for our purposes usually no need to distinguish between any real-world individual, as directly present, or as represented in the data record. Any *future* individuals are, of course, seated in the mind only, but are there being conceived of as part of the real world of the future; we may refer to them as *prospective* individuals, as opposed to *particular* individuals. A particular individual has *its own proprietary*; a prospective individual does not. The proprietary nature of particular individuals either as directly present, or as represented in the historical record, is acknowledged when we identify them by their *proprietary names (proper names)*. Three examples of this are 'Kilimanjaro', 'Nelson Mandela', and 'Krumbein's 1939 data (collected in a road cut through a late Wisconsin drumlin)'. Proper names as such are conventional only, as they then convey identity only. Among Zimbabweans, for instance, Learnmore, Trymore and Wiseman are popular names for a boy child. Again, according to a news report in the *Cape Times* of 25 August 2005, a man named Always Innocent, believed to be of Tanzanian nationality, was to have appeared in the Cape Town Magistrate's Court to face a charge of possession of suspected stolen goods. But he failed to turn up and magistrate Aziz Hamied issued a warrant for his arrest.

The main thrust of this chapter concerns the development of propositions in empirical knowledge, that is to say, propositions in which we point at a data set (at a solitary individual in the real world) and then try to bring the matching *sort* into the human mind. Note, however, that the development might well concern the testing of a proposition in theoretical knowledge. For instance, in order to test the proposition 'cold fronts bring rain' we might consider testing for a positive association between cold fronts and rain, via a 2×2 contingency table. However, such tests (and this is crucial) *involve no forecasting whatsoever*. The following example makes this clear.

Example 1.30.1

Meulepas (1998) considers an example of a 2×2 contingency table with $n = 20$ items classified according to each of two characteristics, as follows:

		Has the item Characteristic 2?	
		Yes	No
Has the item Characteristic 1?	Yes	a = 2	b = 0
	No	c = 0	d = 18

(1.30.1)

Fisher (1970) introduced a classic analysis according to which such data are modelled as a sample of n = 20 independent items, with probabilities respectively as follows:

	Yes	No
Yes	$\theta_1\theta_2$	$\theta_1(1-\theta_2)$
No	$(1-\theta_1)\theta_2$	$(1-\theta_1)(1-\theta_2)$

(1.30.2)

($0 < \theta_1 < 1$; $0 < \theta_2 < 1$). This conceives of no association between row classifications and column classifications, as Fisher's purpose is to test the tenability of this particular aspect of the model. The probability of the sample is then given by the multinomial expression

$$\frac{n!}{a!b!c!d!}(\theta_1\theta_2)^a[\theta_1(1-\theta_2)]^b[(1-\theta_1)\theta_2]^c[(1-\theta_1)(1-\theta_2)]^d.$$

Fisher analyses this into a product of three factors, as follows:

$$\left[\frac{n!}{(a+b)!(c+d)!}\theta_1^{a+b}(1-\theta_1)^{c+d}\right]\times$$
$$\left[\frac{n!}{(a+b)!(c+d)!}\theta_2^{a+c}(1-\theta_2)^{b+d}\right]\times$$
$$\left[\frac{(a+b)!(c+d)!(a+c)!(b+d)}{n!a!b!c!d!}\right]. \qquad (1.30.3)$$

The first factor displays a subsidiary class of binomial models for the row totals; the second factor displays a subsidiary class of binomial models for the column totals. The third factor characterises lack of association, giving the probability of the sample configuration, given the row and column totals, when there is no association between the row and column classifications. For Meulepas' numerical example just three such configurations are possible, with hypothesised probabilities given in Table 1.30.1.

Table 1.30.1: Example for Fisher's exact test for association in a 2×2 table

Configuration 1	Configuration 2	Configuration 3
0 2	1 1	2 0
2 16	1 17	0 18
Probability = $^{153}/_{190}$	Probability = $^{36}/_{190}$	Probability = $^{1}/_{190}$

By advancing from patterns 'less like positive association' to patterns 'more like positive association' we obtain the partial ordering $O_t \sim$ Configuration t (t = 1, 2, 3). Statistical co-ordinates directing us to just where within the test distribution the mental correlate of the given datum of conditional configuration is being placed, are then found to be (189/190, 1/190, ∅). For the pointer, a substantively conceivable alternative must be on

offer. Meulepas envisages the data given at (1.30.1) as arising from n = 20 patients one week after having been admitted to an intensive care unit, as follows:

	Survived	Died
Treated	2	0
Control	0	18

(1.30.4)

That the treatment might increase the frequency of survival, must be considered to be a substantively conceivable alternative, as the medical advisors of the two survivors must have had *prior facts* pointing in that direction. So we attach the pointer to the right co-ordinate, thus obtaining the co-ordination (189/190, 1/190, ∅*) as *a further fact*, and one that turns out to strengthen the case in favour of the treatment, as it points in the same direction as the prior facts.

The reader should note that the test involves a pair of alternative predictions being tested against a given data set, i.e. against a solitary individual in the past, the pair of predictions being:

> M_0: 'Such treatment does not increase the survival rate'
> vs.
> M_1: 'Such treatment increases the survival rate'.

The test, *as such,* does not involve any forecast whatsoever with regard to *prospective* individuals ('those' that may or may not be treated 'in the future'). All that the test produces is a proposition in empirical knowledge (experienced knowledge) stating, as a physically demonstrable fact, that 'The hypothesised model, as tested, matches the given data rather poorly'. This proposition is neither a prediction, nor a forecast; it is *a predication*. It does not differ in epistemological principle from other predications such as 'this is red', 'that is an apple', and 'Nelson Mandela is a man'. In each case we point at a solitary object in the real world and we bring a matching physical sort into the human mind.

We note in passing that the ordering used in the foregoing is incapable of a satisfactory symmetric representation. Such examples present severe difficulties for the ideas of simultaneous statistical inference. Meulepas discusses altogether eight different recipes for complementing Fisher's exact one-tailed significance test for the 2×2 table, with a two-tailed version. One is due to Finney (1948) and defended by Yates (1984). One is due to Irwin (1935). Three are due to Gibbons and Pratt (1975) and discussed by Gibbons (1986). One is due to Cox and Hinkley (1974) and discussed by Cox (1984). One is due to Pike, given in Hill and Pike (1965) and discussed by Hill (1984). Lastly, one is given in Meulepas' paper itself, and intended to be a significance test to counter a randomised two-tailed hypothesis test due to Lehmann (1959, 1986). A randomised one-tailed hypothesis test for 2×2 tables was first proposed by Tocher (1950). We take up this matter in Section 1.34.

Returning to the ideas introduced in this section, we remark that for statistical data analysis to draw the distinction between *predicting* and *forecasting* is especially important, for at least the following three reasons:

A first reason is that the distinction is not drawn in every-day language.

A second reason is that popular notions of a scientist declaring 'Did I not tell you so?' promote habits of thought in which we wrongly tend to think of the evidence favouring a scientific model as a belief in our ability to use that model for forecasting. Yet any data against which a scientific view might be tested, can be seated in the past only (in the data record); it is never seated in the future.

A third reason for distinguishing between prediction and forecasting is that much of the statistical literature would have us (wrongly) present a forecast as if that could be evidence. We will develop this point in Chapter 4. In the meantime the reader may wish to reserve judgement.

1.31 Scientific data analysis

The term 'data analysis' refers to a fundamental principle of scientific investigation, and can be defined on very general grounds, as in Definition 1.31.1.

Definition 1.31.1:

Any question of the kind: 'Does this model fit these data?' must, as far as possible, be analysed into subsidiary questions of that same kind, i.e. into questions of the kind: 'Does this subsidiary model fit these subsidiary data?' If a given model can be analysed into several self-contained subsidiary models, then any test of the given model must preferably be by way of *analytic tests* in the sense of testing each subsidiary model in its own right. That is meant whenever we speak of *data analysis.* (Later we will find that the notion of a subsidiary model must, for the purposes of this definition, be qualified as that of a *relevant* subsidiary model. But, for the time being, that need not concern us.)

It appears at once that for any model, or subsidiary model, we must be able to identify the appropriate data, or subsidiary data, against which it can be tested. That involves a second fundamental principle, namely that a model is a scientific model if, and only if, it predicts conceptual physical experiences corresponding to which actual experiences can then serve as the data against which the model might be tested. We state this principle as Definition 1.31.2.

Definition 1.31.2:

Any model is *a scientific model* if and only if it *predicts*, in the conceptual world of the human mind, conceptual 'physical experiences' for comparison to corresponding real-world physical (bodily) experiences, where such comparisons ultimately constitute the only possible way in which such a model might be tested.

We note in passing that any *prospective* counterparts of the bodily experiences referred to in this definition, then provide for whatever physical rewards a tenable model might be used by way of *forecasting*; but that does not concern us in this chapter.

In previous sections we have already met, in statistical context, the principles defined by the foregoing pair of definitions. At (1.8.1), (1.8.2) and (1.30.3) we saw how certain sampling models can be analysed into subsidiary sampling models. We saw that sampling models are predictive models in the scientific sense of predicting certain conceptual bodily experiences, which experiences we referred to as 'patterns' and 'frequencies'. We also saw how, by comparing a single instance of real-world bodily experience to an ordered ensemble of conceptual bodily experiences, the tenability of a sampling model is

tested. Such tests must be *analytic*, as required by Definition 1.31.1, where this involves a subtlety, as follows: the analyses at (1.8.1), (1.8.3) and (1.30.3) are examples of how certain models and classes of models can be analysed in well-defined mathematical terms. But we have repeatedly found that the alternatives in commencement testing are only broadly envisaged. That does not mean that such alternatives are not capable of being described in meaningful scientific language. It means only that the language of their description is not capable of being expressed in tightly specified mathematical terms. So it has to be understood that *the principle of analysis applies not only to our hypothesised models, but also to any of the alternatives that are brought to mind*. We will now present a number of examples to illustrate this, followed by further discussion. The reader will find that in each of the examples the notions of 'two-tailed tests' and, more generally, of 'simultaneous statistical inference', are *inter alia* rebuffed as being non-analytic notions. That should not come as a surprise, as the notion of 'simultaneous statistical inference' must quite obviously lead to synthesis, rather than to analysis.

Example 1.31.1

Consider Pearson's measure of skewness, $\sqrt{b_1}$, and his measure of kurtosis, b_2, for large samples of size $\underline{n}$. Let

$$Z_1 = (\sqrt{b_1})/\sqrt{(6/n)} \text{ and } Z_2 = (b_2\text{-}3)/\sqrt{(24/n)}.$$

When sampling from a hypothesised normal population, Z_1 and Z_2 are distributed approximately as two independent N(0, 1) variables.

> Z_1 is left sensitive to left skewness and right sensitive to right skewness.
> Z_2 is left sensitive to leptokurtosis and right sensitive to mesokurtosis.

Using different notation, Cox and Hinkley (1974, pp. 71-72) recommend as follows:

> 'If, say, only the symmetry of the distribution is of concern and only departures in one direction are of importance one can use Z_1 directly as a test statistic; if departures in either direction are of approximately equal importance, $|Z_1|$ or equivalently $(Z_1)^2$ can be used.'

A similar recommendation must then of course apply to Z_2, and they make the further recommendation

> 'If both statistics Z_1 and Z_2 are of interest, then some composite function is needed. ... We may for example take $(Z_1)^2+(Z_2)^2$, which has for large $\underline{n}$ approximately a chi-squared distribution with two degrees of freedom.'

Now consider an investigator who has found the following for given data:

> $Z_1 = +1.000$, which co-ordinates at (0.84, ε, 0.16) in the test distribution.
> $Z_2 = +1.645$, which co-ordinates at (0.95, ε, 0.05) in the test distribution. (1.31.1)

Either way the Z_1 test hardly points at skewness. However, the Z_2 test points strongly at mesokurtosis. Suppose, however, that the investigator accepts the recommendations of Cox and Hinkley, not realising that these arise from the doctrine of simultaneous statistical inference. The investigator would then be persuaded that since interest is in either

form of skewness and either form of kurtosis, the tests at (1.31.1) must be disregarded and be replaced by:

$|Z_1| = 1.000$, which co-ordinates at (0.68, ε, 0.32) in the test distribution.
$|Z_2| = 1.645$, which co-ordinates at (0.90, ε, 0.10) in the test distribution.

Moreover, following the further recommendation of Cox and Hinkley, the investigator would be persuaded that these two tests must also be disregarded and replaced by:

$(Z_1)^2+(Z_2)^2 = 3.706$, co-ordinating at (0.83, ε, 0.17) in the test distribution.(1.31.2)

The two tests at (1.31.1) are analytic. The three tests that follow are not analytic. First the sample patterns that are typical of skewness to the left are confounded with those that are typical of skewness to the right. Then the sample patterns that are typical of mesokurtosis are confounded with those that are typical of leptokurtoses. Finally, at (1.31.2), all four types of patterns are confounded with one another, and the result has hidden the evidence displayed at (1.31.1).

Example 1.31.2

Consider the ten waiting times for the early-era eruptions of Vesuvius (Table 1.15.2). We previously found that, as tested by the Cramer-Von Mises ordering, our model of exponential waiting times fits the data well. That test would obviously be sensitive to various non-exponential sample patterns, but it cannot possibly be sensitive to a non-random order of waiting times, as it deals with waiting times in order of magnitude, ignoring order of occurrence. In order to test for non-random order of occurrence, we label the five highest values 'A', and the five lowest values 'B', to obtain the string

AABAABBBBA.

Define $y_x = 1$ or 0, depending on whether the outcome in ordinal position x is A or B, respectively. Let the 10-choose-5 possible sample patterns be ordered by magnitude r, the product moment correlation of x and y_x. Let the hypothesised model be that of a random order of waiting times. For the given data the product moment correlation is r = -0.383, which is a class mark for the interval from r = -0.453 to r = -0.313. Under the hypothesised model, the test statistic is distributed approximately as an $N(0, \sigma^2)$ random variable, where $\sigma^2 = (n-1)^{-1}$ and $\underline{n}$ denotes the number of waiting times (Cf. Example 1.21.1). Using this approximation, the mental correlate of the test datum is to be found at (0.087, 0.092, 0.821) in the test distribution. This kind of test is at its most sensitive if the numbers of A's and B's are as nearly equal as possible. For the 20 waiting times for the later-era eruptions this is achieved by labelling values > 11 as A, and values < 12 as B, to obtain the string

AAABBAABAAAABBBBBBAA,

comprising 11 A's and 9 B's. Using the same type of ordering as before, we find that r = -0.270, which is a class mark for the interval from r = -0.253 to r = -0.288. The mental correlate of the test datum in this case is situated at (0.105, 0.030, 0.865) in the test distribution. Now consider how we might attach the pointers. We see at once that each of the two separate tests is left sensitive to a negative trend in waiting times and right sensitive to a positive trend in waiting times. A data analyst would not wish to confound sample

patterns that point at the one kind of trend with those that point at the opposite trend and therefore two-tailed tests are ruled out. For each era our tests have thus analysed the alternatives into two different possibilities. For the possibility of negative trends, our evidence is very weak:

Early era: (*0.087, 0.092, 0.821). Later era: (*0.105, 0.030, 0.865). (1.31.3)

For the possibility of positive trends, we have no evidence:

Early era: (0.087, 0.092, 0.821*). Later era: (0.105, 0.030, 0.865*).

Finally, now having very little reason to represent the data in terms of a sequence of results, we ignore the sequential order of the waiting times, and use the Cramer-Von Mises statistic to test the quality of fit with which the data could be represented by two Poisson processes. We obtain good fit, as follows (cf. Example 1.15.2):

Early era: (0.85, ε, 0.15*). Later era: (0.87, ε, 0.13*). (1.31.4)

Here the pointer is strictly on the right, as the test is right sensitive to the possibility of a non-exponential pattern of waiting times. We note that for each era, the quality of fit of two distinctly different subsidiary class characteristics have been tested at (1.31.3) and at (1.31.4), respectively.

Example 1.31.3

If the unit vectors $(\sin x_j, \cos x_j)$ for j = 1, 2, 3, ..., n represent a data set of directions, it has been proposed that to test

M_0: 'The directions are a random sample from an isotropic distribution'
vs.
M_1: 'The directions cluster around a preferred direction, α_0',

we might take our test statistic to be the distribution under M_0 of the quantity given by

$$\sum_{j=1}^{j=n} \cos(x_j - \alpha_0). \qquad (1.31.5)$$

We should, however, use this test hesitantly, for the following reasons: vector addition of the given unit vectors yields the resultant vector

$$\sum_{j=1}^{j=n} (\sin x_j, \cos x_j) = q(\sin \tilde{x}, \cos \tilde{x}),$$

whose length and direction are denoted by q and $\tilde{x}$, respectively. The identity

$$\cos(\alpha - \beta) \equiv \sin\alpha \sin\beta + \cos\alpha \cos\beta,$$

then leads to the following expression for the test statistic introduced at (1.31.5):

$$\sum_{j=1}^{j=n} \cos(x_j - \alpha_0) = q \times \cos(\tilde{x} - \alpha_0) \ . \qquad (1.31.6)$$

Hence,

$$\sum_{j=1}^{j=n} \cos(x_j - \tilde{x}) = q. \qquad (1.31.7)$$

A test for isotropic directions using the resultant vector length, q, would simply be an equivalent of Raleigh's test (Raleigh's test as presented at (1.21.3) uses q^2.) Krumbein (1939) introduced the vector mean, $\tilde{x}$, as an estimator of the population mean, say α. Clearly then, a test based on $q \times \cos(\tilde{x}-\alpha_0)$ is not an analytic test; in fact it is even more defective in that the factor q is sensitive to unimodal alternatives by way of unusually *large* values, whereas the factor $\cos(\tilde{x}-\alpha_0)$ is sensitive to $\alpha \neq \alpha_0$ by way of unusually *small* values. An analytic approach must rather separate the two factors, making them the basis of two different tests. A q test is right sensitive to unimodal alternatives, and a $\cos(\tilde{x}-\alpha_0)$ test would be left sensitive to $\alpha \neq \alpha_0$. Note, however, that a co-ordination test for $\alpha \neq \alpha_0$ should rather be based on

$$\text{sign}(\bar{x}-\alpha_0) \times [1-\cos(\bar{x}-\alpha_0)],$$

as that would be left sensitive to $\alpha < \alpha_0$, and right sensitive to $\alpha > \alpha_0$, and would not confound the two different data patterns that underlie these sensitivities. A test based on $\cos(\tilde{x}-\alpha_0)$ would be the two-tailed version of the latter test, where a co-ordination tester will avoid the use of such a two-tailed test as it confounds indicative sample patterns with counter-indicative sample patterns.

Returning now to the test proposed at (1.31.5), let us apply the proposal to Krumbein's data taking α_0 to be the drumlin trend. Using the formula on the right at (1.31.6) together with the information given in Example 1.21.4, we find for the doubled angles

$$q \times \cos(\bar{x}-\alpha_0) = 29.61 \times \cos(177° - 148°), \text{ which equals } 25.90.$$

Cox and Hinkley (1974, p. 67) give the test distribution as approximately N(0, 0.5n) with n = 100 in our case. The test datum, 25.90, co-ordinates at (0.9999, ε, 0.0001*) in the test distribution, pointing at (yes) non-isotropic alternatives, but (no) not at just those whose preferred direction does not deviate substantially from the drumlin trend.

Again, consider a record of the vanishing directions of a number of homing pigeons, when released one by one away from home. Yes! Of course they have flown home! But the home direction might systematically deviate from the vanishing directions, otherwise why have those directions been recorded?

Comparison of the expressions on the left at (1.31.6) and (1.31.7) shows that when $\alpha_0 = \alpha$, Raleigh's tests arises by substituting an estimated value in place of the value estimated. In

that case it can be anticipated that, compared to Raleigh's test, a $q\times\cos(\tilde{x}-\alpha)$ test would have greater sensitivity for almost any, if not any, unimodal alternative, but that is not a tenable criticism of Raleigh's test. Instead, we should recognise that separating characteristics arising within different frames of reference are not *comparable*. In other words, the choice of a test statistic does not begin with a comparison of separating characteristics; it begins with a substantive question.

Example 1.31.4

Snedecor and Cochran (1989, p. 197) reproduce a data set giving the numbers of four distinct types of second-generation plants from a cross between two pure lines in corn. Table 1.31.1 presents these data, along with corresponding expected values that arise, as follows. Using

Table 1.31.1: The numbers of four distinct types of maize plants found in the second generation from a cross between two pure lines. The expected numbers for a 9:3:3:1 hypothesised ratio are given in brackets

	Plain	Striped	Row Totals
Green	a = 773 (731.8)	b = 238 (243.9)	1 011
Gold	c = 231 (243.9)	d = 59 (81.3)	290
Column totals	1 004	1 301	297

the notations introduced at (1.30.1) and (1.30.2) a commonly used Mendelian model arises at (1.30.2) when θ_1 and θ_2 are both = 3/(3+1). In customary genetic terms the model then states that the green:gold segregation ratio θ_1:$(1-\theta_1)$ equals 3:1, that the plain:striped segregation ratio θ_2:$(1-\theta_2)$ equals 3:1, and that the two pairs of genes segregate independently. Together these three statements imply the overall segregation ratio,

$$\theta_1\theta_2:\theta_1(1-\theta_2):(1-\theta_1)\theta_2:(1-\theta_1)(1-\theta_2),$$

equals 9:3:3:1. The expected values given in Table 1.31.1 are those for this 9:3:3:1 ratio. Using the well-known formula:

$$\Sigma[(\text{observed}-\text{expected})^2\div\text{expected}], \qquad (1.31.8)$$

Snedecor and Cochran obtain chi-square = 9.25 on 3 df. (1.31.9)

However, we take the position that this cannot provide an analytical test, as the analysis given at (1.30.3) shows that the 9:3:3:1 ratio arises from three different, self-contained subsidiary models. We begin by fitting the third factor in square brackets at (1.30.3), so as to test for independent segregation. By using a large-sample approximation, the following test statistic can be modelled as an N(0, 1) random variable:

$$Z_1 = (ad-bc)\div\sqrt{[(a+b)(c+d)(a+c)(b+d)\div(a+b+c+d)]}.$$

We obtain

Z_1 = -1.14, which co-ordinates at (0.374, ε, 0.626) in the test distribution.

In order to attach a pointer we would have to know whether the first generation was in linkage phase or in repulsion phase, corresponding to which the pointing co-ordinate would be on the left or the right, respectively. In the present case, however, that does not matter, since either way our hypothesised model of independent segregation, as tested here, fits the data well. Next, to fit the first two factors in square brackets in (1.30.3), and again relying on large-sample approximations, the following two test statistics are modelled as two independent N(0, 1) random variables:

$Z_2 = [(1)(a+b)-(3)(c+d)]\div\sqrt{[(1)(3)(a+b+c+d)]}$.
$Z_3 = [(1)(a+c)-(3)(b+d)]\div\sqrt{[(1)(3)(a+b+c+d)]}$.

Z_2 tests for a 3:1 green:gold segregation ratio, and Z_3 tests for a 3:1 plain:striped segregation ratio. We obtain:

Z_2 = +2.26, which co-ordinates at (0.988, ε, 0.012*) in the test distribution.
Z_3 = +1.81, which co-ordinates at (0.965, ε, 0.035*) in the test distribution.

The pointers arise from the observation that 'striped' and 'gold' represent chlorophyll abnormalities. So, the shortfall in the numbers of such plants is not unexpected, and our analysis produces *further* facts pointing at a lower viability of such plants. Here we have used Z_1, Z_2 and Z_3 to fit three distinctly different, self-contained subsidiary models, each in its own right, as Definition 1.31.1 would have us do. As opposed to that, the chi-square test based on the result at (1.31.9) would have us fit a complex model arising from the three subsidiaries jointly. We note in passing that the sum of squares of the three Z-values, 9.68, differs slightly from the value given at (1.31.9). But that is owing to approximation errors only; i.e. we have for all practical purposes analysed the chi-square statistic arising at (1.31.8) into terms corresponding to three statistically independent degrees of freedom. Clearly, the chi-square test is in this case not analytic. Using the chi-square value at (1.31.9) for a significance test, Snedecor and Cochran find the value too large for the 9:3:3:1 ratio to be tenable (SL = 0.030). They then go on to say that with more than two Mendelian classes, such a test

> '... is usually only a first step in the examination of the data. From the test we have learned that the deviations between observed and expected numbers are too large to be reasonably attributed to sampling fluctuations. But the χ^2 test does not tell us in what way the observed and expected numbers differ. For this, we look first at the individual deviations and their contributions to χ^2.'

So, examining the deviations of observed from expected numbers, they note that these

> '... could be largely explained by a physiological cause, namely the weakened condition of the last three classes due to the chlorophyll abnormality.'

To illustrate this by further analysis, they consider whether the large chi-square value is attributable to poor survivorship of the class with doubly deficient chlorophyll. So they perform two further chi-square tests: (i) a test of whether or not the numbers of plants

in the first three classes reflect a 9:3:3 ratio, and (ii), a test of whether or not the number of plants in the first three classes together and those in the last class reflect a (9+3+3):1 ratio. Again using the recipe given at (1.31.8), they obtain the following:

(i) To test for the 9:3:3 ratio, chi-square = 2.70 on 2 df. SL ≈ 0.25. (1.31.10)

(ii) To test for the (9+3+3):1 ratio, chi-square = 6.53 on 1 df. SL ≈ 0.01. (1.31.11)

They then gather their findings as follows:

> 'To summarize, the high value of χ^2 obtained initially, 9.25 with 3 df, can be ascribed to a deficiency in the number of (doubly deficient) plants, with the other three classes not deviating abnormally from the Mendelian probabilities.'

With regard to the two singly deficient classes, they do, however, remark that:

> '... some deficiencies may also exist in the second and third classes relative to the first class, which would show up more definitely in a larger sample'.

The remark overlooks the following analysis: the 9:3:3 ratio can be viewed as arising independently from a 9:(3+3) ratio and a 3:3 ratio, where only the 9:(3+3) ratio and not the 3:3 ratio is of the type (number normal):(number deficient). Thus the test at (1.31.10) is not analytic. An analytic test for the question it asks is given by the first of the following two test statistics, where the second test statistic gives an analytic test for a distinctly different question:

$$Y_1 = [(3 + 3)(a) - (9)(b + c)] \div \sqrt{[(3 + 3)(9)(a + b + c)]}.$$
$$Y_2 = [(3)(b) - (3)(c)] \div \sqrt{[(3)(3)(b + c)]}.$$

Y_1 tests for a 9:(3 + 3) ratio of normal to singly deficient plants. Y_2 tests for a 3:3 ratio between the two classes of singly deficient plants. We note that these two tests will be conditional with sample totals equal to (a+b+c) and (b+c), respectively; thus each test involves only those plants whose characteristics have a bearing on the question being addressed by that particular test; i.e. each test is appropriately analytic. Under the hypothesised ratios each Y-like statistic is approximately distributed as an N(0, 1) random variable. We find:

$Y_1 = +1.61$, which co-ordinates at (0.946, ε, 0.054*) in the test distribution.
$Y_2 = +0.32$, which co-ordinates at (0.627, ε, 0.373) in the test distribution.

The 1 df chi-square at (1.31.11) is the squared value of

$$[(1)(a+b+c)-(9+3+3)(d)]\div\sqrt{[(3)(3)(a+b+c+d)]} = +2.56,$$

which co-ordinates at (0.995, ε, 005*) in the N(0, 1) distribution.

However, in view of what we have already learned from the tests based on the two Y-like statistics, this last test is of little interest. A more informative test is given by

$$Y_3 = [(1)(b+c)-(3+3)(d)]\div\sqrt{[(1)(3+3)(b+c+d)]}.$$

The rationale for this test is that if the expected shortfall in the observed numbers of deficient plants would be proportionately the same in all three deficient classes, the ratio of singly deficient to doubly deficient plants would remain (3+3):(1). Using the N(0, 1) approximation, we find:

$Y_3 = +2.04$, which co-ordinates at (0.979, ε, 0.021*) in the test distribution.

The Y_1 test shows that the number of singly deficient plants falls short of expectation. The Y_2 test shows no difference in numbers of the two types of singly deficient plants. The Y_3 test shows that the shortfall in the number of doubly deficient plants is greater than the shortfall in the number of singly deficient plants.

Discussion of Examples 1.31.1, 1.31.2, 1.31.3 *and* 1.31.4

Technology is *synthetic*. Investigation is *analytic*. For instance, medical professionals cooperating in the surgical removal of a tumour do so in a *technological capacity*. The question addressed by their reasoning is: 'How might the desired physical outcome *be brought about*?' So, they *synthesise* preparatory procedures, anaesthetic procedures, surgical procedures, and so forth, as an exercise in *the use of knowledge*. But medical professionals cooperating in the diagnosis of an ailment do so in an *investigative capacity*. The question addressed by their reasoning is: 'How might this physical ailment *have come about*?' So, they *analyse* blood samples, urine samples, X-rays, and so on, as an exercise in *the pursuit of knowledge*. Thus the definitive characteristic of investigative science is that it invariably tries to analyse its questions into as many self-contained subsidiary questions as possible. And so it must also be for statistical investigation. In order to achieve this we must recognise the following three binding principles:

> We must try to establish precisely what questions substantive investigation wishes to ask of the data in hand.
>
> We must, as far as possible, try to analyse those questions (and further questions that may thus be raised) into self-contained subsidiary questions.
>
> We must try to deal with the subsidiary questions separately and, for each of the subsidiary questions, try to do so by commencing from a statistic that is minimally sufficient for that particular subsidiary question.

Although these desiderata cannot always be met, we must nevertheless try to do so. Our examples indicate how this may be achieved. Examples 1.31.1 and 1.31.2 show that by asking suitably different questions, the class characteristic is analysed into different subsidiary class characteristics. Example 1.31.3 shows that sample patterns telling us about the class characteristic must not be confounded with sample patterns telling us about the class index. Example 1.31.4 displays a revealing fact: Z_1, Y_1, Y_2 and Y_3 provide four quite differently informative one degree of freedom contrasts, where mathematical statistics can find only three algebraically independent one degree of freedom contrasts. This is because the investigation raised four quite different extra-mathematical questions. In Chapter 6 we will find that there are many such examples whose existence proves that meaningful scientific questions formulated in terms of a given number of mathematical variables may, without redundancy, outnumber those variables.

1.32 Relevant subsidiary models

In order to clarify the parenthetical remark in Definition 1.31.1, we now present four examples, followed by a discussion, further definitions, and a further example.

Example 1.32.1

Consider how an investigator might use the data given in Table 1.1.1 on the half-life of certain fruits, to arrive at an opinion about what values of μ might represent the mean natural ripening time of such fruits. The 0.1 and the 0.9 quartiles of Student's central t distribution for 8 df are –1.397 and +1.397, respectively. So let the equation

$$t = (\bar{x}-\mu) \div \sqrt{\frac{s^2}{n}},$$ where $\bar{x}$ denotes the mean time for the n = 5 control trees,

and s^2 denotes the pooled error variance estimate based on (5-1)+(5-1) = 8 df,

be solved for μ with t = +1.397, and with t = -1.397, respectively. The solutions are

μ = 12.4 days and μ = 13.7 days, respectively.

A co-ordination test of any value of μ between these limits, using Student's t, would result in a test datum whose mental correlate would be situated between (0.1, ε, 0.9) and (0.9, ε, 0.1) in Student's test distribution. Such a value could thereby be judged tenable, provided of course that the two sets of half-life values can be modelled satisfactorily as samples from normal populations that are possibly different in mean only. In Figure 1.1.2 the two sets of responses are plotted against the expected values of the corresponding standard normal-order statistics. For normal sampling each set of points would scatter round a straight-line regression with slope equal to the population standard deviation. Using a table of constants of Shapiro and Wilk (1965), the slope of the regression in the case of the five responses to carbaryl is estimated to be 1.824. For sampling from a *normal population*, this would also estimate the population standard deviation, where that provides the basis for the Shapiro-Wilk test for normality, as the standard deviation of the same responses, 1.062, calculated in the usual way, applies to sampling from *any population*. Here the value of the Shapiro-Wilk statistic, W, is equal to $(1.824 \div 1.062)^2 \div (n-1)$, where n denotes the sample size, five in the present case. The test values for carbaryl and control are W = 0.738 and W = 0.936, respectively, and their mental correlates are to be found within the Shapiro-Wilk test distribution at

(*0.065, ε, 0.935) and (*0.569, ε, 0.431), respectively. (1.32.1)

A lack of 'straightness of scatter' tends to deflate the numerator of W and so the pointing co-ordinates are on the left. However, we must take the position that instead of these two separate tests, our use of Student's t requires a single test, as it relies on a model involving *a compound class characteristic*, whereby the responses to carbaryl and to control were *jointly* modelled in terms of normal sampling. We must therefore employ a single test to judge the quality of fit of that compound. So, using the method of Section 1.22, we combine the results at (1.32.1) to obtain a chi-square value of 6.59 on 4 df. The mental correlate of this value is to be found at

(*0.175, ε, 0.825) in the test distribution. (1.32.2)

Our hypothesised compound class characteristic, here that of joint normal sampling as tested at (1.32.2), fits the given data well.

Example 1.32.2

Consider the eruptions of Vesuvius (Example 1.15.2). The mean waiting times for the two eras were hugely different, namely 155.2 years during the ancient era and 15.65 years during the modern era. Suppose that we must express this difference in terms of a co-ordination test. Let θ_1 and θ_2, n_1 and n_2, and $\overline{X}_1$ and $\overline{X}_2$ denote the expected waiting times, the sample sizes, and the sample means for the ancient and the modern eras, respectively. For the model developed in Example 1.15.2, a minimal sufficient statistic for $(\theta_1 \div \theta_2)$ is given by:

$(\theta_1 \div \theta_2)(n_2 \overline{X}_2 \div n_1 \overline{X}_1)$, which is distributed as Snedecor's *F* on $2n_2$ and $2n_1$ df.

So, in order to test M_0: $(\theta_1 \div \theta_2) = 1$, we calculate the test datum as

$(1)[(10 \times 155.2) \div (20 \times 15.65)] = 4.9$ on 40 and 20 df,

whose mental correlate is situated at (1.00, ε, 0.00) in the test distribution. Here a small left co-ordinate would point at $\theta_1 < \theta_2$, and a small right co-ordinate would point at $\theta_1 > \theta_2$. In the present example the right co-ordinate points overwhelmingly at $\theta_1 > \theta_2$. A commencement test for this development requires that the data for the two eras be *jointly* modelled as the outcome of two independent Poisson processes. So, the two class characteristics that were previously considered separately, and correctly so for the purpose of Example 1.15.2, must for the present purpose be considered to be *the components* of a single *compound characteristic*. The results previously obtained when testing separately for exponential waiting times in Example 1.15.2 were:

For the ancient era: (0.85, ε, 0.15*). For the modern era: (0.87, ε, 0.13*)

Combining these results by the method of Section 1.22, we obtain chi-square = 7.87 on 4 df, whose mental correlate is found at (0.90, ε, 0.10) in the test distribution.

So, for both eras jointly: (0.90, ε, 0.10*). (1.32.3)

As tested at (1.32.3), our hypothesised compound class characteristic, here that of joint exponential sampling, fits the data, though slightly awkwardly.

Example 1.32.3

Recall that in Example 1.31.2 we tested whether or not the two sets of waiting times for the eruptions of Vesuvius can be modelled satisfactorily by two random samples of some or other kind, rather than by two stochastic sequences of some or other kind. In Example 1.31.2 the quality of fit of the two characteristics were tested separately for the two eras, and correctly so for the purpose of that example. There we obtained the following:

Ancient era: (*0.087, 0.092, 0.821). Modern era: (*0.105, 0.030, 0.865) (1.32.4)

However, for the purpose of the *F* test performed in the previous example we must recognise that the two class characteristics *jointly* are the components of a *compound class characteristic*, in which case we must employ a single test to judge the quality of fit of that

compound. So, using the method of Section 1.22 to combine the results given at (1.32.4) we obtain chi-square = 8.42 on 4 df, which is a class mark for the interval from chi-square = 7.25 to chi-square = 9.59.

So, for both eras jointly: (*0.049, 0.057, 0.872). (1.32.5)

Here the magnitude of the rounding is such as to discourage extreme co-ordination. So our hypothesised compound characteristic, here jointly that of a random order as tested at (1.32.5), fits the data, though slightly awkwardly.

Example 1.32.4

In Examples 1.32.2 and 1.32.3 we tested the quality of fit of a class of models for the waiting times for the eruptions of Vesuvius, by way of testing the quality of fit of two entirely different and statistically independent compound class characteristics. That the two are statistical independent follows from the fact that the tests we discussed in Example 1.32.2 are invariant under re-ordering of waiting times within eras, whereas the tests we discussed in Example 1.32.3 rely on, and only on, the consequences of such reordering. It follows that the results at (1.32.3) and (1.32.5) are *mathematically* capable of being combined by the method of Section 1.22. However, to do so would be wrong, as that would run counter to the principle of scientific data analysis expressed by Definition 1.31.1. In fact, the pointing co-ordinates in the two cases point at quite different alternatives. So, the method of Section 1.22 is not available, as it requires the pointing co-ordinates to be identified as those pointing *at the same underlying source of possibly poor fit*, so that they can both be made to be pointing to the right, or both to be pointing to the left.

The foregoing examples show that Definition 1.31.1 requires clarification in that an analytic test can sometimes involve a class characteristic that is 'a compound' in the sense of being capable of analysis into two or more self-contained subsidiaries, but irrelevantly so for the purposes of the test in question. At the same time the examples also show that certain compounds must be ruled out as being synthetic hybrids. Thus for instance, in Example 1.31.1 the test shown at (1.31.2) was ruled out on the grounds that it tries, by way of alternatives, to involve a synthetic hybrid compounded of four entirely different class characteristics. Similarly, in Example 1.32.4, we ruled out the use of the method of Section 1.22 to combine the results at (1.32.3) and (1.32.5), as that would have us involve a synthetic hybrid compounded of two entirely different class characteristics. So, any self-contained components of an admissible compound class characteristic must, in some or other sense, be 'of the same sort'. For instance, Example 1.32.1 involves a compound comprising a pair of mathematically identical components, but Example 1.32.2 shows that our concept 'of the same sort' must also allow for different sample sizes. So the components of an admissible compound will have to be conspecific in some or other sense broader than identical. Again, consider how the termination of a field trial using a completely randomised design might result in inequitably censored treatment groups. A special version of the Shapiro-Wilk test would then be applicable, and we could justifiably use the method of Section 1.22 to combine the results of such tests over different treatment groups. Hence a satisfactory concept of conspecific components must allow for differences in sample size, and for differences arising from censorship or truncation. Perhaps these are the only ways in which the components of a well-conceived compound characteristic might differ. But we avoid making that a strict limitation in Definition 1.32.1. Instead, the phrase 'for

instance' as used in this definition, leaves it up to us to make sensible interpretations of what is meant by 'conspecific components', because that seems to be the best way of avoiding unintended interpretations. Thus, for instance, if a compound consists of nine mathematically identical components, we would not intend that it could be viewed as inadmissible on the grounds that five of the components jointly form a class characteristic 'not of the same sort' as that formed jointly by the other four components. So Definition 1.32.2 arises.

Definition 1.32.1:

A *compound class characteristic* is *an admissible compound* if and only if it comprises *conspecific components*, possibly differing only with respect to for instance different sample sizes, or censorship, or truncation.

Definition 1.32.2:

Let a class characteristic serve as a premise or a partial premise for a further sampling model. Let the characteristic be a compound, thus capable of analysis into several self-contained components. A test of fit of the characteristic can be an analytic test if and only if the compound is an admissible compound, is not a component of a further admissible compound, and serves in its entirety as a premise or a partial premise for the further sampling model. We refer to such a compound as *a relevant subsidiary* of the further sampling model.

Definitions 1.32.1 and 1.32.2 concern the purely conceptual problem of ensuring that a proposed subsidiary model is neither ill conceived, nor irrelevant. The definitions do not concern the empirical problem of *testing* whether or not subsidiary parts of given data are tenably predicated by such a subsidiary model. The following example will help make this clear.

Example 1.32.5

The data set considered in Example 1.32.1 brought into the human mind a compound class characteristic comprising a pair of normal class characteristics, where the mental correlates of the corresponding pair of Shapiro-Wilk test values were situated at

$$(0.065, \varepsilon, 0.935) \text{ and } (0.569, \varepsilon, 0.431) \text{ in the test distribution.} \qquad (1.32.6)$$

The seemingly extreme situation on the left might prompt a heterogeneous incipient alternative to come to mind. A test for such heterogeneity can be based on Theorem 1.22.1, according to which a hypothesised statistical co-ordinate is approximately distributed as a U(0, 1) random variable. An improved approximation replaces U by U+0.5ε, or replaces V by 0.5ε+V (Stone 1969). Let us choose U+0.5ε = Q, so as to transform the information given at (1.32.6) into the following order-statistical values:

$$q_{[1]} = 0.065+0.5\varepsilon \text{ and } q_{[2]} = 0.569+0.5\varepsilon, \text{ where } \varepsilon \approx 0 \text{ in this example.}$$

For samples of size n from a U(0, 1) population, the j^{th} smallest value, $Q_{[j]}$ in present notation, is transformed as follows to Snedecor's *F* (Wilkinson 1933; Ling 1992):

$$\{(1-Q_{[j]})\div Q_{[j]}\}\div\{2(n+1-j)\div 2(j)\} = F \text{ on } 2(n+1-j) \text{ and } 2j \text{ df, for } j = 1, 2, 3, \ldots, n.$$

We find that

> {(1-0.065)÷0.065}÷{2(2+1-1)÷2(1)} = 7.19 on 4 and 2 df, whose mental correlate is found at (0.85, ϵ, 0.15*) in Snedecor's central *F* distribution.

Here the pointer is on the right, as an unduly large *F* value would point at the left co-ordinate on the left at (1.32.6) as being unduly small, which is presently not the case. (Ordering on 0.5ε+V inverts the ordering on U+0.5ε, and so produces the same test.)

1.33 The notion of 'critical regions' as a source of ill-conceived orderings

Let a partial ordering for any co-ordination test or any significance test be denoted by O_1, O_2, O_3, ..., O_{k-2}, O_{k-1}, O_k. The nested regions forming the following array have been termed *the critical regions* for the test (Cox and Hinkley 1974, Section 4.2):

$$W_k = O_k,\ W_{k-1} = O_{k-1} \cup O_k,\ W_{k-2} = O_{k-2} \cup O_{k-1} \cup O_k,\ \ldots .$$

The ensuing development will, however, show that the concept of 'the critical region' is more natural to hypothesis testing than it is to significance testing or co-ordination testing.

Consider a given data set being modelled as the outcome of just $\underline{n}$ independent Bernoulli trials, where the probability of success μ is constant from trial to trial, and $0.5 \leq \mu < 1$. Suppose we wish in some sense to 'test' M_0: $\mu = 0.5$ vs. M_1: $\mu > 0.5$. If $n = 4$ an appropriate partial ordering for a significance test or a co-ordination test is as follows, where S or F denotes success or failure, respectively, and each string of four such letters denotes one of the 2^4 sample patterns that comprise the sample space:

$$\begin{aligned} O_1 &= \{FFFF\} \\ O_2 &= \{SFFF, FSFF, FFSF, FFFS\} \\ O_3 &= \{SSFF, SFSF, SFFS, FSSF, FSFS, FFSS\} \\ O_4 &= \{SSSF, SSFS, SFSS, FSSS\} \\ O_5 &= \{SSSS\}. \end{aligned} \qquad (1.33.1)$$

The attainable significance levels for this ordering are given by:

$$SL_k = \Pr(\text{a sample} \in W_k \,|\, M_0) \text{ for } k = 1, 2, 3, \ldots, 5$$

where each of these can be calculated as a number of favourable cases divided by the total number of cases. For example, $W_3 = O_3 \cup O_4 \cup O_5$. So

$$SL_3 = \Pr(\text{a sample} \in W_3 \,|\, M_0) = [(6 \text{ cases}) + (4 \text{ cases}) + (1 \text{ case})] \div (16 \text{ cases}).$$

However, for a hypothesis test the sample space must be divided into at least two, and at most three, disjoint regions. In the case of just two regions, these will be the critical region, W^r, and its complement, W^a, where *any* region in the sample space may serve as a critical region, and where the corresponding hypothesis test is just a decision rule of the form:

If the sample that models the data is $\in W^r$, reject M_0 and (equivalently) accept M_1.
If the sample that models the data is $\in W^a$, accept M_0 and (equivalently) reject M_1.

The operating characteristics of the decision rule are two *error rates*, these being the frequencies of the two possible kinds of errors (erroneous decisions), as follows:

'Reject $M_0 \mid M_0$' is a Type I error. $\Pr(\text{reject } M_0 \mid M_0) = \alpha$ is the Type I error rate.
'Reject $M_1 \mid M_1$' is a Type II error. $\Pr(\text{reject } M_1 \mid M_1) = \beta$ is the Type II error rate.

Usually the Type I error rate is specified. Consider the specification $\alpha = (0.5)^4$. The critical region may then be chosen to be any one of the 2^4 sample patterns comprising the sample space. The 2^4 different regions involve different Type II error rates, these being, for each region, the frequency of sample patterns that belong to the complement of that region when given M_1. So we can readily calculate that, for instance:

If we choose $W^r = \{SSSS\}$, then $\alpha = (0.5)^4$ and $\beta = 1-\mu^4$.
If we choose $W^r = \{FSSS\}$, then $\alpha = (0.5)^4$ and $\beta = 1-(1-\mu)\mu^3$.
If we choose $W^r = \{FSSF\}$, then $\alpha = (0.5)^4$ and $\beta = 1-(1-\mu)\mu^2(1-\mu)$. Etc.

As $\mu > 0.5$ implies that $1-\mu < 0.5$, it appears at once that for the choice $W^r = \{SSSS\}$ the Type II error rate (the value of β) is less than for any one of the other 2^4-1 possible choices, and uniformly so over all the possible values of $\mu > 0.5$. Such a test is called *a uniformly most powerful hypothesis test* (UMP hypothesis test). Suppose that we now alter the specified value of the Type I error rate for our test to be $\alpha = 2(0.5)^4$. Then, depending on our choice of W^r, there are four different UMP hypothesis tests, each with $\beta = 1-\mu^3$, as follows:

If $W^r = \{SSSS, FSSS\}$, then $\alpha = 2(0.5)^4$ and $\beta = 1-[\mu^4 + (1-\mu)\mu^3]$.
If $W^r = \{SSSS, SFSS\}$, then $\alpha = 2(0.5)^4$ and $\beta = 1-[\mu^4 + \mu(1-\mu)\mu^2]$.
If $W^r = \{SSSS, SSFS\}$, then $\alpha = 2(0.5)^4$ and $\beta = 1-[\mu^4 + \mu^2(1-\mu)\mu]$.
If $W^r = \{SSSS, SSSF\}$, then $\alpha = 2(0.5)^4$ and $\beta = 1-[\mu^4+\mu^3(1-\mu)]$. (1.33.2)

Can any of these UMP hypothesis tests serve our purpose? The answer to the question depends crucially on Definitions 1.2.1 and 1.2.2. On the one hand: should Definition 1.2.1 define our purpose, we are compelled to answer in the negative, because we are then being asked to pick *any arbitrary one* of the four patterns FSSS, SFSS, SSFS, SSSF, and to pretend that it is more indicative of $\mu > 0.5$ than any of the other three would be. Clearly, that amounts to a nonsensical physical 'explanation' in response to the investigative question *How might the value of μ explain how these experiences (data) have come about?* On the other hand: should Definition 1.2.2 define our purpose, we are compelled to answer in the affirmative because then the question to be addressed is: *Depending on the value of μ, how might such experiences be brought about?,* where the experiences in question are a forecasted Type I error rate of $2(0.5)^4$ should μ equal 0.5, and a forecasted Type II error rate of $1-\mu^3$ should μ exceed 0.5, and where any one of the four tests described at (1.33.2) achieve just that. Hence, the language of critical regions and error rates is foreign to investigative statistics; it belongs to a statistical technology of decision-making under predictable risks. Instead of the concept 'critical region', a concept more suitable for significance tests would be that of 'ordinal regions' as provided by Definition 1.33.1. Such regions are always nested in the sense that $W_j \in W_{j-1}$ for j = k, k-1, k-2, ..., 2, and the last member of such an array, W_1, is always the entire sample space. Owing to the formal relationship

between significance tests and co-ordination tests, as given in Theorem 1.20.2, the concept of ordinal regions might well arise in a discussion of co-ordination tests, though for co-ordination tests, *per se*, the concept is redundant.

Definition 1.33.1:

Let $O_1, O_2, O_3, \ldots O_{k-2}, O_{k-1}, O_k$ denote the partial ordering for a significance test. Then *the ordinal regions* for the test are the terms of the array:

$W_k = O_k, W_{k-1} = O_{k-1} \cup O_k, W_{k-2} = O_{k-2} \cup O_{k-1} \cup O_k, \ldots$

Returning to the four UMP hypothesis tests exhibited at (1.33.2) we note that should any ordering for a co-ordination test of M_0: $\mu = 0.5$ vs. M_1: $\mu > 0.5$, assign any of the four patterns FSSS, SFSS, SSFS, SSSF to different ordinal classes, such an ordering would be ill conceived. That is so because in the scientific sense of the term 'test', alternative models can be tested for their tenability by, and only by, testing alternative predictions derived from those models against corresponding data. In order to make this entirely clear we note that at (1.33.2) we are asked *a priori* to choose one of four different tests. Yet in effect we can make the choice *a posteriori,* as follows: let the sample space be partitioned into just three disjoint regions as follows, where X denotes the number of successes:

$$W^r \sim \{X = 4\}.\ W^b \sim \{X = 3\}.\ W^a \sim \{X = 2, 1, 0\}.$$

Let the four sides of a balanced tetrahedral die be marked FSSS, SFSS, SSFS and SSSF, respectively. Then arbitrarily nominate any one of these four patterns, and use the following *randomised decision rule*:

If the sample that models the data is $\in W^r$, reject M_0.
If the sample that models the data is $\in W^b$, roll the die.
    If the die comes to rest on the nominated pattern, reject M_0.
    If the die comes to rest on one of the other three patterns, accept M_0.
If the sample that models the data is $\in W^a$, accept M_0. (1.33.3)

This is an example of *a randomised hypothesis test*. It has exactly the same operating characteristics as those of the four UMP hypothesis tests exhibited at (1.33.2) and so is *a randomised UMP hypothesis test*. The regions W^r, W^b and W^a are often called *the rejection region*, *the boundary region* and *the acceptance region*, respectively. We note in passing that the existence of randomised hypothesis tests should contribute fundamentally to our understanding of the statistical discourse, where, as we will see in Chapters 3 and 4, that has not at all been widely understood. The point here is that hypothesis tests are widely (and mistakenly) seen as part and parcel of data analysis, but then there is almost always extreme reluctance to maximise power, for Type I error rates, as specified by means of randomised tests. On the one hand, such reluctance tacitly admits to *knowing* that there is a defect in the reasoning. On the other hand, to persist with such reasoning, tacitly admits to *not knowing* where the reasoning has gone wrong.

The foregoing has shown that a type of ill-conceived ordering arises when we try to place sample patterns that have the same frequency under both the hypothesised model and the alternative into different ordinal classes. Another type of ill-conceived ordering arises when we consider orderings of the type:

$$O_1 = W^a,\ O_2 = W^b,\ O_3 = W^r. \tag{1.33.4}$$

For instance, let us enlarge our binomial example by taking n = 10 (instead of 4) and specify $\alpha = 0.01$, as is often recommended. Then a randomised UMP hypothesis test requires the following, where X denotes the number of successes, and where a data value of x = 9 would require the use of the auxiliary random device:

$$W^a = \{X < 9\},\ W^b = \{X = 9\},\ W^r = \{X = 10\}. \tag{1.33.5}$$

Considered as an ordering, the defect in this is that the ordering O_{1+X} for X = 0, 1, 2, ..., 10 will often be much more informative, giving for instance the co-ordinates of X = 8 as (0.95, 0.04, 0.01) for $\mu = 0.5$, which is clearly extreme. Opposed to that, the ordering that arises at (1.33.4) from the partitioning at (1.33.5) conceals this extreme by giving the co-ordinates of X = 8 as ($\varnothing$, 0.95, 0.05) for $\mu = 0.5$. So, an ill-conceived ordering arises if we try to place sample patterns that have different frequencies under either the hypothesised model or the alternative, into the same ordinal class.

1.34 Inconceivable orderings

It cannot be overemphasised that science is the discourse of bodily experience, where various forms of mathematical reasoning about science are demonstrably inclined to lose sight of the fact. As a first step toward coming to grips with this very serious defect in some of the literature on our subject, we now re-visit the learned confusion around the idea of a two-tailed significance test for a 2×2 contingency table, as referenced in Section 1.30. We here consider Tocher's one-tailed hypothesis test, and just two of the eight different recipes proposed for a two-tailed significance test. To begin with, we take note of an epistemological usage of the term 'conception'.

Consider the notion of 'a unicorn'. It might be thought that the human mind is able to conceive of a unicorn as being 'a horse-like creature with the tail of a lion, and a long spirally twisted horn growing from its forehead'. However, we take the position that this involves an incorrect usage of the term 'conceive' instead of 'imagine'. We must use conceive as a technical term for describing the process by which a number of bodily experiences make the human mind *conceive* of their *sort*. For instance, imagine being asked to join a panel of tasters who must identify individual instances of 'Pinot Gris wine with a "paraffin-like" taste'. We might protest no experience of Pinot Gris wines, let alone instances of a paraffin-like taste. But if so, the oenologist would explain that we first have to be trained by way of being exposed to samples of Pinot Gris wines with and without the paraffin-like taste. Our minds would be made to *conceive* the requisite *sorts* of bodily experience by our bodies having been exposed to *individual* experiences of those sorts. When we speak of *an inconceivable sort*, we mean that no such sort of bodily experience can be brought to mind.

Now consider Tocher's test. Being a hypothesis test, it relies on the normative prescriptions we met in Section 1.21. As viewed through the spectacles of significance testing, the following rules are prescribed:

(1) Let SL denote the significance level that a one-tailed significance test attaches to the data conveyed by a given 2×2 table, when testing

M_0: 'No association' vs. M_1: 'Positive association'.

(2) Specify a test size, α, *without reference to the data*.

(3) If SL ≤ α, reject M_0 and accept M_1. If SL > α, accept M_0 and reject M_1.

For such a test, Tocher proposed, by way of a hypothesis test, to improve on Fisher's significance test by in effect replacing the natural ordering advocated by Fisher with an artificial ordering, as follows: consider the probability given by the characteristic factor at (1.30.3). It has the form 'a number of favourable cases divided by a number of possible cases'. For Meulepas' numerical example it is convenient to reason as if the number of possible cases equalled 380. (The actual number, given by n!a!b!c!d!, is of course much larger, but that does not affect the reasoning that follows.) The 380 possible cases involve just three configurations whose hypothesised probabilities are then as follows:

Configuration 1	Configuration 2	Configuration 3
0 2	1 1	2 0
2 16	1 17	0 18
Probability = 306/380	Probability = 72/380	Probability = 2/380

For Fisher's significance test the partial ordering is O_t ~ Configuration t for t = 1, 2, 3. Let our specification be α = 0.05. This can be expressed as α = 19/380. So, using the normative prescriptions given above for a hypothesis test, we find that for Meulepas' data:

The observed significance level is given by SL = (2/380)+∅ = 2/380
The specified test size is given by α = 19/380
SL < α
So M_0: 'No association' is rejected, and M_1: 'Positive association' is accepted.

The accept-reject rule would have us reject M_0 if and only if SL ≤ 19/380. So Tocher introduces a device that would in effect have us consider 19/380 to be 'a significance level' attainable by adjoining to 'the two cases favourable for Configuration 3', a further 17 cases poached from 'the 72 cases favourable for Configuration 2'. We are thus to obtain altogether 17+2 = 19 cases that are, so to speak, 'favourable for rejecting M_0'. As seen through the spectacles of a co-ordination tester, this would have us replacing Fisher's partial ordering, i.e.

O_1, O_2, O_3, corresponding to Configurations 1, 2 and 3, respectively,

with O_1, O_{21}, O_{22}, O_3, in that order, where O_{21} and O_{22} are obtained by partitioning O_2 into two disjointed sets of 55 and 17 cases, respectively (55+17 = 72). This is to be done by introducing an extraneous pseudo-random number, x, whose distribution is capable of being modelled as that of a random variable X, such that

Pr(X = 0) = 55/72 and Pr(X = 1) = 17/72.

Consider a bowl containing 72 homogenous chips, 55 marked 0, and 17 marked 1. Then if, for instance, the data in hand had Configuration 2, Tocher's test would have the investigator haphazardly draw one chip to decide whether the given data belong to O_{21}

(should the chip be marked 0) or belong to O_{22} (should the chip be marked 1). In such a case M_0 is accepted or rejected depending on whether the chip is marked 0 or 1, respectively. However, pseudo-random numbers that are subsequently adjoined to given data, cannot possibly contribute any insight whatsoever toward answering the question *How might the given data have come about*? Thus Definition 1.2.1 makes it clear that in the discourse of *the pursuit of knowledge* such numbers must be removed from the given data set. But if those numbers *are* removed, the partitioning of O_2 into the two disjoint sets denoted by O_{21} and O_{22}, respectively, would be asking us to conceive of two inconceivable sorts, because how could the human body then distinguish between

'a Configuration 2 of the O_{21}-like sort' and 'a Configuration 2 of the O_{22}-like sort'?

An investigator's colleagues in substantive science will demand, indeed *must* demand the answer to this question. They must demand to know if the distinction can be tasted, if it can be heard, and if it can be seen. The point here is this: a significance level is by definition a hypothetical long-run frequency; in other words, *its meaning must be capable of being conveyed by simulation*. However, that is possible only if different ordinal sets, such as O_{21} and O_{22}, comprise discernibly different sample patterns.

We note in passing that Chapter 4 will show that Tocher's test originates from reasoning that in the terms of Definitions 1.2.1 and 1.2.2 has mistaken the pursuit of knowledge for the use of knowledge.

We now consider Finney's recipe for a two-tailed test to complement Fisher's one-tailed test for a 2×2 table. In fairness to Finney, we note that his recipe is based on a large-sample approximation given by Fisher (1970, p. 95). Nevertheless, any such approximation requires explication of *what is being approximated*, and on this point Finney's reasoning leads to difficulties. He proposes to calculate the significance level for a two-tailed test as being simply twice that for a one-tailed test. For the present problem, however, the ordering for a one-tailed test is not intrinsically symmetric. So, no matter how large a sample we have in mind, the following difficulties arise. Consider the example given in Table 1.30.1. If we observe Configuration 1, the proposal, when applied either to the ordering used in Example 1.30.1, or to the inverse ordering, will produce a significance level SL, such that SL > 1, where that describes an experience of an inconceivable sort. The difficulty may seem to be avoided by stipulating that the lesser of the two possible one-tailed significance levels must be doubled. However, an insurmountable difficulty then arises, as follows: let the probabilities in Table 1.30.1 be interpreted in terms of the number of 'favourable cases' out of a total of 190 'possible cases'. Suppose that Configuration 3 is observed. Then Finney's recipe would have us calculate SL = 2(1/190). Clearly, that amounts to replacing Fisher's ordering with O_{11}, O_{12}, O_2, O_3, in that order, where O_{11} and O_{12} arise by partitioning O_1 into two disjoint sets comprising one case and 152 cases, respectively (1+152 = 153). It then appears at once that this entails the difficulty of explaining how the human body must establish whether a given Configuration 1 is

'a Configuration 1 of the O_{11}-like sort' or 'a Configuration 1 of the O_{12}-like sort'.

Moreover, whatever device might be introduced in an attempt at a posterior labelling of the given data as being either O_{11}-like or O_{12}-like, will fall foul of Definition 1.2.1 because that definition would compel us to admit that the labelling had nothing to

do with *how the given data came about*. In other words, Definition 1.2.1 would either compel us to explain how O_{11}-like configurations are conceivable as being more 'like' our hypothesised model than O_{12}-like configurations are, or would otherwise compel us to remove the labels as evidentially vacuous for the given investigative problem.

Next we consider the procedure proposed by Meulepas (1998) for a two-tailed significance test in place of a corresponding hypothesis test for 2×2 tables. Meulepas motivates the procedure as one that 'exploits the properties of a two-sided hypothesis test', however, in such a way that there is 'no extraneous quasi-observation such as is needed in the exact randomised uniformly most powerful unbiased (UMPU) test of the hypothesis of independence'. Applying the proposed procedure to the numerical example reproduced at (1.30.4) Meulepas (p. 5) calculates that

$$SL = (9/190)+(1/190)$$

as the proposed value of the significance level for that example, where the method of calculation makes it entirely clear that the procedure must be interpreted as follows in the given instance: there are altogether 190 'possible cases', of which 153, 36, 1, are 'favourable' for Configurations 1, 2, 3, respectively. The significance level is then to be obtained by taking the cases that are, so to speak, 'favourable for the rejection of M_0', as nine of the 153 cases that are 'favourable for Configuration 1', plus the case that is 'favourable for Configuration 3'. So, in effect, we are again dealing with a proposed ordering of the form O_{11}, O_{12}, O_2, O_3, in that order, with O_{11} and O_{12} in this proposal obtained by partitioning O_1 into two disjoint sets comprising nine cases and 144 cases, respectively (9+144 = 153). So we are again faced with the difficulty that, in trying to demonstrate by simulation the bodily meaning of what is being said, we are unable by bodily experience to establish whether a given Configuration 1 is

'a Configuration 1 of the O_{11}-like sort' or 'a Configuration 1 of the O_{12}-like sort'.

We also cannot escape this by resorting to what Meulepas so appropriately refers to as 'extraneous quasi-observations' in an attempt at labelling given results as being either O_{11}-like or O_{12}-like, as that is precisely what Meulepas (quite rightly) would have us avoid. Such 'extraneous quasi-observations' fall foul of Definition 1.2.1 because our fellow investigators will demand, indeed *must* demand, to know how O_{11}-like patterns are more indicative than O_{12}-like patterns of any agency that might have caused an association in the given 2×2 table. 'Surely,' they would be compelled to say, 'these "extraneous quasi-observations" cannot be explained by any such cause, as they have been obtained by for instance drawing from a bowl of chips – *after* the given 2×2 table had *already* come about!'

The direct implication of the foregoing development is that, even if we were to allow for the possibility that 'two-tailed tests' might play some role in statistical data analysis, the recipes of Finney and Meulepas cannot survive scientific scrutiny. That is so because they fail to provide meanings that are capable of being forced upon the human body.

1.35 Developing scientific concepts

The development in the previous section draws attention to the danger of introducing notions not anchored in the discourse of physical (bodily) experience. Three forms of such anchorage were put forth in Section 1.30. One form concerns *predication*. For instance, when we point and say 'This is a cold front', *an experiential concept* (cold front) predicates *a given bodily experience* (this thing here). A second form concerns *forecasting*. For instance, when we then say 'This cold front will bring rain tomorrow', *an experiential concept* (cold front) forecasts *a prospective bodily experience* (rain tomorrow). A third form concerns *prediction*. In the present case, for instance, our forecast relies on the prediction 'cold fronts bring rain', in which *one experiential concept* (cold front) addresses *another experiential concept* (rain). The concepts 'cold front' and 'rain' are both *directly* experiential, as each one can be employed for the predication of bodily experiences. Scientific discourse also involves concepts that are only *indirectly* experiential, in that they cannot be employed for the predication of bodily experiences. Such a concept arose in the early development of genetics. In snapdragons for instance, the numbers of white-, pink-, and red-flowered offspring from the five possible parental combinations of those colours, conform statistically to the multinomial ratios given in Table 1.35.1.

Table 1.35.1: Parental combinations and resulting offspring ratios involving white, pink and red flowering snapdragons

Parental cross	Offspring segregation ratio
White × White	White:Pink:Red::1:0:0
Red × Red	White:Pink:Red::0:0:1
White × Red	White:Pink:Red::0:1:0
Pink × Pink	White:Pink:Red::1:2:1
White × Pink	White:Pink:Red::1:1:0
Pink × Red	White:Pink:Red::0:1:1

By positing the genetic make-up of each individual as comprising two factors – AA if white, Aa if pink, and aa if red – and positing that reproductive cells receive one factor each, these ratios are explained as follows in Mendelian terms:

pink×pink = Aa×Aa and so reproduces as (A+a)×(A+a) = 1AA+2Aa+1aa
white×pink = AA×Aa and so reproduces as (A+A)×(A+a) = 2AA+2Aa
white×red = AA×aa and so reproduces as (A+A)×(a+a) = 4Aa
And so on. (1.35.1)

All the examples in garden peas investigated by Gregor Mendel involved a commonly occurring phenomenon in which one factor is dominant and the other one recessive. In humans for instance, the factor pair for brown eyes (B) and blue eyes (b) results in BB and Bb individuals, both brown eyed, and bb individuals that are blue eyed. Thus B is dominant and b is recessive, so that the Bb×Bb parental combination produces a 3:1 segregation ratio by reproducing as:

(B+b)×(B+b) = 1BB+2Bb+1bb = (1+2)(brown-eyed)+1(blue-eyed).

Mendel's work (published in 1866) went largely unnoticed till its rediscovery in 1900 by, among others, William Bateson, who had then already made similar discoveries in domestic fowl. Bateson made a great contribution to the further development of genetics, as he named the new science. This brings us to a revealing disagreement on the nature of Mendel's 'factors', or 'genes' as they now came to be called. Thomas Hunt Morgan, round about 1910, came to view genes as linearly arranged constituents of cell chromosomes. Bateson rejected this view as overly materialistic, advancing instead his own 'vibratory' notion, founded on ideas of force and motion. Morgan's view prevailed, but that does not concern us here. What does concern us is that the disagreement proves beyond reasonable contest that in the early development of genetics, the concept 'gene' was clearly experiential, but clearly also only *indirectly* so, otherwise there could not have been such a disagreement. For another example, consider the concept named 'energy' by basic physics. Energy can neither be created, nor destroyed; it can only be converted into various different forms. 'Potential energy' is 'stored work'. If we are pumping water up into a storage tank, the potential energy that is chemically stored in our bodies is converted into kinetic energy. Up in the tank, our work is again stored as potential energy. It can then by the mechanical energy of falling water be converted into electrical energy. That energy can again be stored in a battery as potential energy, or it can be converted into heat energy, and so on. These examples show that 'energy' is an *experiential* concept. They also show that the concept is *indirectly* experiential, as we cannot 'predicate energy' by pointing and saying 'See for yourself that "energy" there'. The need for *indirectly* experiential concepts is obvious when we consider that all our ultimate data are sensory data (seeing, smelling, tasting, etc.) by means of which we try to come to grips with the nature of an 'external world' that 'lies beyond'. So, science can but try to understand the 'external world' as it is experienced through the human body. We must be careful not to introduce concepts that are not anchored in bodily experience. This point is especially important, as all of us also conduct extra-scientific discourses involving, for instance, ethical concepts, aesthetic concepts and legal concepts. It follows that a scientist, as such, must demand to know how any proposed concepts are anchored in physical (bodily) experience, and why they are needed above and beyond any concepts already put into scientific place, as required by Definition 1.35.1. This definition holds that science is 'physics' in the broad sense of being the discourse of bodily experience (physical experience). In terms of this definition the concepts of the extra-scientific discourses (ethical concepts, aesthetic concepts, legal concepts, and so on) are *extra-physical* concepts. Some literature, indeed much of the statistical literature, holds that the discourse of science can be advanced by the introduction of concepts that are not extra-physical in the foregoing sense, but that also do not satisfy Definition 1.35.1. Such concepts may thus reasonably be referred to as *metaphysical* concepts. This book holds, as a binding principle, that concepts and rules proposed for statistical discourse must be capable of having their proposed meaning either directly, or indirectly, forced upon the human body, otherwise they must be rebuffed as being incapable of scientific meaning (as being metaphysical). It is a principle that arises from science in general, as underscored by Definition 1.35.2.

Definition 1.35.1:

The discourse of science is concerned with, and only with, the development of predictions, predications and forecasts, all concerning the world as experienced through the human body (experienced *physically*). So the concepts and rules of science are those that are minimally sufficient for that discourse. We refer to those concepts, and to the rules for discourse in terms of those concepts, as *physical* concepts and rules.

Definition 1.35.2:

A scientific concept is *directly experiential* if and only if it can be used to predicate a bodily experience, that is to say, if and only if its meaning is capable of being directly forced upon the human body by pointing and predicating. All other scientific concepts are *indirectly experiential*, being requisite for the discourse of bodily experience.

In retrospect it can now be seen that Definition 1.2.1 leads to the discourse of pointing and predicating, that Definition 1.2.2 leads to the discourse of forecasting, and that we can now complement those definitions by Definition 1.35.3 that leads to the discourse of predicting.

Definition 1.35.3:

'How might *such* and *such* bodily experiences be related to each other?' is the definitive question of *theoretical science*. In science it proclaims the discourse of the explication of knowledge.

1.36 The role of simulation

The term 'simulation', as used in this book, refers to the use of an analogue to explain or study a physical system. Simulation is sometimes resorted to because the system of interest is too complex to be capable of being studied analytically. An example of this is the use of small-scale models of alternative plans for a harbour development in order to simulate complex consequences in respect of oceanic wave action. Similarly, in our own subject we may resort to simulation, in order to obtain the distribution of a statistic whose distribution defies mathematical derivation. Scientific investigation, however, would employ an analytical approach whenever possible. So it must be firmly grasped that simulation also provides such investigation with an important analytical tool. In agricultural research, for instance, experimental design often provides an analogue of some or other subsidiary aspect of a farming system. In fact, in many fields of scientific investigation, analogy is all that is possible. Consider, for instance, investigation of the evolution of ring species, such as the herring gull/lesser black-backed gull ring. In Britain these gulls are clearly distinct species. Not only do they differ in appearance, but they also do not interbreed. Yet, if we trace the population of herring gulls westward around the North Pole, from Britain to Canada, to Alaska, to Siberia, to Europe, and back to Britain, we find interbreeding populations of herring gulls along the way, slowly but surely transforming to lesser black-backed gulls. So why do the two kinds not interbreed in Britain? Biology answers that as we travel from one end of the ring to the other, there is a gradual change in ecological niches occupied by the different populations we come across. So, back again in Britain, the two ends of the ring occupy ecological niches that are different to the extent that any bird that may arise from a herring gull × lesser black-backed gull cross will not be adapted to either one of the two niches. So, natural selection has, in Britain and Europe, resulted in a behavioural barrier that eliminates any tendency to interbreed. Obviously the theory of how such barriers evolved can only be tested by analogy. Consider for instance the snapdragons in Table 1.36.1. Should we maintain through successive generations an open pollinating population of such plants, removing the pink flowering plants in each generation, such a barrier may arise, perhaps by way of pollen incompatibility. Such tests have in fact been conducted. In a well-known test using vinegar flies, two genetically different types of flies stopped interbreeding after 50 generations. The reader should note (and this is the point of the present discussion) that simulation not only provides *explanatory*

reasoning, but it also provides *tests of meaning*, as the acts of simulation are physical acts that have physical consequences. So, if a purportedly 'scientific' assertion has no scientific meaning, we will be unable to simulate that which it asserts. In Section 1.34 such tests of meaning were conducted in order to expose certain metaphysical ideas. For a further example, let us bring into the human mind two continuous U(0, 1) random variables, *U'* and *V'* = 1-*U'*. Then

$$\Pr(U' < u') = u' \text{ and } \Pr(V' < v') = v'.$$

These equations resemble the first and third of the three equations of Theorem 1.22.1, in which sense the hypothesised statistical co-ordinates of any given test statistic are *approximately* those of a U(0, 1) random variable, but never *exactly* so, as an *actual* real-world rounding cannot be of measure zero. When, for instance, we are given the statistical co-ordinates of *U'* = *u'* as

$$\begin{pmatrix} u' & u' & 1 \\ \int dU', & \int dU', & \int dU' \\ 0 & u' & u' \end{pmatrix} = (u'\text{-}0,\ 0,\ 1\text{-}u') = (u',\ 0,\ v'),$$

it is a *conceptual* rounding, a 'rounding in the human mind', that is being made to be of measure zero. Now let the notation of Theorem 1.22.1 be abbreviated as follows:

$$[\mathrm{U}(\mathrm{T}, 0),\ \varepsilon(\mathrm{T}, 0),\ \mathrm{V}(\mathrm{T}, 0)] = (\mathrm{U}, \varepsilon, \mathrm{V}),$$

and consider the approximation

$$(\mathrm{U}, \varepsilon, \mathrm{V}) \approx (U', 0, V').$$

Stone (1969) proposes to improve on this approximation by using instead

$$(\mathrm{U}+0.5\varepsilon,\ 0,\ 0.5\varepsilon+\mathrm{V}) \approx (U', 0, V'). \qquad (1.36.1)$$

The reader may recall that in Example 1.32.5 we used this approximation to provide an *indirectly* experiential random variable. But, as will now be explained, Stone tries to make the approximation provide a *directly* experiential random variable by way of 'a significance level', which is a very different kettle of fish. In the present notation, the true significance level, considered as a statistic, is defined by Definition 1.19.1 as

SL = U+ε when the pointing co-ordinate is on the left, and
SL = ε+V when the pointing co-ordinate is on the right.

Hence, by virtue of Theorem 1.22.1, SL has the property

$$\Pr[\mathrm{SL} \le \mathrm{sl} \mid \mathrm{M}_0] = \mathrm{sl}.$$

A significance tester interprets this property as follows:

> Suppose we were to regard the observed significance level as just decisive against M_0. Then we would be bound to regard any smaller value of the significance level as even more decisive against M_0. Hence the observed significance level, sl, is the

theoretical frequency with which we would mistakenly regard such a significance level as being decisive against M_0.

That this interpretation has *direct* experiential meaning is beyond reasonable contest, because, if needs be, we could by simulation force that meaning onto the human body. However, Stone considers the improved approximation obtained via the recipe given at (1.36.1) as a sound motivation for proposing to re-define the significance level as

U+0.5ε when the pointing co-ordinate is on the left, and
0.5ε+V when the pointing co-ordinate is on the right.

This fails to grasp that such a 'significance level' cannot be interpreted as 'a datum of physical evidence', as the value calculated for it in any given instance cannot have *direct* experiential meaning. This must be firmly grasped. So, consider just five replicate attempts to discriminate by taste between two wines, with four apparently successful (S) attempts, and one apparent failure (F). Let the hypothesised model be the binomial with equal chances of S or F at each attempt, thereby positing M_0: 'No discriminative ability', the alternative being M_1: 'Some discriminative ability'. Taking the number of successes, X, as test statistic, the mental correlate of the test datum, X = 4, is situated within the test distribution at

$$(U, \varepsilon, V) = \left(\frac{32\text{-}5\text{-}1}{32}, \frac{5}{32}, \frac{1}{32} \right). \qquad (1.36.2)$$

The pointing co-ordinate is on the right. In order to simulate the meaning conveyed by this test, we could repeatedly flip five coins of a balanced sort, interpreting heads as S and tails as F. Suppose we then obtain the following:

Replicate 1	Replicate 2	Replicate 3	Replicate 4	Replicate 5	Replicate 6	...
F F F S S	F S S F S	S S S S S	S F F S F	S F S S S	F F S F F	...

Simulation of the co-ordination test at (1.36.2) assigns

Replicates 1, 2, 3, 4, 5, 6, ..., to U, U, V, U, ε, U, ..., respectively.

As the pointing co-ordinate is on the right, simulation of a corresponding significance test, as defined by Definition 1.19.1, assigns

Replicates 1, 2, 3, 4, 5, 6, ... to U, U, ε+V, U, ε+V, U, ..., respectively.

But attempted simulation of Stone's 'significance test' here founders at Replicate 5, as the placing of the mental correlate of the test datum '*in the centre of the rounding*' is incapable of simulation. The problem is this:

Does S F S S S belong to U+0.5ε, or does S F S S S belong to 0.5ε+V?

In other words:

Is S F S S S a 'left-of-the-mental-correlate' sort of sample pattern, or
is S F S S S a 'right-of-the-mental correlate' sort of sample pattern?

If we try to resolve this by flipping an auxiliary coin, we will be asked 'What does the auxiliary simulate? Does it simulate an "after-taste" of some or other sort? If so, why in the first place is there no mention of such an "after-taste" in the data record?' Here again, as in Section 1.34, mathematical reasoning has lost track of reality, and so ends up asking the human mind to conceive of an experience of inconceivable sort. However, in the present instance that is so for a reason that differs somewhat from that in Section 1.34. The present reason is that mathematical thought has failed to grasp the importance of Theorem 1.36.1.

Theorem 1.36.1:

The real continuum can be of service to science as an *indirectly* experiential concept *only*.

Section 1.14 issued an advance caution on this point. That our use of the continuum is limited by Theorem 1.36.1 is often overlooked because, as a matter of convenience, development of statistical theory and method is usually accomplished by loosely reasoning *as if* real-world sample spaces can be continuous, and then treating certain small sample problems involving discrete sample spaces *as if* they are special cases rather than cases that actually reflect *a universal reality*. So it must be firmly grasped that any set of descriptions used for recording data, will always necessarily be a discrete and essentially finite set. Overlooking this fact makes Stone's 1969 paper fall into metaphysical reasoning. In addition, we have obtained proof of that simply by asking of the outcome of that reasoning 'How can it be simulated?' and showing that the answer is 'It cannot!'

We note in passing that the foregoing example, together with Example 1.32.5, shows how simulation might enable us to distinguish a directly experiential concept (a predicative concept) from an indirectly experiential concept (a concept not predicative but necessitated by predictive reasoning).

We must distinguish *simulation* from *formal analogy*. Consider, for instance, the influence of two different sources of light on the orientation of a copepod with just one cyclopean eye. Fraenkel (1927) posited that the animal's path might be predicted by formal analogy in which the animal is drawn toward the two light sources like iron toward two magnets. By then formally treating light intensities as physical forces, the animal's path would be predicted by the resultant force. However, as pointed out by Maynard Smith (1972, p. 45), 'Analogies' and he means *formal* analogies, 'may be helpful in suggesting theories, but are irrelevant when it is a question of confirming or disproving them.' Thus for instance, Fraenkel's predictive model came to be viewed as tenable, not because of how he arrived at the model, but because the model seemed to match the corresponding data.

In statistical debate we are fortunate in that appropriately formulated proof by statistical simulation is so straightforward that it can hardly be challenged as yielding no more than formal analogy. Certainly, the proofs by statistical simulation we have developed in the present section and in Section 1.34 cannot be challenged thus, as they can be executed as laboratory experiments. That is to say, they can be executed such that one can call upon the human body, as the ultimate arbiter of science, and demand of it to 'Experience this for yourself'.

1.37 Spanning the interface

In the case of data arising from a randomised design, the design is part and parcel of *how the given data came about* and so cannot be ignored by investigative reasoning. On the contrary, such reasoning almost always, if not always, relies on the properties of the design. In the case of our analysis of the gladiolus data in Example 1.20.1, for instance, that is largely so. We must, however, be leery of the idea that randomisation obviates the need for commencement tests. The reason for that is that a sound theory of data analysis must necessarily span the interface between mathematical statistics and substantive science. In this section, and in the next three sections, we take steps toward clarifying this point.

Consider the data given in Table 1.1.1. Inspection of those data suggests that carbaryl treatment shortened the half-life of the fruit involved. In order to judge that more incisively, consider, as hypothesised model,

M_0: The whole of the sampling variation is attributable to randomisation only.

Let the ostensible effect of carbaryl on the half-life of the fruit, say d, be measured as the mean half-life for carbaryl-treated trees minus the mean half-life for control trees. Randomisation then brings to mind a corresponding random variable, D, such that under M_0 the expected value of D is zero. In order to test the quality of fit of

$$M_0 \text{ such that } E(D \mid M_0) = 0 \text{ vs. } M_1 \text{ such that } E(D \mid M_1) \neq 0,$$

we order the possible sample patterns according to the magnitude of D. The resulting test is left sensitive to reduced half-life owing to carbaryl. As the five smallest of the observed half-life values are also those five that arose with the carbaryl treatment, the observed value of d is the smallest possible under M_0. As there are just 10-choose-5, i.e. 252 different field plans, of which just one accounts for the observed value of d, and as the different plans arise with equal frequency, the mental correlate of the test datum is situated at

$$(*\varnothing,\ 1\div 252,\ 251\div 252) = (*\varnothing,\ 0.004,\ 0.996) \tag{1.37.1}$$

in the test distribution. This strongly suggests that carbaryl shortened the half-life of the fruit involved. Such reasoning extends to other hypothesised values for the effect of carbaryl. For instance, to test whether carbaryl treatment might have shortened by one day the half-life of the fruit involved, we modify the given data by adding one day to each of the five observed half-life values arising from the carbaryl-treated trees. This replaces the test datum and the test statistic by d+1 and D+1, respectively. Then, in order to test whether the modification cancels the ostensible effect of carbaryl, we order the modified sample patterns according to the magnitude of D+1 and test:

$$M_0 \text{ such that } E(D+1 \mid M_0) = 0 \text{ vs. } M_1 \text{ such that } E(D+1 \mid M_1) \neq 0.$$

Let T and C denote carbaryl treatment and control, respectively. Then the given value of D+1 arises as at (1.37.2) in the following, where the first row gives the original half-life values for carbaryl-treated trees, the second row gives in order of magnitude the half-life values as modified to test the new hypothesised model, and the third row corresponds to the field plan:

8.8	10.8		11.1	11.2	11.4					
9.8	11.8	11.9	12.1	12.2	12.4	12.8	13.1	13.1	14.4	
T	T	C	T	T	T	C	C	C	C	(1.37.2)

Smaller values of the test statistic arise from just three other field plans, as follows:

8.8	10.8		11.1	11.2	11.4					
9.8	11.8	11.9	12.1	12.2	12.4	12.8	13.1	13.1	14.4	
T	T	T	C	T	T	C	C	C	C	
T	T	T	T	C	T	C	C	C	C	
T	T	T	T	T	C	C	C	C	C	(1.37.3)

The sample pattern at (1.37.2) accounts for the statistical rounding. The three sample patterns at (1.37.3) account for the left co-ordinate. Owing to randomisation, there are just 252 different sample patterns, and they arise with equal frequency. So, under the hypothesised model, the mental correlate of the test datum is in this case situated at:

$$(*3\div252,\ 1\div252,\ 248\div252) = (*0.012,\ 0.004,\ 0.984) \qquad (1.37.4)$$

in the test distribution.

It now appears that many different hypothesised models can be tested in this manner. Let Δ denote the hypothesised effect of carbaryl, where $\Delta = 0$ vs. $\Delta < 0$ is tested at (1.37.1), and $\Delta = -1$ vs. $\Delta < -1$ is tested at (1.37.4). Let $-\infty < \Delta < +\infty$. Then we obtain a class of models indexed by Δ. The design is a random variable that ranges over 252 different field plans, where each plan occurs with frequency $1\div252$, *regardless of the value of* Δ. So (and this is the point of the foregoing development), *the design is a class characteristic*. Typically the design will have been generated by the use of an extensively tested source of pseudo-random numbers, such as the 10 000 pseudo-randomly assorted digits given in Table A1 of Snedecor and Cochran (1989). So it might be thought that the foregoing development does not need commencement testing. However, that is not the case. Suppose, for instance, that the ten experimental trees were evenly spaced in a straight row, and we then assigned carbaryl treatment to the trees whose row positions were given by the first five different digits in Snedecor and Cochran's table. In that case our field plan would have been:

Tree in position number	0	1	2	3	4	5	6	7	8	9
Treatment	C	C	T	T	T	T	T	C	C	C

Would we have used this plan? Or would we have discarded it, drawing another plan instead? Had we decided to draw another plan by using the *next* five different digits in Snedecor and Cochran's table, our field plan would have been:

Tree in position number	0	1	2	3	4	5	6	7	8	9
Treatment	T	C	T	C	C	T	T	C	C	T

Surely this would have seemed a better plan than the previous one? The issue is this: no sensible person would be prepared to base an investigation on the following plan:

Tree in position number	0	1	2	3	4	5	6	7	8	9	
Treatment	T	T	T	T	T	C	C	C	C	C	(1.37.5)

So, it is a fact of every-day statistical life that an investigator will occasionally discard a field plan drawn by valid randomisation and draw another plan instead. The issue cannot be evaded by arguing that it arises only with an inappropriate design. It arises with *any* design involving adequate slack to provide for statistical analysis. In case of the gladiolus trial of Example 1.27.1 for instance, the field plan giving the positions of older corms (O) and younger corms (Y) might physically, on the ground, in a straight row of seven pairs of plots, turn out to be:

(O, Y) (O, Y) (O, Y) (O, Y) (O, Y) (O, Y) (O, Y). (1.37.6)

This plan would surely be discarded in favour of drawing another plan.

In order to come to grips with the foregoing it must be firmly grasped that the problem belongs to *investigative* statistics. That is so, as it arises when we ask 'How may these *particular* data have come about?' and where we then wish to be able to discourage the unsavoury answer 'Possibly by way of a field plan that has confounded treatment effects with systematic variation owing to the substantive subject matter'. So, inspection of the field plan for its suitability is an inescapable *commencement test*. Clearly, it is a test that spans the interface between statistics and substantive science, as it amounts to judging the quality of fit of a class characteristic (a statistical design) with respect to a given data set (the field plan in relation to the substantive nature of the experimental material). Just as clearly, such judgement is of an axiomatic nature, as it cannot involve mathematical definition and deduction only. It must also involve, as exemplified at (1.37.5) and (1.37.6), judgement on possible systematic variation in the experimental material.

It might be thought that the foregoing involves self-contradictory reasoning, as it might be contended that rejection of the plans at (1.37.5) and (1.37.6) presupposes knowledge of the experimental material that, by the choice of the design, was tacitly denied. But that is easily disproved. If, for instance, forethought indicated a completely randomised design for allotting eight replicates of each of two treatments, A an B, to 16 trees evenly spaced in a 4×4 array, then the following field plans would be discarded:

A A A A	A A A B	A A B B
A A A A	A A B B	A A B B
B B B B	A A B B	A A B B
B B B B	A B B B	A A B B

The point here is that we would not be discarding such a plan because of knowledge of the *particular* fertility trends in the proposed experimental material. We would be discarding such a plan because of knowledge of the *sort* of fertility trends that often occurs in such material. We return to this point at the end of the next section, and again at the end of the section after that.

We note in passing that we would reluctantly discard a field plan, as we must be sure that, having thereby modified our design, the original design is an adequate approximation of the modified design. In practice, that is made possible by knowing that the number of plans that would be discarded is very small compared to the total number of possible plans. Stated otherwise, the practising statistician will make sure that field plans obtained by valid randomisation are very seldom discarded in favour of replacement. This is aided by avoiding, when possible, designs with little slack.

We also note that, instead of the randomisation tests presented at (1.37.1) and (1.37.4), corresponding tests using Student's *t* instead are, for the particular data set, justified. This is so partly by insight that anticipates Gaussian-like errors, and partly also by the commencement test at (1.32.2), which shows that a model of Gaussian-like errors, as tested, fits the half-life data well. In such cases we tend to favour Student's test over the randomisation test, as the former then has superior separating characteristics. Such use of Student's *t* amounts to modelling the error structure as having, above and beyond certain properties, owing to randomisation, also certain properties of a pseudo-Gaussian nature, owing to the substantive subject matter. All the various properties are then brought into account by treating the error structure *as if* Gaussian.

Clearly then, randomisation cannot obviate the need for commencement tests.

1.38 Beyond the interface

As statistical data analysis concerns the development of scientific evidence by way of physical facts, and as the development of such facts originates in substantive science, that is where our understanding of such matters must begin. Consider, for instance, a substantive scientist who wants to conduct an experiment with sheep. Examination of any two sheep will show that they differ in countless ways; they are *individuals* of the sort named 'sheep'. Such individuality is unavoidable. Consider, for instance, a group of sheep. Comparison to other groups of sheep will show that such groups are *further individuals* of the sort named 'group of sheep'. So the question arises: 'How can an investigator obtain "a group of sheep" that is suitable for a proposed experiment?' An inferior statistical literature tries to deal with such questions by prescribing random sampling of a target population, where the prescription is a circular one, as random sampling cannot turn sheep of 'an unsuitable sort' into sheep of 'a suitable sort'. The point here is that a substantive investigator may require for instance a representative variety of two-tooth merino ewes, veld-reared on typical Karoo range, normal, healthy and in good condition. Then the selection of suitable sheep will require appropriately experienced bodily perception. The meanings conveyed by 'a representative variety', 'in good condition', 'normal', 'typical Karoo range', etc. are familiar to the farmers and the substantive scientists involved. A novice joining that community initially has to *learn* those meanings by having this, that and the other matter to be grasped by bodily experience, being pointed out. The outcome is, so to speak, a language of the body, where there is no place for demands of the kind, *define* 'a representative' variety of sheep, *define* 'a normal' sheep, or *define* the line that separates a sheep 'in good condition' from a sheep 'in poor condition'. Such demands are out of place in the discourse of bodily experience, as they demand definable realities where only demonstrable realities are available.

We note in passing that use of Student's *t* as described in the last paragraph of the previous section does *not at all* imply that the experimental trees would thereby be modelled as 'a random sample from a more extensive population of trees'. The ten trees were, in the manner explained in this section, selected as ten suitable trees and not as a pseudo-random sample of ten trees. This point is often overlooked despite the excellent explanations by for instance Hinkelmann and Kempthorne (2007).

The matters discussed in this and the previous section are essentially the same in that we must judge whether *a particular individual* represents *the sort* of individual we have in mind, and where particular individuals of *that sort* demonstrably exist, but *that sort* is not capable of being defined.

1.39 The receding Ideal

Recall how we pointed at the standard representations in Figure 1.23.2 so as to force the human body to agree. When pointing thus, we are often trying to force the human body to grasp that a given co-ordination is extreme. Consider, for the sake of simplicity, a near-to-zero rounding. We might then be trying to force agreement by pointing at a standard representation where the ratio is:

(one of the two co-ordinates)**:**(the other co-ordinate)**::**95**:**5 (1.39.1)

This particular ratio is a widely accepted *norm of disproportion*. Similarly, the ratio

(one of the two co-ordinates)**:**(the other co-ordinate)**::**99**:**1 (1.39.2)

is a widely accepted *norm of extreme disproportion*. These particular norms are often considered an embarrassment, or are ridiculed on grounds that, for instance, a 16-fingered species would have chosen different norms. Yet, the norms survive, and not only in the context of the silly prescriptions considered in Section 1.21. They survive also as approximate benchmarks for practical data analysis where that is not at all a laughing matter.

The issue here is simply this: in the same way that the sheep farmers and their associated scientific community need to have examples pointed out so that a language of *bodily understanding* can be established, we too need to establish such a language. So when we point at physical representations of the type appearing in Figures 1.23.1 and 1.23.2, asking for co-ordinations '*like*' (0.90, 0.10), (0.95, 0.05) and (0.99, 0.01), to be recognised as 'slightly extreme', 'extreme', and 'very extreme', respectively, we are not laying down prescriptions or strict rules. We are offering *examples* of a kind of bodily experience so as to communicate in a language of such bodily understanding. And we must grasp that it is essentially the same kind of bodily understanding without which our customers in the substantive sciences would not in the first place have been able to conduct their investigations. So, instead of having nods and winks poke fun at what are after all just illustrative examples, we should instead take careful note of the following incontestable fact: when examples are physically displayed as in Figures 1.23.1 and 1.23.2, they *do in fact* communicate in 'a language of the body'. Similarly, when such figures are used to *physically* convey the norms *numerically* conveyed at (1.39.1) and (1.39.2), they *do in fact* communicate in that 'language' by exemplifying 'extreme' and 'very extreme' co-ordinates. We only have to bear in mind that, in that language of the body, 'extreme' and 'very extreme' are not definable. So the norms at (1.39.1) and (1.39.2) must not be thought of as *defining* extreme and very extreme co-ordinates, they must be understood as *exemplifying* those extremes. After all, we know full well that for various reasons, calculated statistical co-ordinates are more often than not approximate only. One good reason for that is that any sensible investigator must occasionally discard a field plan arising from valid randomisation, owing to which the mathematics of randomisation theory is, in respect of reality, approximate only.

This section and the previous two sections dealt with essentially just different versions of the same epistemological problem, namely that of the receding ideal. It is a very old problem and can be traced back to the works of Plato. In answer to a question raised by Socrates, Plato held that a horse, for instance, is recognised as such because of its resemblance to 'the Ideal horse'. That, however, leads to difficulty: what determines 'the Ideal horse'? If, on the one hand, 'the Ideal horse' is a horse, its recognition as such would have to be by virtue of its resemblance to 'the Ideal Ideal horse', whose recognition in turn would require 'an Ideal Ideal Ideal horse', and so on, into infinitely circular reasoning. And if, on the other hand, 'the Ideal horse' is not a horse, how is it to be distinguished from for instance 'the Ideal goat'? Again: when using a randomised design, do we accept a field plan because it resembles 'the Ideal field plan'? Is 'the Ideal field plan' itself a field plan? It would seem that Aristotle subsequently developed the correct solution to the problem of the receding ideal: individual, real-world horses are to be predicated of the indefinable concept (ultimate concept) 'horse', which concept can demonstrably be conceived by the human mind in response to the human body having experienced various real-world horses (Metaphysics, Z: 13-14; $1038^{b}8$ — $1039^{b}19$). Aristotle emphatically denies that the universal sorts of particular individuals can have any substantiality; so he holds that universals have no existence beyond cognition (i.e. beyond the human mind); the universals do not partake of the individuals, and the individuals do not partake of the universals. Thus, for example, the universal called '*random sample*' is a concept brought into the human mind in response to the human body having experienced particular real-world individuals such as *this* flip of coin, *that* roll of die, or *yonder* shuffle of a pack of cards. Here clearly the universal does not partake of the real-world individuals, and the real-world individuals do not partake of the universal.

The receding Ideal exemplifies the slippery nature of axiomatic problems. That slipperiness arises from an inclination to cast around for definitions, whereas axioms do not concern what can be defined, but what can be demonstrated only. This point must be grasped very firmly, otherwise we either become confused and fall into circular reasoning of the type that would have us turn 'sheep' into 'suitable sheep' by means of 'random sampling', or we succumb to normative ideas of the type that would try to define the meanings of 'tenable', 'hardly tenable', and 'untenable' in terms of the cut-and-dried prescriptions of the silly kind discussed in Section 1.21.

1.40 The scientific status of randomised designs

Section 1.37 contained a very simple demonstration of an important fact we formulate as Theorem 1.40.1. Note (as proof) that the theorem would apply no matter what class of models one may try to defend for the given data set in question, otherwise one will fall into self-contradiction.

Theorem 1.40.1:

The use of a randomised statistical design for the purpose of obtaining any given data set implies that that design is a class characteristic of any class of models that one might wish to defend as a possible explanation of how those particular data might have come about.

Chapter 2

ELIMINATION TESTS

Populations being deleted from the human mind

2.1 Introduction

The main thrust of the previous chapter was the development of tests for the quality of fit of isolated singletons when used for the representation of given data. Isolated singletons are the most elemental statistical models, and at the very outset Examples 1.7.1 and 1.8.1, and Examples 1.7.2 and 1.8.2 were introduced to demonstrate that class characteristics are instances of such singletons. In this chapter we wish to take first steps toward developing tests for the quality of fit of the alternative *members* of any given class, where it must then be clearly understood that a comprehensive theory of data analysis must, for its viability, present a cumulative development of consistent ideas. It simply will not do to commence by introducing certain ideas for the initial analysis of given data, and then to in effect jettison those initial ideas by proceeding with further ideas in conflict with the initial ones. This must be firmly grasped, as it is a source of much slippage in the statistical literature. So here, at the very outset of this chapter, we challenge the reader to note carefully that the following development is a seamless continuation of the reasoning that was developed in the previous chapter.

2.2 Null hypotheses as opposed to hypothesised models

Hypothesis tests address certain problems in decision-making under risk; they do not address problems in data analysis properly at all, although they are often mistakenly thought to do so. Owing to that mistake, certain concepts of hypothesis testing invade much of the literature on data analysis, and as those concepts are not motivated by the needs of investigative science, the invasion sows confusion. Amongst those sources of confusion are the concepts *simple hypothesis* and *composite hypothesis*. In order to come to grips with this, consider 25 measurements that can satisfactorily be modelled as a random sample of $N(\mu, 5^2)$ values for $\mu \in \{-2, -1, 0, +1, +2\}$. *A hypothesis test* is a decision rule for accepting *the null hypothesis* H_0, or else *the alternative hypothesis* H_1, where a hypothesis is a set of models. It is a *simple* hypothesis if it comprises just one member of a class, and a *composite* hypothesis if it comprises several members of a class. For example, in the present case:

$$H_0: \mu \in \{0\} \text{ versus } H_1: \mu \in \{-2, -1, +1, +2\} \qquad (2.2.1)$$

proposes to test a simple null hypothesis against a composite alternative, whereas

$$H_0: \mu \in \{-2, -1, 0\} \text{ versus } H_1: \mu \in \{+1, +2\} \qquad (2.2.2)$$

proposes to test a composite null hypothesis against a composite alternative. Despite the simplicity of our example, a large variety of such pairs of hypotheses can clearly be proposed for it. Moreover, for any particular pair there are many hypothesis tests from which to choose. In the case of the pair at (2.2.1) for instance, and even after the value of the Type I error rate has been specified as $\alpha = 0.05$ say, there remain infinitely many tests from which to choose, as we can then specify infinitely many different rules for accepting H_1. In the present case for instance that rule might be any one of the following, where $\overline{X}$ denotes the sample mean:

$$\overline{X} \leq -1.645. \quad \overline{X} \geq +1.645. \quad |\overline{X}| \geq 1.960. \quad \overline{X} \text{ either} \leq -2.575 \text{ or} \geq +1.696, \ldots .$$

Our example is not exceptional in this respect. The literature on hypothesis testing and related developments for decision-making under risk constitutes a vast pharmacopoeia of technological recipes. In terms of Definition 1.2.2, they are recipes for *how this, that and the other experience can be brought about*. By comparison, the literature on data analysis must be anticipated to be much smaller, as we must then, in terms of Definition 1.2.1, be concerned with the question of *how given data might have come about*, where that is a matter of *investigation*, rather than one of *specification*. Thus, for instance, instead of being able to specify infinitely many hypothesis tests for the present example, we are able to defend for it just five co-ordination tests. These tests arise as follows: any question about the tenability of different values for μ can in the present case best be addressed by way of the sample mean, as that is here the minimal sufficient statistic for μ. So there are then only five singletons:

$$N(\mu, 1) \text{ for } \mu = -2, -1, 0, +1, +2,$$

wherein the mental correlate of the datum mean might then alternatively be placed. If for instance the datum mean equals $\overline{x} = 0.30$ with negligible rounding, its mental correlate is to be found within these five singletons, at:

$$(0.99, 0.01), (0.90, 0.10), (0.62, 0.38), (0.24, 0.76), (0.04, 0.96) \qquad (2.2.3)$$

respectively; and for the given problem, these are the only tests one can defend successfully. Note that the two extremes fit the given datum poorly. There is no need for a pointer, as we have arrived at a closed system of alternatives. Now consider (and this is crucial) how to reason in order to grasp what the foregoing array of co-ordinates tells us. It then appears that we do not need to reason in terms of more than just two singletons at a time. To give examples:

In order to test M_0: $\mu = -2$ versus M_1: $\mu = 0$, the co-ordinates to be compared are (0.99, 0.01) versus (0.62, 0.38), strongly favouring M_1 over M_0.

In order to test M_0: $\mu = 0$ versus M_1: $\mu = +2$, the co-ordinates to be compared are (0.62, 0.38) versus (0.04, 0.96), strongly favouring M_0 over M_1.

In order to test M_0: $\mu = -1$ versus M_1: $\mu = +1$, the co-ordinates to be compared are (0.90, 0.10) versus (0.24, 0.76), only slightly favouring M_1 over M_0. (2.2.4)

and so on. Also, as previously noted in connection with Definition 1.23.1, when M_0 and M_1 are both fully specified singletons, they are on an equal footing, and can therefore be

relabelled M_1 and M_0 without other consequence if done consistently throughout. This has brought us to a fundamentally important fact:

> Any tests of co-ordination can be analysed into pair-wise comparison of singletons for their tenability as alternative explanations of how a given datum might possibly have come about (and where, of course, in the case of commencement testing, one member of the pair is only a broadly envisaged incipient model).

The reason for this is simply that the outcome of a co-ordination test is never qualified by probability. So the outcome of a suite of such tests, for instance as exemplified at (2.2.3), is also never qualified by probability. *Any notions about decisions subject to 'error rates', or about interval estimates subject to 'coverage probabilities', are simply out of place in the theory of co-ordination tests*. That does not imply that co-ordination testing cannot involve a *variable* alternative. For instance, instead of the formulation at (2.2.1) a co-ordination tester might well be interested in testing:

M_0: $N(0, 5^2)$ versus M_μ: $N(\mu, 5^2)$, *pair-wise* for μ = -2, -1, +1, +2, *one by one*.

And as we have explained at (2.2.4), that has already been provided for at (2.2.3). In other words, a model can be a member of an array of models, but there can be no such creature as a 'composite model' for a given data set, because a model cannot be a set of *contradictorily* different models for the given data set. In order to grasp this, one need only grasp that, given the datum $x = 0.30$, a co-ordĩnation tester who has mistaken H_0 as specified at (2.2.2) for a hypothesised model, would be trying to calculate three 'numbers' given by:

$$U = \Pr(\overline{X} < 0.30 \mid \mu \in \{-2.\ -1,\ 0\})$$
$$\varepsilon = \Pr(\overline{X} = 0.30 \mid \mu \in \{-2.\ -1,\ 0\})$$
$$V = \Pr(\overline{X} > 0.30 \mid \mu \in \{-2.\ -1,\ 0\})$$

where no such 'numbers' exist (cf. Kempthorne and Folks 1971, pp. 336-337).

There is more to come.

Let the test proposed at (2.2.1) be for $N(\mu, \sigma^2)$ with $0 < \sigma^2 < \infty$. Then a very different case arises because the null hypothesis corresponding to the one at (2.2.1), i.e.

$$H_0: \{\mu, \sigma^2\} \in \{0, \sigma^2\} \qquad (2.2.5)$$

is now, in the language of hypothesis testing, 'a composite hypothesis' rather than 'a simple hypothesis', as it is a set of different singletons corresponding to the different values of σ^2. However, a co-ordination tester cannot conceive of 'a composite model', as that refers to a set of *contradictorily* different models for a given data set. A co-ordination tester conceives of the set specified at (2.2.5) as an array of *alternative* models for the selfsame data; we call such an array *a suite of models*, which (in this case) is indexed by *a nuisance parameter*. In other cases such a suite of models might be indexed by *a parameter of interest* only, or by parameters of both kinds. We have preferred to speak of a *suite* of models rather than a *class* of models because, as indicated by Definition 1.15.1, elimination tests do not involve class characteristics. So, a suite of models proposed for any particular data set will have been derived from a minimally sufficient statistic for the index of a class of models.

The foregoing development shows that from a data analyst's point of view, the term composite hypothesis fails to separate two very different cases, as exemplified at (2.2.2) and (2.2.5). That is so because hypothesis tests do not pursue the problem of modelling an individual data set; instead they pursue the problem of producing a host of individuals that satisfies certain specifications.

2.3 Measuring the quality-of-fit of alternative members of a class of models

When a data set, its substantively circumstantial details and suitable commencement tests have brought a class of models into the human mind, it is often (even usually) the case that not all the members of the class are tenable models of 'how the given data might have come about'. In such cases the untenable index values must then be weeded out. We refer to that as *elimination testing*. Now recall that, for any class of models, the only predictions that can differentiate amongst the different members of the class are those that depend on the index. So it would be silly to use an elimination test whose outcome depends on any prediction of the class characteristic. Such silly tests were ruled out by Definition 1.15.1. (See also Section 1.33.)

In this section and sections that follow, we develop a number of examples for the elucidation of elimination tests. Although we will consider tests that make sense, we are not at this stage seriously concerned about whether or not these tests are 'best' in any sense. At this stage of development our main purpose is to provide sufficient insight into the ideas of such tests for the reader to be able to grasp the extent to which the ideas differ, and even conflict, with ideas to be considered in subsequent chapters. We introduce certain terminology as we go along.

Example 2.3.1

Consider the *suite of elimination tests* leading to the five sets of co-ordinates given at (2.2.3). Given *the test datum* $\bar{x} = 0.30$, we wished to test each member of *a suite of hypothesised models* $N(\mu, 1)$ indexed by $\mu = -2, -1, 0, +1, +2$. This was done for each hypothesised model by calculating a set of co-ordinates of the form:

$$C(\bar{x}, \mu) = [U(\bar{x}, \mu), \varepsilon(\bar{x}, \mu), V(\bar{x}, \mu)]$$

giving directions to just where within that particular hypothesised model, the mental correlate of the test datum is being situated. We refer to the function that maps the index onto the co-ordinates as *the co-ordinate trace (of the mental correlate)*. Using the results at (2.2.3), the co-ordinate trace for the present example is found to be:

$$[\mu, C(\bar{x}, \mu)] = [-2, (0.99, 0.01)], [-1, (0.90, 0.10)], [0, (0.62, 0.38)], \ldots .$$

We note that, for the given data set, the function traces the different situations of the mental correlate from model to model in the human mind. The test datum, $\bar{x} = 0.30$ in the present case, usually arises from a larger data set in the real world. We refer to the corresponding random variable brought into the human mind as *the elimination quantity*, previously denoted as $\bar{X}$ in the present example. An elimination quantity provides

indexed predictions to be tested against the test datum. These predictions might be indexed in one or both of two different ways. One way is for the predicted distribution to be indexed. For instance, as $\bar{x} = 0.30$ in the present case, the values of $V(\bar{x}, \mu)$ can, with $\overline{X}$ as the elimination quantity, be calculated as:

$$\Pr(\overline{X} > 0.30) = \frac{1}{\sqrt{2\pi}} \int_{0.30}^{\infty} \exp[-\frac{1}{2}\ (z-\mu)^2]dz \text{ for } \mu \in \{-2, -1, 0, +1, +2\}.$$

The other way is for the elimination quantity itself to be indexed. For instance, in the present case, the values of $V(\bar{x}, \mu)$ can, with $\overline{X} - \mu$ as the elimination quantity, also be calculated as:

$$\Pr(\overline{X} - \mu > 0.30 - \mu) = \frac{1}{\sqrt{2\pi}} \int_{0.30-\mu}^{\infty} \exp[-\frac{1}{2}\ (z)^2]dz \text{ for } \mu \in \{-2, -1, 0, +1, +2\}.$$

So, in this particular example, there are two different elimination quantities for the self-same suite of tests:

$\overline{X}$ whose distribution is $N(\mu, 1)$, and
$\overline{X} - \mu$ whose distribution is $N(0, 1)$.

The distribution of $\overline{X}-\mu$ is independent of the index; we call such an elimination quantity *an elimination pivot*.

Example 2.3.2

Let us take up the problem raised in Example 1.8.2, where n = 6 municipal buses numbered 1, 2, 3, 4, 9 and 10, and spotted in a town T, prompted the development of a model giving the minimal sufficient statistic for θ, the total number of municipal buses in T, as follows: letting $x_{(n)}$ denote the largest number observed, and $X_{(n)}$ denote a corresponding random variable in the human mind, we found that if the n = 6 numbers are modelled as a random sample obtained without replacement, $X_{(n)}$ is a sufficient statistic for θ. In fact, $X_{(n)}$ is minimally sufficient for θ. The probability of $X_{(n)} = x_{(n)}$ was given by the first factor in square brackets at (1.8.2) as:

$$\begin{pmatrix} x_{(n)}-1 \\ n-1 \end{pmatrix} \begin{pmatrix} \theta \\ n \end{pmatrix}^{-1} \text{ for } x_{(n)} = n, n+1, n+2, \ldots, \theta.$$

For the number of observed buses in our example, n = 6, this probability is given by:

$$6[(x_{(n)}-1)(x_{(n)}-2)(x_{(n)}-3) \ldots (x_{(n)}-5)] \div [\theta(\theta-1)(\theta-2) \ldots (\theta-5)]$$
$$\text{for } x_{(n)} = 6, 7, 8, \ldots, \theta. \qquad (2.3.2)$$

As our datum is $x_{(n)} = 10$, we cannot have $\theta < 10$. Consider M_0: $\theta = 11$. By using the recipe at (2.3.2) we obtain the model given in Table 2.3.1, where small $x_{(n)}$ values point at a 'poor

Table 2.3.1: Hypothesised distribution of $X_{(n)}$ for n = 6 and θ = 11

$x_{(n)}$	6	7	8	9	10	11
$Pr[X_{(n)} = x_{(n)} \mid \theta = 11]$	0.00	0.01	0.05	0.12	0.27	0.55

fit'. So, in order to test this model, let the $x_{(n)}$ values be ordered from largest to smallest. The modelled counterpart of the datum $x_{(n)} = 10$, is by means of Table 2.3.1 found to be co-ordinated well within the crowd at (0.18, 0.27, 0.55) in the test distribution. The singletons M_0: $\theta = J$ for J = 10, 11, 12, ... can now be tested *one by one* in this way. Some of the resulting co-ordinates for the mental correlate of the datum are displayed in Table 2.3.2, showing that, by the tests here performed, $\theta = 10, 11, 12$ fit the data well, $\theta = 13$ fits slightly awkwardly, and $\theta = 14, 15, 16, \ldots$ fit the data poorly.

Table 2.3.2: Co-ordinating trace arising from a suite of elimination tests

Hypothesised model	Co-ordinates of the mental correlate
θ = 10	(0.40, 0.60, Ø)
θ = 11	(0.18, 0.27, 0.55)
θ = 12	(0.09, 0.14, 0.77)
θ = 13	(0.05, 0.07, 0.88)
θ = 14	(0.03, 0.04, 0.93)
θ = 15	(0.02, 0.02, 0.96)
θ = 16	(0.01, 0.02, 0.97)
...	...

We note again that the pointer was introduced as a *scaffolding* symbol. It served only to remind us that any unusual co-ordination has to be interpreted as nothing more than just that, unless a substantively conceivable alternative explanation of how the given datum might have come about is thereby being pointed out. So the pointer, having served its purpose, may now be dropped, as an elimination test deals with a closed system of explicit alternatives.

Now consider how the co-ordinates for testing M_0: $\theta = 11$ arose. A data set in the real world indicated that the value of θ might be 11. Further thought on how the given data might have come about then brought into the human mind a population of many samples. The question of how to then test whether or not the value of θ might be 11, brought into the human mind an ordering of the many $x_{(n)}$ values arising from the many samples. We then considered the real-world value $x_{(n)} = 10$, and calculated that the co-ordinates (0.18, 0.27, 0.55) would give directions to just where amongst the many $x_{(n)}$ values in the human mind, the mental correlate of that real-world value was being situated. This situation is depicted in Figure 2.3.1.

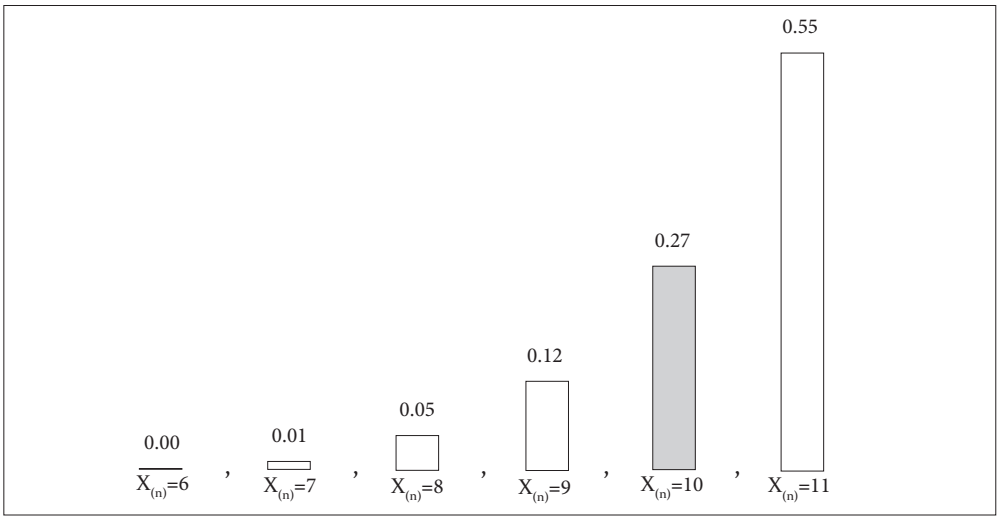


Figure 2.3.1: Testing the quality of fit of a member of a class of models for a number of municipal buses

The shaded bar depicts a statistical rounding of size 0.27, where the mental correlate is to be found. Clearly then, we can point at Figure 2.3.1 or, if needs be, at a simulated counterpart, and declare: 'See for yourself how the hypothesised model, as *here* being tested, fits the given data well. See how snugly the mental correlate of the datum is being placed within the crowd.' We would thus use the method once used by Galileo at Pisa. It is the method that is invariably used by substantive science for the marshalling of its ultimate evidential facts, which facts are of course *physical* facts, that is to say, facts capable of being forced upon the human body.

It might be objected that we are neglecting the outcome of the test that we displayed in Figure 1.8.2 where, using the method of runs, we found that our model, as tested *there*, does not fit the data well. However, we reply that this is no cause for embarrassment: the two tests are tests of the quality of fit of two different subsidiary models with respect to two different subsidiary data sets. In principle, that does not differ at all from how, for instance, we might point and say: 'The *shape* of this spoor *is unlike that* of an aardwolf. But *if it were to be* the spoor of an aardwolf, the *size* of the spoor *is like that* of a juvenile.' In other words, we must firmly rebuff any arguments to the effect that the test here performed rests on the *assumption* that the six numbers were drawn as if by random sampling without replacement. When we say '... *if it were to be* the spoor of an aardwolf', we are not in any way *assuming* it to be the spoor of an aardwolf. The point being made here must be firmly grasped: the term *assumption* does not belong to data analysis, as data analysis is the discourse of the pursuit of knowledge. The term belongs to technology, as technology is the discourse of the use of knowledge, *which discourse often needs to be assisted by assumptions*. In other words, scientific investigation is the pursuit of *physical facts*, whereas scientific technology is the pursuit of *physical rewards*. So, the former does not want presumed 'facts' to be based on assumed 'facts', whereas the latter will often settle for the expected rewards of augmenting facts with reasonable assumptions. In short, a technology may *act upon assumptions*, but an investigation must *point at facts*.

The foregoing dealt with the very simple case in which a scalar statistic is minimally sufficient for a scalar parameter. The next example displays a more complex situation in that the minimal sufficient statistic is a vector, rather than a scalar, although the index is a scalar.

Example 2.3.3

Consider a sample of n integers drawn as if at random without replacement from the set of all the integers straddled by two integers, θ and 2θ, such that $\theta > 3$ and $2 \leq n \leq \theta\text{-}1$. Let $x_{(1)}$ denote the smallest number drawn, and let $x_{(n)}$ denote the largest number drawn. Then the corresponding random variables $X_{(1)}$ and $X_{(n)}$ are (jointly) minimally sufficient for θ. Their joint distribution was previously given by the first of the two factors in the square brackets at (1.12.1). For n = 3 the distribution is given by:

$$\Pr\{[X_{(1)}, X_{(3)}] = [x_{(1)}, x_{(3)}]\} = 3(3\text{-}1)\ \frac{[x_{(3)}\text{-}x_{(1)}]\text{-}1}{(\theta\text{-}1)(\theta\text{-}2)\ (\theta\text{-}3)}, \text{ where}$$

$x_{(1)} = \theta+1, \theta+2, \theta+3, \ldots, 2\theta\text{-}3$
$x_{(3)} = \theta+3, \theta+4, \theta+5, \ldots, 2\theta\text{-}1$
and $x_{(1)}+2 \leq x_{(3)}$, because n = 3. (2.3.3)

Let the given data be $(x_{(1)}, x_{(3)}) = (16, 19)$.

As $\theta+1 \leq x_{(1)}$ and $x_{(1)} = 16$, we cannot have $\theta > 15$.
As $2\theta\text{-}1 \geq x_{(3)}$ and $x_{(3)} = 19$, we cannot have $\theta < 10$.

So, our class of models comprises just six different members indexed by $\theta = 10, 11, 12, \ldots, 15$. Consider M_0: $\theta = 11$. Using the formulae given at (2.3.3), the joint distribution and the marginal distributions of $X_{(1)}$ and of $X_{(3)}$ are easily found (Table 2.3.3). It appears from

Table 2.3.3: Joint and marginal distributions of $X_{(3)}$ and $X_{(1)}$ when $\theta = 11$ and n = 3.
$X_{(1)}$ ranges from $\theta+n\text{-}2$ to $2\theta\text{-}n$ inclusive, i.e. from 12 to 19 when $\theta = 11$ and n = 3.
$X_{(3)}$ ranges from $\theta+n$ to $2\theta\text{-}1$ inclusive, i.e. from 14 to 21 when $\theta = 11$ and n = 3.

$X_{(1)}$	12	13	14	15	16	17	18	19	Total
$X_{(3)}$									
14	1/120								1/120
15	2/120	1/120							3/120
16	3/120	2/120	1/120						6/120
17	4/120	3/120	2/120	1/120					10/120
18	5/120	4/120	3/120	2/120	1/120				15/120
19	6/120	5/120	4/120	3/120	2/120	1/120			21/120
20	7/120	6/120	5/120	4/120	3/120	2/120	1/120		28/120
21	8/120	7/120	6/120	5/120	4/120	3/120	2/120	1/120	36/120
Total	36/120	28/120	21/120	15/120	10/120	6/120	3/120	1/120	1

these distributions that either large values of $x_{(1)}$, or small values of $x_{(3)}$ point at a 'poor fit'. As θ bounds $x_{(1)}$ from below, that implies that a large value of $x_{(1)}$ is sensitive to a value of θ that is too small to supply a good fit (i.e. try a larger θ value – one closer to $x_{(1)}$). Similarly, as 2θ bounds $x_{(3)}$ from above, that implies that a small value of $x_{(3)}$ is sensitive to a value of θ that is too large to supply a good fit (i.e. try a smaller 2θ value – one closer to $x_{(3)}$). Let us then consider the $X_{(1)}$ test statistic for M_0: $\theta = 11$, as given by the $X_{(1)}$ margin of Table 2.3.3, and let us calculate that the mental correlate of the given datum, i.e. of $x_{(1)} = 16$, is being situated within the $X_{(1)}$ test distribution, at

$$\left\langle \frac{36+28+21+15}{120}, \frac{10}{120}, \frac{6+3+1}{120} \right\rangle = (0.833, 0.083, 0.083).$$

Here the pointing co-ordinate is on the right, corresponding to a large $x_{(1)}$ value. For the $X_{(3)}$ test statistic, the pointing co-ordinate is on the left, corresponding to a small $x_{(3)}$ value. Using the $X_{(1)}$ and $X_{(3)}$ test statistics separately, in each case to test, one by one, the quality of fit of the six models M_0: $\theta = J$ for J = 10, 11, 12, ... 15 for the given data, Table 2.3.4 arises. On the

Table 2.3.4: The co-ordinating traces of the $X_{(1)}$ and $X_{(3)}$ tests of $\theta = J$ for J = 10, 11, 12, ..., 15. The data for the two suites of tests are $X_{(1)} = 16$ for the $X_{(1)}$ suite and $X_{(3)} = 19$ for the $X_{(3)}$ suite with sample size being n = 3.

Hypothesised model	Co-ordinates of $x_{(1)}$	Co-ordinates of $x_{(3)}$
$\theta = 10$	(0.95, 0.04, 0.01)	(0.67, 0.33, Ø)
$\theta = 11$	(0.83, 0.08, 0.08)	(0.29, 0.18, 0.53)
$\theta = 12$	(0.66, 0.13, 0.21)	(0.12, 0.09, 0.79)
$\theta = 13$	(0.45, 0.16, 0.38)	(0.05, 0.05, 0.91)
$\theta = 14$	(0.23, 0.19, 0.58)	(0.01, 0.02, 0.97)
$\theta = 15$	(Ø, 0.26, 0.74)	(0.00, 0.01, 0.99)

one hand, M_0: $\theta = 10$, as tested by the $X_{(1)}$ test, fits the data poorly, pointing at larger values of θ. On the other hand, M_0: $\theta = 15$ and M_0: $\theta = 14$, as tested by the $X_{(3)}$ test, fit the data poorly, pointing at smaller values of θ.

The present example nicely exemplifies why co-ordination tests never require us to reason in terms of more than a pair of hypothesised singletons at a time. In order to see this, note that in Table 2.3.4 we can pick out co-ordinates arising from:

$$2\times \left\langle \begin{matrix} 6 \\ 2 \end{matrix} \right\rangle = 30 \text{ pairs of } X_{(j)} \text{ tests } (j = 1, 2).$$

Consider the pair of $X_{(1)}$ tests for M_0: $\theta = 10$ versus M_1: $\theta = 14$. Table 2.3.4 tells us that for this particular pair, the mental correlate of the $x_{(1)}$ datum is to be found at (0.95, 0.04, 0.01) and (0.23, 0.19, 0.58) in the respective test distributions. The two situations are displayed in Figure 2.3.2.

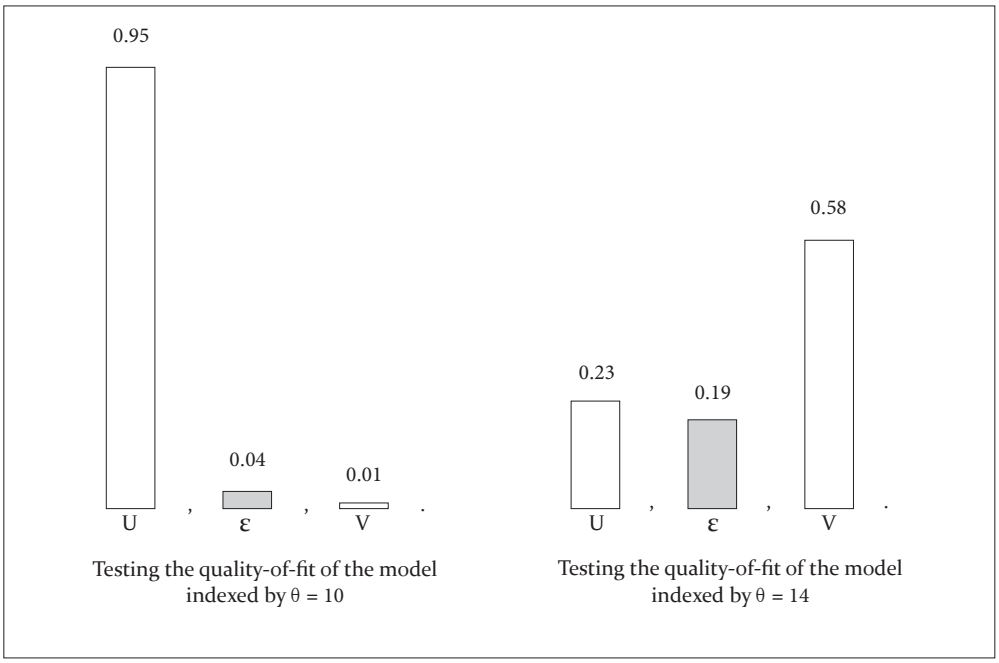


Figure 2.3.2: Comparing the quality of fit of two alternative members of a class of models for the value of an unknown integer θ

The meanings displayed in Figure 2.3.2 can, if needs be, be forced upon the human body by simulation. They are meanings that belong to the discourse of physical evidence, as we can point and say:

> 'Look at Figure 2.3.2, and see for yourself how awkwardly the mental correlate of the $x_{(1)}$ datum is situated on the outskirts of the crowd that θ = 10 brings to mind.'
>
> 'Look at Figure 2.3.2, and see for yourself how snugly the mental correlate of the $x_{(1)}$ datum is situated within the crowd that θ = 14 brings to mind.'

The human body is thus forced to grasp that, by the two tests performed here, the θ = 14 explanation of 'how this $x_{(1)}$ datum might have come about' fits the datum better than does the θ = 10 explanation.

We challenge the reader to note that, just as in the previous chapter, so also in the present chapter, co-ordination tests produce findings whereof the veracity cannot be qualified by probability. The findings are facts whose veracity is absolute, having by physical perception (bodily perception) been placed beyond reasonable contest. As this point must be firmly grasped, let us state and consider precisely what the findings of the previous paragraph are, as follows:

> By the tests performed, the explanation offered by θ = 10 fits the data poorly, whereas the explanation offered by θ = 14 fits the data well.

Should we be compelled to attach a 'probability of truth' to this statement, we would perforce have to declare that it be 1 (unity), as the statement is plainly a fact. We must,

however, be exceedingly reluctant to introduce such a 'probability', as it can serve no positive purpose for an irrelevant concept to be dragged into our development.

Example 2.3.4

The index of a class of models often represents some or other quantity of interest such as a number of municipal busses, or a percentage of viable seed, or a variance of human statures, or the effect of a vitamin supplement on the egg production of laying hens. In each of these examples the quantity of interest ranges over an ordered scale of values. It is important to understand that such is not always the case. So, consider a yield trial with tomato cultivars named Money Maker, Beauty, Bonny Best and Juicy Lucy, with equal numbers of replication in a completely randomised design, where interest is in potential fruit production, the higher the better. (If interest were to be in the percentage of defective fruits, the defining phrase would be the lower the better.) For any of the entries, say Beauty, an unknown θ names the best of the other entries. We consider a case for which the standard analysis of variance can be defended. So, the following random variables are brought into the human mind:

$\overline{X}_{\theta}$, representing the mean yield of θ, where $E(\overline{X}_{\theta}) = \mu_{\theta}$

$\overline{X}_{Beauty}$, representing the mean yield of Beauty, where $E(\overline{X}_{Beauty}) = \mu_{Beauty}$

$\overline{X}_{\max \bullet \text{ Beauty}}$, representing the highest mean yield excluding that of Beauty

s^2, representing the error variance of the trial.

The extent to which Beauty's yield potential falls short of that of its unknown best competitor named θ, is represented by the shortfall parameter:

$$\mu_{\theta} - \mu_{Beauty} = \delta_{Beauty}, \tag{2.3.4}$$

which shortfall may of course be negative, as Beauty might be better than θ. It is impossible to estimate this shortfall untrammelled by nuisance parameters involving the expected yields of entries other than Beauty and θ. But with all other parameters fixed, the larger the shortfall of Beauty would be, the larger the expected value of

$$\overline{X}_{\max \bullet \text{ Beauty}} - \overline{X}_{Beauty}.$$

would be. So to test whether Beauty must be indexed by positive shortfall, we employ the *t*-like statistic of Dunnett (1955) given by:

$$t' = \frac{\overline{X}_{\max \bullet \text{ Beauty}} - \overline{X}_{Beauty}}{\sqrt{\frac{2s^2}{n}}}, \text{ where } n \text{ is the number of replications.} \tag{2.3.5}$$

In this expression it is wrong to interpret the denominator as the standard error of the numerator, because the first term of the numerator is of an order-statistical nature. In order to underscore this, the denominator is in some of the literature taken to be the standard error of a single treatment mean. In fact, as it merely serves to eliminate the error variance as a nuisance parameter, it may be taken to be Cs, for C any positive constant. Let us, for the moment, imagine that in order to test whether or not a tenable model for the data in

hand needs a positive shortfall value for Beauty, we can supply the values of the remaining nuisance parameters. Then we would be able to supply the test distribution within which the mental correlate of the calculated *t'* value is to be seated, and small values of the right co-ordinate would then point at Beauty as being lower than best. In order to overcome the problem of not knowing what values are to be supplied for the remaining nuisance parameters, we replace the probability of *t'* exceeding any given value, with the *supremum* of that probability over the nuisance parameter space. So when referring a given *t'* value to Dunnett's test distribution, the statistical co-ordinates obtained are *the leftmost co-ordinates* for the given *t'* under all the various possible nuisance parameter values. Consider the hypothetical data given in Table 2.3.5.

Table 2.3.5: A suite of tests for the elimination of entries that are lower than best

Entry name	Mean yield (kg/plot)	Shortfall	*t'* value	Left-most co-ordinate when modelling the entry as best
Money Maker	44	-7	-0.99	(0.040, ε, 0.960)
Juicy Lucy	37	+7	+0.99	(0,662, ε, 0.338)
Bonny Best	29	+15	+2.13	(0.934, ε, 0.066)
Beauty	26	+18	+2.55	(0.968, ε, 0.032)

The estimated standard error of an entry mean equals 4.99 kg on 12 df.

We wish to test:

$$M_0: \mu_\theta - \mu_{Beauty} \leq 0 \text{ versus } M_1: \mu_\theta - \mu_{Beauty} > 0.$$

Note that M_0 involves two different possibilities:

$\mu_\theta - \mu_{Beauty} < 0$ means Beauty is the sole best entry, whereas
$\mu_\theta - \mu_{Beauty} = 0$ means Beauty is one of several best entries.

So we may express the alternatives more explicitly, as follows:

M_0: 'Beauty is a best entry' versus M_1: 'Beauty is lower than best'.

The observed mean for Beauty falls short of the highest mean observed amongst the other entries by +18 kg. Using the formula given at (2.3.5), the corresponding value of Dunnett's -like statistic is found to be:

$$t' = 18 \div (4.99 \times \sqrt{2}), \text{ i.e., } t' = 2.55 \text{ on 12 df. with three competitors.} \qquad (2.3.6)$$

From a table in Dunnett (1955) we find that in this case the right co-ordinate of t'=2.29 equals 0.05 and that of t'=3.19 equals 0.01. Linear interpolation gives the co-ordinates of the mental correlate of the value obtained at (2.3.6) as (0.962, ε, 0.038). More exact co-ordinate values, given in Table 2.3.5 as (0.968, ε, 0.032), are obtained by adapting the PROBMC function of SAS (1992). The pointing co-ordinate in such a test is always on the right. As the computed co-ordinates are *the leftmost co-ordinates* that arise when taking account of all possible nuisance parameter values that might arise from entries other than Beauty and one or more unknown best competitors, we should strictly speaking say:

> The mental correlate of the test datum is to be found *at least* as far down as at (0.968, ε, 0.032) in the right-hand tail of the hypothesised distribution.

In other words, Dunnett's *t'* has enabled us to pick out from amongst the models in the human mind, that model which is most tenable under the proposition 'Beauty is a "best" entry.' As this 'most tenable' model, by the test performed, fits the given data poorly, every one of the models in which Beauty is modelled as a 'best' entry, must then, by the test performed, fit the given data poorly. Table 2.3.5 displays the results of a suite of such elimination tests. In this suite of tests the parameter of interest is:

$$\theta \in \{\text{Beauty, Bonny Best, Juicy Lucy, Money Maker}\} \qquad (2.3.7)$$

The parameter ranges over *a set of identities*, and *not over an ordered scale of values*. Our findings are that by the suite of elimination tests performed in Table 2.3.5, *all the models* in which Beauty is modelled as a best entry fit the given data poorly, and *all the models* in which Bonny Best is modelled as a best entry fit the given data poorly, although not quite as poorly as Beauty, and *not all the models* in which Money Maker and/or Juicy Lucy are modelled as best entries fit the given data poorly. We note that Table 2.3.5 effectively displays only three tests, not four tests, as the highest yielding entry can never, by the method used, be judged 'lower than best'. So the co-ordinates given in the table for Money Maker are redundant and may be omitted.

Example 2.3.6

Shortfall testing sometimes involves a number of entries and a control. An example involving six entries and a control is provided by a yield trial concerning the recovery of *Pythium* when bits of infested lucerne tissue are plated on growth media made up of various kinds of agar. Recovery rates (from 50 bits per replicate) were transformed to angles leading to the analysis given in Table 2.3.6.

Table 2.3.6: Shortfall tests involving a control

Growth medium	Mean	Test 1 (7 entries in all)		Test 2 (6 entries in all)	
	(3 reps)	Shortfall	Co-ordinate	Shortfall	Co-ordinate
Corn meal agar	28.5°			-0.9°	(0.11, ε, 0.89)
Lima bean agar	27.6°			+0.9°	(0.25, ε, 0.75)
Water agar (control)	27.6°	+0.9°	(0.22, ε, 0.78)		
Potato-carrot agar	26.1°			+2.4°	(0.41, ε, 0.59)
V-8 agar	25.0°			+3.5°	(0.54, ε, 0.46)
Oat agar	22.4°			+6.1°	(0.81, ε, 0.19)
Potato-dextrose agar	16.7°			+11.8°	(0.99, ε, 0.01)

The estimated standard error of an entry mean equals 2.709° on 96 df.
(The error source was a wider analysis involving factorial interaction.)

Because water agar, a standard medium, has a low nutrient content, interest was directed at the possibility of increasing yields by using any of six alternative nutrient-enriched media. As the alternatives require extra effort to prepare and are somewhat difficult to work with, we must begin by asking Question 1:

> Does the recovery rate for water agar evidently fall short of that for the unidentifiable best one amongst the six nutrient-enriched media?

An affirmative answer would then prompt us to ask Question 2:

> Can any members of the group of six nutrient-enriched media be ruled out as evidently lower than best within that group?

Question 1 leads to Test 1 in Table 2.3.6, and it so happens that in this case the evidence would not persuade us to replace the standard medium (control) with any of the other media. In other words, there is no indication that the standard falls short of its unknown best competitor amongst the nutrient-enriched media. It may nevertheless be noted that amongst the nutrient-enriched media only, Test 2 provides forceful evidence that potato-dextrose agar is lower than best within that group. We note that Test 1 has been singled out as the only test of interest from amongst a suite of seven possible shortfall tests. However, for Test 2 the full suite of the six possible shortfall tests involving the six nutrient-enriched media is of interest. We note again that a full suite of shortfall tests always involves at least one redundant test.

In Table 2.3.6 the mean for water agar is straddled by the other entry means. A shortfall test does not presuppose that the population mean for an entry under test may not similarly be straddled by the other population means. For instance, a test based on the contrast:

> $\overline{X}_{\text{aver}\bullet j}$-$\overline{X}_j$,
> where $\overline{X}_{\text{aver}\bullet j}$ denotes the *average* over all of the entries other than Entry *j,*

would be entirely unsuitable for tests of the kind appearing in Tables 2.3.5 and 2.3.6, as an entry can be lower than best without being lower than average.

We remark in passing that Dunnett's *t*-like statistic appears in the literature on decision-making under risk in connection with a variety of different problems, and is also widely – and wrongly – viewed as *necessarily* being subject to certain ideas arising from the doctrine of simultaneous statistical inference. In the subsequent development we will meet and critically examine that view. For the time being, however, we need only note that shortfall testing is not a procedure for decision-making under risk and also does not in any way involve the doctrine of simultaneous statistical inference.

2.4 Some definitions

Various ideas exemplified in the previous section motivate a series of definitions that we now introduce along with further explanations as we proceed. To begin with, note that if a problem in elimination testing involves a vector parameter, interest is often (even usually) in some or other scalar function of that vector. For instance, let the two sets of

waiting times for the eruptions of Vesuvius during the ancient era and the modern era be modelled as two independent, random samples of exponential waiting times (Example 1.32.2). If θ_1 and θ_2 denote the expected waiting times and $R_{1.}$ and $R_{2.}$ the sample totals for the two eras, respectively, ($R_{1.}$, $R_{2.}$) is minimally sufficient for (θ_1, θ_2). Here interest might be in $\theta_1 \div \theta_2 = \theta$, for which $R_{1.} \div R_{2.} = T(\theta)$ is an elimination quantity. This example serves to clarify Definition 2.4.1.

Definition 2.4.1:

Let given data $S_x = \{x_1, x_2, x_3, ..., x_n\}$ be modelled by a random sample $S_X = \{X_1, X_2, X_3, ..., X_n\}$, whose probability is given by a class of models indexed by a vector, $\Theta = (\theta_1, \theta_2, \theta_3, ..., \theta_m)$. Let k transformations:

$r_j\{x_1, x_2, x_3, ..., x_n\}$ for j = 1, 2, 3, ..., k,

define a set of statistics $S_R = \{R_1, R_2, R_3, ..., R_k\}$, that is minimally sufficient for Θ. Let two further transformations:

$\theta = \theta(\theta_1, \theta_2, \theta_3, ..., \theta_m)$ and $t(\theta) = t(r_1, r_2, r_3, ..., r_k, \theta)$,

define a scalar random variable $T(\theta)$, whose value and/or distribution depends on Θ through θ only. Then we call $T(\theta)$ *a (well-conceived) elimination quantity* for θ.

We remark on four aspects of this definition.

Remark 1: The expression 'well conceived' is used to underscore the principle that an elimination quantity may not depend on the class characteristic. The present definition thus provides for elimination tests in the sense of Definition 1.15.1. The point here is that when we judge the quality of fit of alternative members of a class of models in respect of predictions that depend on the index of the class, those predictions must be untrammelled by predictions that do not discriminate amongst different index values.

Remark 2: In order to show that the elimination quantity supplies predictions indexed by θ, the definition denotes the quantity as $T(\theta)$, which is not to say the quantity itself is indexed by θ. Possibly only its distribution is indexed by θ. For instance, let X be an exponentially distributed random variable with expected value θ^{-1} ($0 < \theta < \infty$), in which case X is minimally sufficient for θ. So a well-conceived elimination quantity for θ is $T(\theta) = X$, which is not indexed by θ, but whose distribution,

$$\theta^{-1} e^{-x\theta} \text{ for } 0 < x < \infty,$$

is indexed by θ. In this example an equivalent suite of tests is obtainable via an elimination pivot $T(\theta) = X\theta$, which is itself indexed by θ, but whose distribution,

$$e^{-x} \text{ for } 0 < x < \infty,$$

is independent of θ. So the co-ordinates of X = x, for any given x, are calculable either via X or via $X\theta$, as (1-V, ε, V) where:

$$V = \int_{x}^{\infty} \theta^{-1} \exp(-y\theta) dy = \exp(-t\theta), \text{ or } V = \int_{x\theta}^{\infty} \exp(-y) dy = \exp(-t\theta), \text{ respectively}$$

However, not every elimination quantity is capable of an equivalent elimination pivot. For the elimination quantity used in the number-of-busses problem in Example 2.3.2 for instance, an equivalent elimination pivot seems non-existent. We note in passing that X is a statistic, and Xθ is not a statistic. In general elimination pivots are not statistics.

Remark 3: The definition requires T(θ) to be scalar. The reason for this can be shown by means of Example 2.3.3. It is perfectly possible, in the case of that example, to calculate statistical co-ordinates that will direct us to precisely where within the hypothesised model, indexed by say θ = 11, the mental correlate of the vector datum

$$[x_{(1)}, x_{(3)}] = (16, 19)$$

is to be found. Consider for that purpose Table 2.3.3. It then appears that the required directions partition the probability content of the table into nine parts corresponding to the Cartesian product:

$$[U(x_{(1)}; \theta), \varepsilon(x_{(1)}; \theta), V(x_{(1)}; \theta)] \times [U(x_{(3)}; \theta), \varepsilon(x_{(3)}; \theta), V(x_{(3)}; \theta)] \text{ for } \theta = 11.$$

These nine parts, in a layout corresponding to that of Table 2.3.3, are as follows:

34/120	1/120	0/120
18/120	2/120	1/120
48/120	7/120	9/120

(2.4.1)

In this layout, the column totals:

$$(110/120, 10/120, 10/120) = (0.83, 0.08, 0.08) \text{ are the co-ordinates of } x_{(1)}$$

and the row totals:

$$(35/120, 21/120, 64/120) = (0.29, 0.18, 0.53) \text{ are the co-ordinates of } x_{(3).}$$

These co-ordinates have already been given in Table 2.3.4 as those arising for θ = 11, and the joint co-ordination at (2.4.1) dissipates the evidence conveyed by them. Such dissipation is avoided by gathering evidence of any given sort together into 'a single degree of freedom' so to speak. So we must always try to employ *univariate* ordering of evidence, which in this case is supplied by ordering first the one, and then the other of the two marginal sample patterns.

Remark 4: An elimination quantity T(θ) is brought to mind by a real-world data set. Hence, by definition, we make T(θ) depend on a real-world datum t(θ). This is obviously the case in Examples 2.3.2 and 2.3.3. However, these examples are not exceptional on that account. It is fundamentally important to grasp that co-ordination tests always test models that have arisen in response to and that involve just one particular real-world data set, as that is a fundamental distinction between such tests and hypothesis tests. This is underscored by Definition 2.4.2.

A co-ordinate trace represents an array of self-contained predications. Consider, for instance, the trace listed in Table 2.3.2 and depicted in Figure 2.4.1 by a bar diagram whose subsidiary parts have areas proportional to the frequencies they represent. The human

Definition 2.4.2:

Let T(θ) denote a (well-conceived) elimination quantity for θ. Let t(θ) denote the corresponding real-world elimination datum. For every given value of θ, the distribution of T(θ) is a singleton whose range includes the value t(θ) = t. All the singletons thus formed together constitute *a suite of hypothesised models* indexed by θ. For each member of that suite, the calculable ordered number triplet:

$$C(t; \theta) = [U(t; \theta), \varepsilon(t; \theta), V(t; \theta)],$$

provides the requisite statistical co-ordinates to direct us to just where, within that particular member, the rounding that contains the mental correlate of t(θ), i.e. of the real-world elimination datum, is to be found. For *the suite of tests performed*, we call C(t; θ), when considered as a function of θ, *the co-ordinate trace of the mental correlate*.

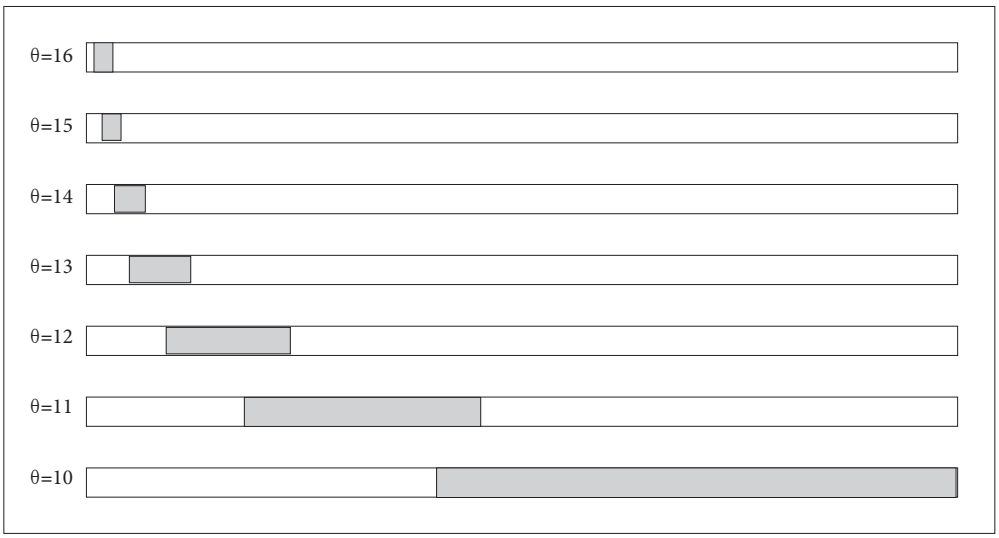


Figure 2.4.1: Bar diagram displaying the values of U (left), ε (centre) and V (right) when tracing the situation of the mental correlate of a given datum $X_{(n)} = 10$, the correlate being transported from model to model indexed by θ = 10, 11, 12, ..., respectively

body is thereby forced to physically grasp the nature of the evidence being put to it because, pointing at Figure 2.4.1, or at a simulated equivalent if needs be, we force the human body to make the corresponding array of predications when we say:

> 'See for yourself how snugly the mental correlate of the test datum is placed within the crowds that $\theta = 10$, $\theta = 11$, and (yes) also $\theta = 12$, bring into the human mind.'
>
> 'See for yourself how the placing of the mental correlate within the crowd that $\theta = 13$ brings into the human mind, is slightly maladroit.'
>
> 'See for yourself how the placing of the mental correlate within the crowd that $\theta = 14$ brings into the human mind, is distinctly maladroit.'
>
> 'See for yourself ...'

The reader should note carefully that the predications are not qualified by probability. Note in particular that we make no qualification of the kind:

'{10, 11, 12, 13} covers θ with coverage probability = ...' (2.4.2)

Herein we are taking the first steps toward ultimately showing that such qualifications rely on poor epistemology, regardless of whether the coverage probability in question is a confidence coefficient, a fiducial probability, or a personal probability.

We usually view a co-ordinate trace as an *ordered* succession of predications. Note, however, that in Table 2.3.5 the ordering is by co-ordinates, as listed in the last column on the extreme right, whereas in Table 2.3.2 the ordering is by index values, as listed in the left-hand column.

An investigator is usually not interested in an entire trace; instead interest will usually be directed at several specially selected terms, as in the following examples.

Example 2.4.1

A much simplified version of the problem leading to the trace shown in Figure 2.4.1, but nevertheless retaining certain essentials for the present discussion, is obtained by considering just n = 1 measurement t, made using a continuous scale with negligible rounding, and modelled by a continuous U(0, θ) random variable T. An observation cannot be modelled as one that could not have occurred. So the members of our class of models are exhaustively indexed by all θ in the interval $t \le \theta < \infty$. With $T(\theta) = T$ as elimination quantity, the trace of the mental correlate of the datum t, is thus given by:

$$[U(t;\theta), \varepsilon(t;\theta), V(t;\theta)] = \left[\frac{t}{\theta}\ ,\ \varepsilon,\ 1-\ \frac{t}{\theta}\right] \text{ for } t \le \theta < \infty, \text{ where } \varepsilon \approx 0 \qquad (2.4.3)$$

This is depicted in Figure 2.4.2, showing how the co-ordinates of the mental correlate

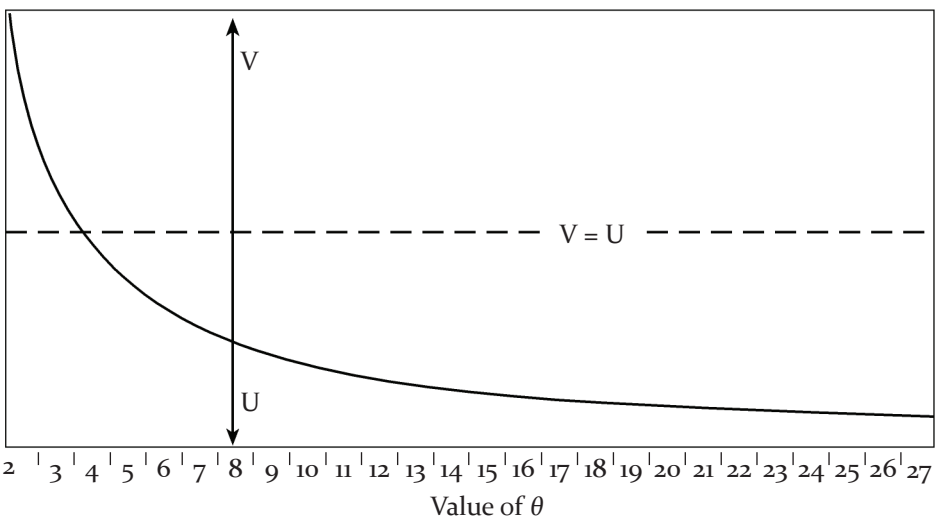


Figure 2.4.2: Tracing the situation of the mental correlate of a given datum T = 2, when modelled as a sample from a U(0, θ) population for various θ, where $2 \le \theta < \infty$

range from (1, ε, 0) at θ = t toward (0, ε, 1) as θ increases. An abbreviated account of the trace is obtained by specifying certain co-ordinate values and solving for θ. Thus, for instance, with t = 1.8, we obtain the following abbreviated trace:

(0.98, ε, 0.02) and (0.02, ε, 0.98) are attained at θ = 1.84 and 90.0, respectively
(0.96, ε, 0.04) and (0.04, ε, 0.96) are attained at θ = 1.88 and 45.0, respectively
(0.92, ε, 0.08) and (0.08, ε, 0.92) are attained at θ = 1.96 and 22.5, respectively.

This abbreviation has been tailor-made by specifying co-ordinates of special interest and then solving for the corresponding index values. In other cases we might specify index values of special interest and then solve for the corresponding co-ordinates. The two possibilities might be expressed by saying that we can specify certain *elimination levels* of special interest and then solve for the corresponding *elimination bounds*, or *vice versa*. We note that, beginning at θ = t, the pointing co-ordinate of the trace given at (2.4.3) is on the right until the elimination level,

(0.50, ε, 0.50), is attained at θ = 3.6

Thereafter the pointing co-ordinate is on the left. The pointing co-ordinate in the trace given in Table 2.3.2 and depicted in Figure 2.4.1 is on the left for every index value. We might express this by saying that the trace depicted in Figure 2.4.1 arises from 'a suite of inherently one-sided tests', whereas the trace depicted in Figure 2.4.2 arises from 'a suite of inherently two-sided tests'. But we will prefer to describe such tests as *unilateral* and *bilateral*, respectively. At (2.4.3) we see yet again that a suite of models and the resulting trace are always, by definition, data dependent.

Example 2.4.2

Snedecor (1956, p.57) gives a data set concerning the efficacy of an insecticide spray for control of the European corn borer. Corn yields, from both sprayed and unsprayed strips in each of 14 fields of corn on different farms, were recorded (Table 2.4.1).

Table 2.4.1: Yields of corn (bushels/acre) from sprayed (S) and unsprayed (U) strips on each of 14 different farms in Boone County, Iowa, 1950 (D = S-U).

S	64.3	78.1	93.0	80.7	89.0	79.9	90.6	102.4	70.7	106.1	107.4	74.0	72.6	69.5
U	70.0	74.4	86.6	79.2	84.7	75.1	87.3	98.8	70.2	101.1	83.4	65.2	68.1	68.4
D	–5.7	3.7	6.4	1.5	4.3	4.8	3.3	3.6	0.5	5.0	24.0	8.8	4.5	1.1

The mean difference in yield (sprayed minus unsprayed) is 4.7 bu./acre with an estimated standard error of 1.73 bu./acre. Let δ denote the expected mean difference. Pesticides are routinely tested on non-infested material of the kind to be protected, thus to ensure against deleterious consequences. This was evidently also the case here as Snedecor, apart from using a notion different to our δ, states:

> 'It had already been established that the spray, at the concentration used, could not decrease yield. ... Consequently, if δ is not zero then it must be greater than zero.'

Following Snedecor, we employ Student's t as an elimination pivot for δ by referring $(4.7\text{-}\delta)\div 1.73$ to Student's test distribution on 13 df. Snedecor calculates that in order to break even, the cost of spraying requires a yield increase of about 2 bu./acre. So, the question: 'Does spraying seem to increase yield?' asks whether or not a good fit needs a hypothesised value > 0 bu./acre, and the question: 'Does spraying seem to pay?' asks whether or not a good fit needs a hypothesised value > 2 bu./acre. We find that:

$$\text{at } \delta = 0 \text{ bu./acre the trace attains } (U, \varepsilon, V) = (0.992, \varepsilon, 0.018), \tag{2.4.4}$$

which evidence strongly points at larger δ values, and

$$\text{at } \delta = 2 \text{ bu./acre the trace attains } (U, \varepsilon, V) = (0.913, \varepsilon, 0.087), \tag{2.4.5}$$

which evidence points at yet larger δ values, but not strongly so. Note that, in contrast to the previous example, the abbreviated trace at (2.4.4) and (2.4.5) was tailor-made by specifying the *elimination bounds* of special interest, rather than by specifying *co-ordinate levels* of special interest.

Example 2.4.3

A visitor to the Kgalagadi Transfrontier National Park learns from Maclean (1985) of a 'rare' yellow-breasted variant of the crimson-breasted shrike. Subsequently, during a game viewing drive, the visitor spots n = 11 birds of that species, all crimson. Modelling the number of yellow individuals as a binomial count T = 0, with $E(T) = n\theta$ when the population proportion of yellow birds is denoted by θ, the trace of the mental correlate of the test datum is found to be given by:

$$[\emptyset, (1\text{-}\theta)^{11}, 1\text{-}(1\text{-}\theta)^{11}] \text{ for } 0 < \theta < 1.$$

We note again that a trace depends on a real-world data set. Here interest would be in a few selected terms of the trace, perhaps specified by way of the elimination levels

$$(\emptyset, 0.04, 0.96), (\emptyset, 0.06, 0.94), (\emptyset, 0.08, 0.92)$$

corresponding to which the elimination bounds are:

$$\theta = 0.25, 0.23, 0.21, \text{ respectively.}$$

As the value of θ is thereby bounded from above by about 1 in 5, and as 1 in 5 could hardly be interpreted as rare, the statistical analysis does not add to what is already known according to Maclean. We note in passing that if Maclean's information is not based on records capable of statistical analysis, the visitor's two sources of information are incapable of being combined in a physically meaningful way. Such situations, as we shall see later, have prompted the introduction of 'Bayesian' recipes for combining two such sources of information in terms of certain *metaphysical* notions of probability, i.e. notions of probability whose meanings are incapable of being grasped by the human body. But, for the time being, that need not concern us.

Example 2.4.4

We revisit Example 2.3.4 and note that reversing the sign of the parameter defined at (2.3.4) expresses the gain or loss when Beauty is substituted for its unknown best contender named θ. So let us call that gain or loss 'Beauty's *substitution value*', and let it be denoted by:

$$\Delta_{Beauty} = \mu_{Beauty} - \mu_{\theta}, \text{ where } \Delta_{Beauty} = -\delta_{Beauty}.$$

Inspection of Table 2.3.5 indicates that Beauty's substitution value may well be -11, as that equals the difference between Beauty's mean performance (judged lower than best within the group) and Juicy Lucy's mean performance (not judged lower than best within the group). In general, a shortfall pivot for Beauty's substitution value is obtained by expressing the many-one t statistic at (2.3.4) in a general form, as follows, where Δ_{Beauty} denotes Beauty's substitution value:

$$t' = \frac{\overline{X}_{\max \bullet Beauty} - (\overline{X}_{Beauty} - \Delta_{Beauty})}{\sqrt{\frac{2s^2}{n}}}.$$

The test at (2.3.6) is thus interpretable as a test of:

$$M_0: \Delta_{Beauty} = 0 \text{ versus } M_1: \Delta_{Beauty} < 0.$$

Similarly, to test say:

$$M_0: \Delta_{Beauty} = -10 \text{ versus } M_1: \Delta_{Beauty} < -10,$$

we calculate the test datum:

$$t' = [+18+(-10)] \div (4.99 \times \sqrt{2}), \text{ i.e., } t' + 1.13 \text{ on 12 df. with three competitors}$$

whose mental correlate is situated at (0.71, ε, 0.29) in the test distribution. Recall that the co-ordinate thus obtained is the *leftmost* of the possible right co-ordinates under the hypothesised model. So, for leftmost right co-ordinates here specified by:

$$(1\text{-}0.08, \varepsilon, 0.08), (1\text{-}0.04, \varepsilon, 0.04), (1\text{-}0.02, \varepsilon, 0.02), \tag{2.4.6}$$

the corresponding upper bounds for Beauty's substitution value are the solutions of:

$$t' = [+18+\Delta_{Beauty}] \div (4.99 \times \sqrt{2}) = 1.98, 2.41, 2.83, \text{ respectively.}$$

These solutions turn out to be:

$$\Delta_{Beauty} = -4, -1, +2, \text{ respectively.} \tag{2.4.7}$$

For Juicy Lucy, the corresponding upper bounds are the solutions of:

$$t' = [+7+\Delta_{Juicy\ Lucy}] \div (4.99 \times \sqrt{2}) = 1.98, 2.41, 2.83, \text{ respectively,}$$

which turn out to be:

$$\Delta_{Juicy\ Lucy} = +7, +10, +13, \text{ respectively.} \tag{2.4.8}$$

The abbreviated trace conveyed at (2.4.6) and (2.4.7) jointly, underscores the result of the test appearing in the last line of Table 2.3.5. The abbreviated trace conveyed at (2.4.6) and (2.4.8) jointly, discourages the view that on the evidence given in Table 2.3.5 Juicy Lucy may be discarded in favour of Money Maker. Terminology now introduced by Definition 2.4.3 underscores the tailor-made nature of trace abbreviation.

Definition 2.4.3:

A tailor-made trace comprises *several* of, rather than *all*, the terms of a co-ordinate trace arising from testing a suite of models for a given real-world data set. The terms of such an abbreviation are tailor-made in the sense of having been selected to serve the purposes of the particular investigation. A tailor-made trace comprises particular index values that we call *elimination bounds.* We call their corresponding co-ordinate values, *elimination levels*.

Note that the definition emphasises 'several', that is to say, 'more than one'. A tailor-made trace is the co-ordination tester's answer to the idea indicated at (2.4.2). Thus, a data analyst does not bundle together a motley of models ranging from some that fit the given data well to others that fit the given data poorly, so as to describe the motley by a just one number called a confidence coefficient, or a credibility coefficient, or a fiducial probability, etc. It should be perfectly obvious from the outset that such bundling together is symptomatic of poor epistemology, of epistemology that at best is marching to the beat of a mistaken drum.

2.5 One-to-one transformations of the index

Theorem 2.5.1 states a useful fact about one to one transformations of the parameter space.

Theorem 2.5.1:

Let $T(\theta)$ denote a well-conceived elimination quantity for θ, the index of a suite of hypothesised models. Let $\Omega(\theta)$ denote the range of θ. Let:

$C(t; \theta) = [U(t; \theta), \varepsilon(t; \theta), V(t; \theta)]$ for $\theta \in \Omega(\theta)$

denote the trace of the mental correlate of the given datum $t(\theta) = t$. Let $\phi = \phi(\theta)$ be any one to one transformation of θ. Then $C(t; \phi) = C(t; \theta)$ identically in θ.

The theorem is obvious, as the one to one transformation simply re-labels every singleton in the suite of hypothesised models without altering the situation of the mental correlate within that singleton. This enables useful transformation of elimination bounds, as in the following examples.

Example 2.5.1

In order to estimate population size, a game farmer catches, marks and releases 104 springbok. A subsequent survey finds eight marked animals amongst 884 spotted while traversing the farm. Let N denote the total number of animals, of which M = 104 are marked, and let n = 884 denote the number subsequently spotted, of which m = 8 are

marked. Solving from m/n ≈ M/N, we find N ≈ 11500 animals, indicating a negligibly small sampling fraction, 884/11500 = 0.08. So, we may consider binomial sampling as an approximate model for the given data set. Using the normal approximation for the binomial, the U, V = 0.050, 0.075, 0.100 elimination bounds for M/N are given by:

$$(m/n)\text{-}Z\sqrt{(m/n)[1\text{-}(m/n)]/n} \le M/N \le (m/n)\text{+}Z\sqrt{(m/n)[1\text{-}(m/n)]/n},$$

with Z = 1.645, 1.439, 1.282, respectively. These bounds are:

$$0.00381 \le M/N \le 0.0143$$
$$0.00447 \le M/N \le 0.0136$$
$$0.00497 \le M/N \le 0.0131, \text{ respectively.} \qquad (2.5.1)$$

As $(M/N)^{-1} \times M$ is a one to one transformation from M/N to N, we apply this transformation to the bounds given at (2.5.1) to obtain the following elimination bounds for N:

7 300 animals ≤ N ≤ 27 000 animals, at U, V = 0.050
7 600 animals ≤ N ≤ 23 000 animals, at U, V = 0.075
7 900 animals ≤ N ≤ 21 000 animals, at U, V = 0.100

In order to quarter the length of the intervals spanned by such bounds at the specified elimination levels, the farm must be traversed independently about 16 times.

Example 2.5.2

Saunders and Rayner (1951, p. 13) present data giving the number of suckers per plant for n = 600 maize plants from each of two different susceptible strains (Table 2.5.1).

Table 2.5.1: Numbers of suckers per plant for n = 600 maize plants from each of two different susceptible strains

Number of suckers per plant	0	1	2	3
Number of plants of strain 1	446	130	22	2
Number of plants of strain 2	507	85	7	1

They note that such data is often satisfactorily modelled in terms of Poisson sampling, stating that apparently 'the various influences tending to produce a sucker in these strains of maize each operate with very small probability of producing a successful event, i.e. a single sucker.' In order to try out this suggestion by way of a commencement test, let us model their data set as two independent Poisson samples of size n = 600 counts each, with the expected number of suckers per plant denoted by μ_j for Strain J (J = 1, 2). The joint probability of two such samples then factors similarly to the factoring displayed at (1.8.1), except that in this case there are four factors, say as follows in an abbreviated notation:

$$[\Phi\{\mu_1\}\times\Gamma_1]\times[\Phi\{\mu_2\}\times\Gamma_2], \qquad (2.5.2)$$

where the Φ-like factors correspond to the first of the two factors at (1.8.1) and the Γ-like factors correspond to the second of the two factors at (1.8.1). In order to test the quality of fit of the Γ-like factors, we might order each population of samples by the magnitude of the variance-to-mean ratio (Example 1.11.1). The calculated chi-square values for the two tests are 606.7 and 615.6, each on 599 df. The mental correlates of the datum ratios are then found to be situated in the test statistic at approximately:

$$(0.59, \varepsilon, 0.41) \text{ and } (0.55, \varepsilon, 0.45), \text{ for Strains 1 and 2, respectively.} \qquad (2.5.3)$$

By these tests the Poisson characteristic fits each of the two data sets very well.

Turning now to the Φ-like factors at (2.5.2), the expression at (1.8.1) indicates how the total counts are modelled as the realisations of two Poisson random variables, denoted by X_1 and X_2 for Strains 1 and 2, respectively. (X_1, X_2) is then minimally sufficient for (μ_1, μ_2) and the data in hand indicate that a good fit requires values of μ_1 in excess of those of μ_2. In order to investigate this, we transform as follows:

$$\rho = \mu_1 \div \mu_2 \text{ and } \mu_\bullet = \mu_1 + \mu_2,$$

where ρ is the parameter of interest and $\mu_\bullet$ is a nuisance parameter. The probability of $(X_1, X_2) = (x_1, x_2)$, previously given as $\Phi\{\mu_1\} \times \Phi\{\mu_2\}$, is then transformed, in terms of $X_1 + X_2 = X_\bullet$ and $X_1 = X$, to the probability of $(X_\bullet, X) = (x_\bullet, x)$, which is given by:

$$\left[\frac{e^{-\mu_\bullet} \mu_\bullet^{x_\bullet}}{x_\bullet!} \right] \times \left[\binom{x_\bullet}{x} \left(\frac{\rho}{1+\rho} \right)^{x} \left(1 - \frac{\rho}{1+\rho} \right)^{x_\bullet - x} \right],$$

$$\text{where } x_\bullet = 0, 1, 2, \ldots, \text{ and } x = 0, 1, 2, \ldots, x_\bullet. \qquad (2.5.4)$$

The essence of the expression at (2.5.4) is that, conditional on $X_\bullet = x_\bullet$, we may treat X as a binomial random variable for a sample of size $x_\bullet$ with probability of success equal to $\rho \div (1+\rho)$. In our case the data are now:

$$X = 0(446)+1(130)+2(22)+3(2) = 180 \text{ successes out of just}$$
$$x_\bullet = 0(446+507)+1(130+85)+2(22+7)+3(2+1) = 282 \text{ trials}$$

and our suite of models comprises binomials indexed by $\rho \div (1+\rho)$ over the range $0 < \rho \div (1+\rho) < 1$. In order to obtain the usual $N(\mu, \sigma^2)$ approximation for the binomial total, we must take:

$$\mu = x_\bullet \left(\frac{\rho}{1+\rho} \right) \text{ and } \sigma^2 = x_\bullet \left(\frac{\rho}{1+\rho} \right) \left[1 - \left(\frac{\rho}{1+\rho} \right) \right] = \frac{x_\bullet \rho}{(1+\rho)^2}.$$

Using Yates's correction for continuity, the approximation then gives the required right co-ordinate as the frequency with which a N(0, 1) variable exceeds:

$$z = [x - (x_\bullet - x)\,\rho + 0.5] \div \sqrt{x_\bullet \rho}, \text{ in our case } [180 - 102\rho + 0.5] \div \sqrt{282\rho}.$$

For instance, to test whether a good fit is obtained with μ_1 twice the size of μ_2, we test M_0: $\rho = 2$, finding $z = -0.990$ and $V = 0.84$. In order to evaluate the rounding we must repeat this procedure, after subtracting (rather than adding) 0.5 to the numerator of

the expression for z. We thus find z = -1.03 giving ε = 0.85-0.84 = 0.01. The requisite co-ordinates are thus found to be approximately (0.15, 0.01, 0.84). By the test performed, the model indexed by $\rho = 2$ fits the given data well. Alternatively, we may specify certain interesting values of V and establish the corresponding values of ρ. For instance, to obtain (U, ε, V) = (0.05, ε, 0.95) we must take z = -1.645. Squaring the expression for z, we must then find the roots of the quadratic equation:

$$(102)^2\rho^2-[2(180.5)(102)+(-1.645)^2(282)]\rho+(180.5)^2 = 0.$$

The roots are $\rho = 1.44$ and $\rho = 2.17$, of which the larger root is the one asked for, and the smaller root is for (U, ε, V) = (0.95, ε, 0.05), i.e. these two roots are the elimination bounds at the U, V = 0.05 levels of elimination.

We note in passing that, in order to justify testing alternative values of ρ thus, the two tests leading to the co-ordinates at (2.5.3) should strictly speaking be replaced by testing the compound consisting of both class characteristics (Section 1.32). Such a test is provided by the sum of the two chi-square values previously obtained, i.e.:

chi-square = 1222.3 on 599+599 df.,

whose mental correlate is situated close to the median of the test distribution.

2.6 Unilateral and bilateral tests of co-ordination

Hypothesis testing can often be specified to be either 'one-sided repetitive testing' or 'two-sided repetitive testing'. For instance, let X denote a $N(\mu, 1)$ random variable, where $-\infty < \mu < +\infty$, and where we require an accept-reject rule for testing $H_0: \mu = 0$ subject to a Type I error rate specified to be $\alpha = 0.05$ in repetitive testing. Consider the following three different ways of specifying the alternative:

$$H_1: \mu < 0. \quad H_1: \mu = \pm\delta \neq 0. \quad H_1: \mu > 0.$$

The matching three decision rules for minimising the Type II error rate in repetitive testing are reject H_0 and accept H_1 if and only if:

$$X \leq -1.645. \quad |X| \geq 1.96. \quad X \geq +1.645.$$

The first and the last of these three rules exemplify one-sided hypothesis testing, and the second exemplifies two-sided hypothesis testing. Now note (and this is crucial) whether we then perform one-sided testing or two-sided testing *is a matter of choice*. One must *specify* whether the repetitive testing is to be one-sided or two-sided. As opposed to that, two different suites of co-ordination tests sometimes involve a superficially similar, but in fact fundamentally different distinction. In order to make this clear the following example is introduced for comparison to Example 2.4.3.

Example 2.6.1

A visitor to the Kruger National Park learns from Maclean (1985) of a 'rare' yellow-headed variant of the black-collared barbet, usually red headed. Subsequently, during a game

viewing drive, the visitor spots n = 11 birds of that species, all but one of them red-headed. Modelling the number of yellow-headed individuals as a binomial count, T = 1, with E(T) = nθ when the population proportion of such birds is denoted by θ, the trace of the mental correlate of the datum count is given by:

$$[(1-\theta)^{11},\ 10(1-\theta)^{11-1}\theta^{1},\ 1-(1-\theta)^{11}-10(1-\theta)^{11-1}\theta^{1}] \text{ for } 0 < \theta < 1.$$

When θ increases from zero, we find the index being bounded from below at round about θ = 0.0037, 0.0056, 0.0076, the corresponding levels attained by the trace being:

(0.960, 0.036, 0.004), (0.940, 0.053, 0.007), (0.920, 0.070, 0.010), respectively.

For comparison to Example 2.4.3, note that here ε+V = 0.04, 0.06, 0.08, respectively. The quality of fit steadily improves as θ increases, until the best-fitting models are found at round about θ = 0.097, at which point the mental correlate is being placed at:

(0.326, 0.350, 0.324) in the test distribution.

Thereafter the quality of fit deteriorates as θ increases, until we find the index being bounded from above at about θ = 0.32, 0.34, 0.39, the corresponding levels attained by the trace being:

(0.01, 0.07, 0.92), (0.01, 0.05, 0.94), (0.01, 0.03, 0.96), respectively.

For comparison to Example 2.4.3, note that here U+ε = 0.08, 0.06, 0.04, respectively.

Comparison of Examples 2.4.3 and 2.6.1

In both examples the elimination quantity used is the total number of successes in just n = 11 Bernoulli trials, but the results are two very different traces. In Example 2.4.3 the tests are *inherently* one sided, as the co-ordinates of the mental correlate of the test datum T = 0 range from just right of (Ø, 1, 0) to just left of (Ø, 0, 1) when θ ranges from just more than θ = 0 to just less than θ = 1. So, as large rounding fails to provide informative directions to just where within the test distribution the mental correlate of the test datum is to be found, an informative co-ordination can here only take the form:

(Ø, 1-V, V) for small 1-V, thus pointing at smaller values of θ.

In Example 2.6.1, however, the tests are *inherently* two sided, as the co-ordinates of the mental correlate of the test datum T = 1 range from just right of (1, 0, 0) to just left of (0, 0, 1) when θ ranges from just more than θ = 0 to just less than θ = 1. So in that example an extreme co-ordination can take one of the two different forms:

(U, ε, 1-ε-U) for small U and ε, thus pointing at smaller values of θ, or

(1-ε-V, ε, V) for small ε and V, thus pointing at larger values of θ.

Instead of describing a trace as an inherently one-sided trace, let us rather describe it as a unilateral trace, as that is more concise and (more importantly) helps to combat the importation of inappropriate ideas from the theory of hypothesis testing. So let us avoid the terms 'one-sided' and 'two-sided' and, instead using the terms unilateral and bilateral, respectively, defined as in Definition 2.6.1.

Definition 2.6.1:

Let a real-world data set bring into the human mind a well-conceived elimination quantity for θ, the index of a suite of hypothesised models. Let θ range over an ordered space. Consider the trace of the mental correlate of the test datum.

If every term of the trace points at θ larger than the hypothesised value, we call the trace a *lower-bounding unilateral trace*. Such a trace comprises an array of *lower-bounding unilateral tests of co-ordination*.

If every term of the trace points at θ smaller than the hypothesised value, we call the trace an *upper-bounding unilateral trace*. Such a trace comprises an array of *upper-bounding unilateral tests of co-ordination*.

If there exists θ_1, θ_2, $\theta_1 \leq \theta_2$, such that for any hypothesised value $< \theta_2$ the trace points at θ larger than that value, and for any hypothesised value $> \theta_1$ the trace points at θ smaller than that value, we call the trace a *bilateral trace*. Such a trace comprises an array of *bilateral tests of co-ordination*.

In Examples 2.4.3 and 2.6.1 the traces are unilateral and bilateral, respectively, owing to the particular data involved despite arising from the selfsame test statistic. In other cases such distinctions may be owing to the test statistic. In Table 2.3.4 for instance, we find a unilaterally lower-bounding trace arising from the $X_{(1)}$ test statistic, and a unilaterally upper-bounding trace arising from the $X_{(3)}$ test statistic, where that would be the case for any appropriate data. Again, at (2.4.3) we find a formula for a bilateral trace, no matter what appropriate data might be under consideration. In the case of the trace appearing in Table 2.3.5, the foregoing definition is not applicable, as the index in that case ranges over the unordered set of identities given at (2.3.7). Definition 2.6.2 now arises.

Definition 2.6.2:

Let a real-world data set bring into the human mind a well-conceived elimination quantity T(θ) for θ, the index of a suite of hypothesised models. Let θ range over an ordered space. Consider the trace of the mental correlate of the test datum.

If the trace is a lower-bounding unilateral trace, regardless of the value of the test datum, we call T(θ) a *strictly lower-bounding elimination quantity*.

If the trace is an upper-bounding unilateral trace, regardless of the value of the test datum, we call T(θ) a *strictly upper-bounding elimination quantity*.

If the trace is a bilateral trace, regardless of the value of the test datum, we call T(θ) a *bilateral elimination quantity*.

In this definition, as in the previous definition, the phrase 'elimination quantity' may of course be replaced by the phrase 'elimination pivot'.

2.7 Some principles of shortfall testing

In this section we develop a platform from which we can, in the next section, remove certain sources of possible confusion arising from the literature on Dunnett's many–one-*t* statistic. Here it will suffice to consider that statistic in the case of a known error variance, which can be formulated very simply, as follows: let a real-world data set,

$$\{x_1, x_2, x_3, ..., x_k\}$$

comprising the yields of k different entries in a yield trial, bring into the human mind a corresponding set of independently and continuously distributed random variables:

$$\{X_1, X_2, X_3, ..., X_k\}$$

with corresponding density functions:

$$\{f_1(x_1-\mu_1), f_2(x_2-\mu_2), f_3(x_3-\mu_3), ..., f_k(x_k-\mu_k)\} \quad (2.7.1)$$

fully specified apart from being indexed by corresponding location parameters:

$$\{\mu_1, \mu_2, \mu_3, ..., \mu_k\}, \text{ for } -\infty < \mu_1, \mu_2, \mu_3, ..., \mu_k < +\infty.$$

We wish to test whether or not any particular entry may be lower than best, where it will be convenient to label that particular entry as Entry 1. The apparent shortfall of Entry 1 is calculated as:

$x_{max\bullet 1}-x_1$ in the notation of Example 2.3.4.

Conceptual calculations bring into the human mind a corresponding random variable:

$X_{max\bullet 1}-X_1$.

Consider the statistical co-ordinates that direct us to just where within the $X_{max\bullet 1}-X_1$ distribution the mental correlate of the $x_{max\bullet 1}-x_1$ datum is to be found. For the present purposes we may consider a near-to-zero statistical rounding. The right co-ordinate then suffices to provide the required directions and is given by:

$$\Pr(X_{max\bullet 1}-X_1 > x_{max\bullet 1}-x_1) = 1-\Pr(X_{max\bullet 1}-X_1 \leq x_{max\bullet 1}-x_1). \quad (2.7.2)$$

The expression on the right is spelled out more explicitly as:

$$1-\Pr(X_j-X_1 \leq x_{max\bullet 1}-x_1 \text{ for all of } j = 2, 3, 4, ..., k). \quad (2.7.3)$$

As $X_1, X_2, X_3, ..., X_k$ are independently distributed, their joint density is given by:

$$\prod_{j=1}^{k} f_j(x_j-\mu_j). \quad (2.7.4)$$

Consider the transformation:

$$X_1-\mu_1 = Y \text{ and } X_j-X_1 = Y_j \text{ for } j = 2, 3, 4, ..., k$$

and the corresponding notations:

$$(x_1-\mu_1) = y \text{ and } (x_j-x_1) = y_j \text{ for } j = 2, 3, 4, ..., k.$$

The Jacobian of the transformation equals unity. So, owing to the identity:

$$x_j-\mu_j = (x_j-x_1)+(x_1-\mu_1)+\mu_1-\mu_j = y_j+y+\mu_1-\mu_j,$$

it follows from the expression at (2.7.4) that the joint density function of Y and Y_j for $j = 2, 3, 4, ..., k$ is given by:

$$f_1(y)\times\left[\prod_{j=2}^{k} f_j(y_j+y+\mu_1-\mu_j)\right]. \qquad (2.7.5)$$

It also follows from the equation at (2.7.2) and the expression at (2.7.3) that:

$$\Pr(X_{\max\bullet 1}-X_1 > x_{\max\bullet 1}-x_1) = 1-\Pr(Y_j \le x_{\max\bullet 1}-x_1 \text{ for all of } j = 2, 3, 4, \ldots, k) \qquad (2.7.6)$$

Let the cumulative distribution functions corresponding to the density functions at (2.7.1) be denoted by:

$$\{F_1(x_1-\mu_1), F_2(x_2-\mu_2), F_3(x_3-\mu_3), \ldots, F_k(x_k-\mu_k)\}. \qquad (2.7.7)$$

Then it follows from the expression at (2.7.5) and the equation at (2.7.6) that:

$$\Pr(X_{\max\bullet 1}-X_1 > x_{\max\bullet 1}-x_1) = 1-\int_{-\infty}^{+\infty}\prod_{j=2}^{k} F_j[(x_{\max\bullet 1}-x_1)+y+(\mu_1-\mu j)]\,\mathrm{d}\,F_1(y). \qquad (2.7.8)$$

Example 2.4.4 introduced the notion of the substitution value of an entry under test as being the gain or loss in yield potential, should that entry be substituted for its unknown 'best' competitor. Define:

$$\mu_{\max\bullet 1} \text{ as the largest value in the set of mean values } \{\mu_2, \mu_3, \mu_4, \ldots, \mu_k\}. \qquad (2.7.9)$$

Then the substitution value of Entry 1 is given by Δ_1 as defined by:

$\mu_1 = \mu_{\max\bullet 1}+\Delta_1$.
If $\Delta_1 > 0$, Entry 1 is a sole best entry.
If $\Delta_1 = 0$, Entry 1 is one of two or more best entries.
If $\Delta_1 < 0$, Entry 1 is lower than best.

The right-hand side of the equation at (2.7.8) can now be expressed as:

$$1-\int_{-\infty}^{+\infty}\prod_{j=2}^{k} F_j[y+(x_{\max\bullet 1}-x_1)+\Delta_1+(\mu_{\max\bullet 1}-\mu j)]\,\mathrm{d}\,F_1(y), \qquad (2.7.10)$$

this being the right statistical co-ordinate directing us to just where in the $X_{\max\bullet 1} - X_1$ distribution the mental correlate of the $x_{\max\bullet 1}-x_1$ datum is to be found.

Before proceeding further, it must be underscored that the derivation of the co-ordinate in (2.7.10) is mathematically correct beyond any reasonable contest. This is important, as we will have to concern ourselves with different interpretations of the result. So, note that the derivation is by a well-known method for obtaining the distributions of order-statistical quantities. In fact, the mathematical result as such has in one form or another appeared in the statistical literature for close on a half a century. In the form given here, it goes back at least as far as Gupta (1965).

Now, in order to grasp what the co-ordinate tells us, three facts must be borne in mind:

(1) $x_{max \bullet 1}$-x_1 is a constant and y is a bound variable at (2.7.10), whereas the values the co-ordinate may take, depend on disposable variables only with the disposable variables being the substitution value of the entry under test Δ_1 and the nuisance parameters ($\mu_{max \bullet 1}$-μ_j) for j = 2, 3, 4, ..., k.

(2) ($\mu_{max \bullet 1}$-μ_j) ≥ 0 for all of j = 2, 3, 4, ..., k, owing to the definition in (2.7.9).

(By the way, $\mu_{max \bullet 1}$ does not denote the expected value of $X_{max \bullet 1}$.)

(3) The functions defined at (2.7.7) are distribution functions, and any distribution function is non-decreasing. So, the term on the right at (2.7.10) will decrease, thus causing the value of the co-ordinate to increase whenever the value of any one or more of the disposable variables decreases, and *vice versa*.

It appears at once from the expression given at (2.7.10) that if, for any specified substitution value Δ_{10} we wish to test:

M_0: $\Delta_1 = \Delta_{10}$ versus M_1: $\Delta_1 > \Delta_{10}$,

the leftmost of the possible values of the hypothesised right co-ordinate arises when the values of all of the k-1 nuisance parameters are zero (as small as possible). So, the leftmost right co-ordinate is obtained from the result at (2.7.10) and given by:

$$1- \int_{-\infty}^{+\infty} \prod_{j=2}^{k} F_j [y+(x_{max \bullet 1}-x_1)+\Delta_{10}]d\, F_1(y), \qquad (2.7.11)$$

If this leftmost right co-ordinate is small, the actual right co-ordinate is equally small or smaller. And if that is judged as untenably small, the term on the right at (2.7.11) is judged as untenably large, implying that Δ_{10} is judged as untenably large. In the case of Δ_{10} = 0, that would imply that Entry 1 is judged lower than best. There is nothing whatsoever to prevent us from then also testing whether Entry 2 may be lower than best, whether Entry 3 may be lower than best, whether Entry 4 may be lower than best, ..., whether Entry k may be lower than best. We will thus obtain a trace that, like any other trace, comprises the outcomes of an array of different tests of co-ordination. The point here is that results of the kind displayed in Table 2.3.5 do not involve the notion of simultaneous statistical inference. On the contrary, they are the results of separate tests of co-ordination, each test possessing its own integrity. This is underscored by Example 2.4.4, showing that each test concerns the substitution value of one particular entry only. Note for instance that when writing:

$X_{max \bullet 1}$ as in the foregoing, or $X_{max \bullet Beauty}$ as in Example 2.4.4,

the notations 'max•1' and 'max•Beauty' do not identify any of the entries under test. Instead, they identify certain *outcomes* where *different* entries may be involved from outcome to outcome. In Example 2.4.4, for instance, the mean yield identified as the 'max•Beauty mean', may be the mean of Juicy Lucy in one sample, that of Bonny Best in the next sample, and so forth.

We have shown how the substitution value of Entry 1 is bounded from above by finding, within the $X_{max\bullet 1}$-X_1 distribution, the leftmost of those positions that may be occupied by the mental correlate of the $x_{max\bullet 1}$-x_1 datum. We now consider whether the substitution value of Entry 1 may be bounded from below by finding, within the $X_{max\bullet 1}$-X_1 distribution, the rightmost of those positions that may be occupied by the mental correlate of the $x_{max\bullet 1}$-x_1 datum. In the case of a negligible statistical rounding, as here being considered, the expression at (2.7.10) shows that the left statistical co-ordinate directing us to just where in the $X_{max\bullet 1}$-X_1 distribution the mental correlate of the $x_{max\bullet 1}$-x_1 datum is to be found, is given by

$$\int_{-\infty}^{+\infty} \prod_{j=2}^{k} F_j\,[y+(x_{max\bullet 1}-x_1)+\Delta_1+(\mu_{max\bullet 1}-\mu_j)]\,d\,F_1(y). \qquad (2.7.12)$$

So the value of the rightmost left co-ordinate for any fixed value of Δ_1 is obtained by inserting into the expression at (2.7.12) the largest possible values of the k-1 nuisance parameters denoted by

$(\mu_{max\bullet 1}-\mu_j)$ for j = 2, 3, 4 ..., k.

However, as these parameters are not bounded from above, the value of the rightmost left co-ordinate is then vacuously given by 1 (one). The point here is simply this: the elimination quantities used for shortfall tests are strictly upper-bounding elimination quantities. Traces of the type arising from a suite of shortfall tests for the substitution value of any one particular entry, for instance the abbreviated traces of Example 2.4.4, are upper-bounding unilateral traces.

In order to relate the foregoing development to Dunnett's many-one-*t*, we note that the '∞ degrees of freedom' form of the expression at (2.3.5) is given by:

$$t' = \frac{\overline{X}_{max\bullet Beauty} - \overline{X}_{Beauty}}{\sqrt{\frac{2\sigma^2}{n}}}.$$

Taking n = 1, σ^2 = 0.5, and Entry 1 = Beauty, we find

$t' = X_{max\bullet 1}$-X_1, as developed in the present section.

The requisite mathematics for extending the development to any fixed-model analysis of variance situation, as required for Examples 2.3.5, 2.3.6 and 2.4.4, is well established. The reader is referred to Gupta (1965), or Gupta and Panchapakesan (1979), or, for a treatment closer to the present development, Van Aarde (1994).

2.8 Shortfall tests *vis-à-vis* Dunnett's 'multiple comparison' procedures

Often the selfsame test statistic arises in otherwise unrelated contexts, a case in point being Snedecor's F statistic arising in Example 1.32.2 – an example that otherwise has little, if anything, in common with analysis of variance. Moreover, unrelated contexts may involve formal similarities that sow confusion, a case in point being widespread failure to grasp the distinction between significance tests and hypothesis tests. So, in this section, we wish to make it clear that shortfall tests, despite involving Dunnett's many–one-*t* statistic, are not to be confused with certain procedures for 'simultaneous statistical inference' proposed by Dunnett (1955) and involving the same test statistic. We do so by way of two contrasting examples, followed by a comparative discussion.

Example 2.8.1

Consider the following problem in decision-making under risk: an extensive array of sets, each comprising a standard item, and of k-1 other items to be tested against the standard, is brought into the real world. The performance of each item is measured, and within each set of items the performance of each non-standard item (test item) is compared to the performance of the standard item, so as to decide whether or not that test item is to be accepted or rejected. To provide a concrete example, consider a large number of crates of fruit being taken from cold storage. Let each crate contain a fruit (the standard) that is inedible owing to a prior treatment that prevents further ripening, and k-1 edible fruits whose measured ripeness, allowing for possible measurement error, must not be more than that of the standard. Let the set of random variables:

$\{X_1, X_2, X_3 ..., X_k\}$, as defined in the previous section,

reasonably describe the real-world array of sets of measurements being made, with X_1 representing that of the standard. Then there has been brought into the real world an extensive array of sets of measurements, say 1 000 or some such large number of sets, of the form:

$\{x_j\text{-}x_1 \text{ for } j = 2, 3, 4 ..., k\}_r$ for r = 1, 2, 3 ..., 1 000,

where these sets can be viewed as 1 000 realisations of the set of random variables:

$\{X_j\text{-}X_1 \text{ for } j = 2, 3, 4 ..., k\}$.

Consider the case:

$X_j\text{-}X_1$ is an $N(\mu_j\text{-}\mu_1, 1)$ random variable for each of $j = 2, 3, 4 ..., k$.

For each fruit to be tested, the decision-maker must decide whether that fruit is:

'not too ripe' (accepting $H_0: \mu_j\text{-}\mu_1 = 0$), or
'too ripe' (accepting $H_1: \mu_j\text{-}\mu_1 > 0$).

Consider a disposable number $d(\gamma)$, and an accept-reject rule of the form:

reject H_0 and accept H_1 if x_j-$x_1 \geq d(\gamma)$
accept H_0 and reject H_1 if x_j-$x_1 < d(\gamma)$
j = 2, 3, 4 ..., k, and r = 1, 2, 3 ..., 1 000

where the value of $d(\gamma)$ arises by specifying the Type I error rate as:

$\gamma = \Pr[X_j\text{-}X_1 \geq d(\gamma) \mid \mu_j\text{-}\mu_1 = 0]$, the same for all of j = 2, 3, 4, ..., k.

Then, for each crate of fruit, the decision-maker performs *k* different hypothesis tests, each of size γ. The expected value of the number of fruits erroneously rejected as too ripe, is thereby, for the entire array of crates, specified to be γ × k × 1 000. If γ = 0.05, the critical value of the X_j-X_1 test statistic, here an N(0,1) random variable, is given by $d(\gamma)$ = +1.645 for each of j = 2, 3, 4, ..., k.

However, as opposed to the foregoing, Dunnett (1955) has the decision-maker specify γ per item, such that α = 0.05 per set equals the population proportion of *sets* where *one or more* items are erroneously rejected. For such a specification, α = 0.05 *per crate* in our case, the value of γ *per fruit* in our case is approximately such that:

$$1\text{-}(1\text{-}\gamma)^k = 0.05. \qquad (2.8.1)$$

This formula arises as follows: let the probability of not erroneously classifying a fruit as too ripe be:

(1-γ) for a 1st fruit, (1-γ) for a 2nd fruit, (1-γ) for a 3rd fruit, ... (2.8.2)

Then, supposing for the moment that these events are statistically independent, the probability of not erroneously classifying any of those fruits as too ripe equals:

(1-γ) × (1-γ) × (1-γ) ...

It follows that the probability of erroneously classifying *one or more* of those fruits as too ripe then equals:

1-[(1-γ) × (1-γ) × (1-γ) ...]

Thus, for example, for k = 9 , α = 0.05 when γ = 0.0057, and the critical value of the N(0, 1) test statistic, i.e. (X_j-X_1) under H_0, then turns out to be approximately $d(\alpha)$ = +2.53. This value is approximate, as events such as those envisaged at (2.8.2) are not statistically independent for fruits from the same crate. The exact critical value is given in Table 1a of Dunnett (1955) as $d(\alpha)$ = +2.42. The derivation of the approximate $d(\alpha)$ value has been given here in order to make the nature of Dunnett's reasoning entirely clear. The derivation of the exact value is explained below.

Dunnett (1955, p. 1096) calls the decision-making procedure we are here discussing, 'a multiple comparison procedure' because he recognises (p. 1097) that it has much in common with the so-called 'multiple comparison procedures' of Scheffé and Tukey (as described by Scheffé, 1959). In Chapter 13 we will come to grips with such procedures in general. Here we merely want to relate Dunnett's procedure to the mathematics of shortfall testing, so as to make it clear that shortfall testing is not a multiple comparison procedure. To that end, we note that Dunnett (1955, using his Equation (4) as simplified on p. 1105)

develops his procedure as directed at finding a (1-α)-confidence region for the values of μ_j-μ_1 for all of j = 2, 3, 4, ..., k, *simultaneously*, as follows. Make the notation:

$Z = X_j$-μ_j, so that $E(Z_j) = 0$, for all of j = 1, 2, 3, ..., k.

Corresponding to our equations at (2.7.2) and (2.7.8) we then have:

$$\Pr[Z_{\max \bullet 1}\text{-}Z_1 \geq d] = 1 - \int_{-\infty}^{+\infty} \prod_{j=2}^{k} F_j[d+y]\,dF_1(y)$$

$$= \alpha \text{ when } d = d(\alpha).$$

This can be expressed as:

$$1\text{-}\alpha = \Pr[Z_{\max \bullet 1}\text{-}Z_1 < d(\alpha)]$$

$$= \Pr[Z_j\text{-}Z_1 < d(\alpha) \text{ for all of } j = 2, 3, 4, ..., k],$$

which corresponds to Dunnett's simplified version of his Equation (4). Hence, for each one of the 1 000 crates, we obtain lower confidence bounds for each one of the k-1 fruits in that crate, by means of the inequalities:

z_j-$z_1 < d(\alpha)$ for j = 2, 3, 4, ..., k, respectively.

Rearranging the terms of these inequalities in more explicit form, the requisite bounds are found to be the left-hand sides of the k-1 inequalities:

$$x_j\text{-}x_1\text{-}d(\alpha) < \mu_j\text{-}\mu_1 \text{ for } j = 2, 3, 4, ..., k, \text{ respectively.} \qquad (2.8.3)$$

We thus obtain, for each crate, a simultaneous confidence region defined by:

$$L_j < \mu_j\text{-}\mu_1 < +\infty, \text{ where } L_j = x_j\text{-}x_1\text{-}d(\alpha), \text{ for } j = 2, 3, 4, ..., k. \qquad (2.8.4)$$

For instance, if k = 9 and we specify α = 0.05, Dunnett's Table 1a gives $d(\alpha)$ = +2.42, as before, and the decision rule developed in the previous paragraph is equivalent to:

> Reject any given crate if and only if $0 \leq L_j$ for *one or more* of the fruits labelled j = 2, 3, 4, ..., k in that crate.

Example 2.8.2

Brownlee (1949) considered the throughput obtained from units of plant before failure through corrosion at each of three foundries (Table 2.8.1). The data prompted him to consider whether there was sufficient justification for 'regarding the Foundry A pots as giving smaller average throughputs than those from the other foundries' (p. 55). It appears at once that a shortfall test of the extent to which the Foundry A throughputs fall short within the group, will provide – in the form of a set of statistical co-ordinates – exactly the kind of measurement required for such consideration. By adapting the PROBMC function of SAS (1992), the mental correlate of the appropriate test datum is found to be situated at (0.916, ε, 0.084), or further to the right, in the many-one-*t* distribution.

Table 2.8.1: Throughput recorded at three different foundries

Foundry	Mean throughput	Number of replicates
C	79.7	3 pots
B	78.3	9 pots
A	53.0	5 pots
Mean square between pots within foundries = 528.5 on 14 df.		

Comparative discussion of Examples 2.8.1 and 2.8.2

(1) The mathematics of a shortfall test of Entry 1 (Section 2.7), and that of Dunnett's procedure with Entry 1 as standard (this section), both involve the parameters:

$$(\sigma^2, \mu_2-\mu_1, \mu_3-\mu_1, \mu_4-\mu_1, ..., \mu_k-\mu_1).$$

In Dunnett's procedure, σ^2 is the only nuisance parameter. But in the shortfall test, the only parameter that is *not* a nuisance parameter is Δ_1, i.e. the substitution value of Entry 1, which value is a special function of:

$$\mu_2-\mu_1, \mu_3-\mu_1, \mu_4-\mu_1, ..., \mu_k-\mu_1,$$

and which value is not explicitly involved by Dunnett's procedure. So, although the hypothesised models in both cases lead to the use of Dunnett's many-one-*t* statistic, the alternatives against which the hypothesised models are being tested are different.

(2) With Entry 1 as the standard, Dunnett's procedure explicitly involves k-1 contrasts amongst the entry means, these being given by:

$$(-1)X_1+\lambda_2X_2+\lambda_3X_3+\lambda_4X_4+...+\lambda_kX_k \text{ when } (\lambda_2, \lambda_3, \lambda_4, ..., \lambda_k) \text{ equals}$$
$$(1, 0, 0, ... 0), (0, 1, 0, ... 0), (0, 0, 1, ... 0), ... (0, 0, 0, ... 1). \quad (2.8.5)$$

A shortfall test of Entry 1 explicitly involves just one contrast amongst the entry means, this being given by:

$$(-1)X_1+\delta_{2j}X_2+\delta_{3j}X_3+\delta_{4j}X_4+...+\delta_{kj}X_k \text{ when j labels } X_{max\cdot 1}, \text{ and } \delta_{ij} \text{ is}$$
$$\text{Kronecker's delta, i.e. } \delta_{ij} = 1 \text{ when i = j, and } \delta_{ij} = 0 \text{ when i} \neq \text{j.} \quad (2.8.6)$$

At (2.8.5) the coefficients of X_2, X_3, X_4, ..., X_k are specified constants not depending on the sample values. In (2.8.6) the coefficients of X_2, X_3, X_4, ... X_k are the values of random variables arising from the sample values.

(3) In Dunnett's procedure, the error rate α, or the corresponding confidence coefficient 1-α, can only be meaningful in repetitive testing, as in say 1 000 repetitions. It would be meaningless to speak of a *rate* with reference to a single item, such as one crate. For just a single crate we could either have made the correct decision ('$\alpha = 0$' so to speak) or we could have made the wrong decision ('$\alpha = 1$' so to speak). But to speak of an error rate of $\alpha = 0.05$ with reference to just one, isolated decision is nonsense. In contrast, shortfall

testing is not repetitive, it does not involve the notion of an error rate, and it does not lead to a decision. In Example 2.8.2 for instance, there is no repetitive testing, there is no error rate involved, and the test does not lead to a decision; it leads to a conclusion in the sense of a fact, which fact is the following:

> The mental correlate of the many-one-t value for foundry A is situated rather far down at (0.916, ε, 0.084) or further down in the right-hand tail of the hypothesised distribution. So, by the test performed, the hypothesised model M_0: 'The foundry A population mean does not fall short within the group', fits the data awkwardly. Whereas, by the same test, instances of the alternative model, M_1: 'The foundry A population mean falls short within the group', would fit the data well. (2.8.7)

This is just a fact and is not in any way prescriptive; the investigator must consider it in conjunction with other facts. Was the layout of foundry A ergonomically inferior to the layouts of B and C? Were the pots at foundry A inferior to those at B and C? If so, the fact cited at (2.8.7) may strengthen the investigator's opinion that foundry A is of inferior type. However, it may be that the investigator finds nothing to explain the awkward fit reported at (2.8.7), in which case the investigator might be of the opinion that the observed co-ordination is extreme by mere coincidence. After all, the hypothesised model *recognises* the possibility of such coincidence.

(4) Dunnett's procedure involves a control and other entries (Example 2.8.1), whereas a shortfall test need not involve a control (Example 2.8.2). Moreover, when shortfall testing *does* involve a control, the natural question to be addressed for the purposes of data analysis is not the one that Dunnett would have us address (Example 2.3.6).

(5) Dunnett (1955) also developed *two-sided* confidence limits for the expected values of the k-1 differences arising at (2.8.5). In contrast, we found in Section 2.7 that the elimination quantities involved by shortfall testing are *strictly upper-bounding* elimination quantities.

Recall a passing remark at the end of Example 2.3.6. Clearly, shortfall testing provides the data analyst with a very useful tool. Yet most statisticians seem to be unaware of such tests. They are not found in Snedecor and Cochran (1989), despite many examples where their use would be appropriate. Other authors, such as Steel and Torrie (1980), present the many–one-*t* statistic as if it were of no more use than being part and parcel of Dunnett's multiple comparison procedure. It is therefore important to grasp that shortfall testing is not a multiple comparison procedure.

2.9 A DESTRUCTIVE PRESCRIPTION REVISITED

In Section 1.21 we saw, in the context of commencement testing, that inappropriate importation of certain ideas, from hypothesis testing into data analysis, results in a normative prescription with destructive consequences. In this section we display, in the context of elimination testing, further examples of such destruction. Recall that the essence of the normative prescription is embodied by the following rules:

> (1) Let SL denote the significance level attached to given data by a significance test of a hypothesised model M_0 against an alternative M_1.

(2) Specify a test size α, which is a small fraction selected *without reference to the data*. A much favoured value is $\alpha = 0.05$.

(3) If SL $\leq \alpha$, reject M_0 and accept M_1. If SL $> \alpha$, accept M_0 and reject M_1.

We begin with an example that displays destruction of statistical evidence, followed by two examples that display destruction of substantive evidence.

Example 2.9.1

A randomised block design was used to compare the tolerance of six different kinds of rootstock to waterlogged conditions. Each plot comprised a waterlogged plant and an own control. Response, per plot, was measured as total shoot length of the own control minus that of the waterlogged plant, relative to that from the own control. Three plots were lost owing to rootstocks that failed to take root. Table 2.9.1 displays a suite of shortfall tests on the responses.

Table 2.9.1: Shortfall tests for elimination of root-stocks that are 'lower than best' in respect of toleration of waterlogged conditions

Root-stock	No. of replicates	Mean	Many-one-t	Co-ordinates
A: Red leaf prunus 1	5	-46.49		
B: Peach-almond	5	-58.36	1.0929	(0.606; 0.394)
C: Red leaf prunus 2	5	-60.82	1.3194	(0.700; 0.300)
D: Red leaf prunus 3	4	-68.14	1.8794	(0.874; 0.126)
E: Red leaf prunus 4	5	-72.34	2.3801	(0.945; 0.055)
F: Peach	3	-85.20	3.0866	(0.988; 0.012)
Error mean square = 294.9 on 17 degrees of freedom				

As 'lowest is best' in this example, the sign of each treatment mean is reversed and then tested as if 'highest is best'. The evidence points strongly at F as lower than best, and points less strongly at E as also lower than best. In order to see that it is impossible to convey this by way of just one single cut-off, note that for each member of the suite of tests, the right statistical co-ordinate is interpretable as a significance level subject to negligible rounding. The cut-off for a corresponding hypothesis test of size α is then given by:

Accept as lower than best if and only if SL $\leq \alpha$, and not so if SL $> \alpha$.

Two widely favoured test sizes are $\alpha = 0.01$ and $\alpha = 0.05$. But let us for good measure also consider $\alpha = 0.075$.

If $\alpha = 0.01$ is specified, the result is:
None of the six entries are lower than best. (2.9.1)

If $\alpha = 0.05$ is specified, the result is:
F is lower than best. The other five entries are not lower than best. (2.9.2)

If $\alpha = 0.075$ is specified, the result is:
F and E are lower than best. The other four entries are not lower than best. (2.9.3)

At (2.9.1) the poor performances of both F an E have been concealed. At (2.9.2) the poor performance of E has been concealed. At (2.9.3) the difference in performance of F an E has been concealed. In all three cases, it was also concealed whether or not the significance level for any entry classified as not lower than best, is *much* larger than α rather than *just* larger than α. This is bad. Yet there is worse to come. In order to come to grips with that, we must first note that on the evidence conveyed by Table 2.9.1, horticultural opinion could hardly be divided on the following facts:

> As tested, all the models with F as a best entry fit the data poorly. The same is true of E, though the fit is not as poor as in the case of F. With D as a best entry, the fit is very slightly awkward, but hardly poor. As tested, many models with A as a best entry, and many models with B as a best entry, fit the data well. (2.9.4)

Bear in mind that the value of α is a *single* specification *to be made without reference to the data*. So some advocates of the accept–reject rules recommend that, rather than employ the rules directly, the 'realised' significance levels should be reported 'so that,' they explain, 'a recipient of the report can specify his/her own value of α' (Wackerly, Mendenhall and Scheaffer 1996, p. 432). If three different members of a horticultural society should then specify:

$\alpha = 0.01$, $\alpha = 0.05$, $\alpha = 0.075$, respectively, (2.9.5)

we would, instead of promoting informed opinion *by way of facts arising from given data* as at (2.9.4), be instigating a 'scientific' controversy *by way of values specified without reference to those data*.

We note in passing that owing to the missing plots, the sample means for this example are correlated in ways that do not conform to standard facilities for shortfall testing. The results given in Table 2.9.1 were obtained by simulation (Sadie 1996).

Example 2.9.2

Consider an experiment concerning the use of sulphur for controlling downy mildew in vineyards. Suppose two methods of application and an untreated control lead to the results in Table 2.9.2, all applications having involved the same amounts of sulphur per plot.

Table 2.9.2: Analysis of yields from a hypothetical experiment with grape vines to compare different methods of applying a fixed amount of sulphur per plot

Method	Mean yield	Estimated effect	Student's t	Co-ordinates of t in test distribution
Dusting once a week	7.96	+3.04	+2.64	(0.993, ε, 0.007)
Drifting once a week	6.91	+1.99	+1.73	(0.905, ε, 0.095)
Untreated control	4.92	–	–	–
Standard error of a treatment difference = 1.15 on 28 df.				

'Drifting' means that the dust is allowed to settle down over the plants from above. 'Dusting' (the usual procedure) means that the sulphur is forced down among the vines by a blast of air, where the purpose of the experiment is to establish whether or not that *does* make the application more effective. Sulphur application is known to control downy mildew, and we must bear in mind that a positive but weak response to sulphur might to some extent be masked by experimental error. So, based on the evidence presented in Table 2.9.2, it is reasonable to conclude that both the sulphur treatments increased yield as expected and that, perhaps also as expected, dusting *does* seem to be more effective that drifting. However, as it is known that sulphur applications in limited amounts do not harm the plants, a widely used normative prescription would have us consider for each method:

M_0: E(effect) = 0 as hypothesised model, versus M_1: E(effect) > 0 as alternative,

and would then have us apply the following accept-reject rule:

If $\varepsilon+V \leq 0.05$, reject M_0 and accept M_1. If $\varepsilon+V > 0.05$, accept M_0 and reject M_1.

This rule would have us conclude that the sulphur has a beneficial effect if applied by dusting, but no effect if applied by drifting, where that asks us on statistical grounds to draw a conclusion that on substantive grounds we believe to be fallacious. (For a real-world example, refer to Goulden 1939, Example 34, p. 149.)

Example 2.9.3

In the arid region of South Africa known as the Great Karoo, dust devils arise in the summer heat. Let us observe just eight of them, with just one whirling anti-clockwise, and let us model these data as a binomial sample from a population of clockwise and anti-clockwise winds. Then a co-ordination test with

M_0: P(clockwise) = 0.5 as hypothesised model
versus
M_1: P(clockwise) ≠ 0.5 as alternative

places the mental correlate of the observed number of clockwise winds at

(0.965, 0.031, 0.004) in the test distribution.

This would point to the right in favour of the popular belief that, owing to the rotation of the earth round its own axis, such winds more often than not whirl anti-clockwise in the northern hemisphere, and clockwise in the southern hemisphere. However, we may well be sceptical of such mechanics, and consider the following model instead. As the sun beats down, the temperature over certain terrain, say a patch of bare earth, rises to a higher level than that over surrounding parts shaded by plant cover. Hot air thus rises from the bare patch and cooler air is drawn in from surrounding parts, resulting in an unstable equilibrium. Subsequently, the equilibrium is destabilised by some or other co-incidence – by the flight of a bird, or by the shape of a bush, or by the disturbance of some dry leaves – and so a dust devil whirls into being. Thus, owing to the coincidental nature of the destabilising cause, there is then no reason to expect that the two different directions of whirl occur in unequal proportions. So a normative prescription such as

> 'If and only if $\varepsilon+V \leq 0.05$, reject M_0 and accept the rotational mechanics'

would have us reject a model we believe to be correct, and have us accept a model we suspect of being ill-conceived. So again, as in the previous example, insistence on the normative prescription would have us on statistical ground draw a conclusion that on substantive ground we believe to be fallacious.

Discussion of Examples 2.9.1, 2.9.2 and 2.9.3

These examples show that the use of accept–reject rules based on fixed cut-offs selected without reference to the data, leads to two difficulties:

> (1) It might happen that the data strongly point at a possibility to which, for good reasons based on other considerations, we are unwilling to grant credence.
>
> (2) It might happen that the data weakly point at a possibility to which, for good reasons based on other considerations, we are willing to grant credence.

Difficulty (1) is uncommon, as it arises from coincidence. Difficulty (2) is common, as it arises from the limitations of finite data. In neither case, however, can there be any merit in arguments that would refuse us the right to try to recognise such cases by way of scientific reasoning that takes into account facts not arising from the numerical data under analysis. In other words, statistical epistemology based on the notion that, in respect of the subject matter under investigation, a substantive investigator *is entirely ignorant beyond the numerical data being subjected to scrutiny*, is arrant nonsense. It is a sad state of affairs that this has not been understood in much, even most, of the statistical literature.

2.10 On a notion of objectivity

Sometimes prescriptions that are formally similar to those discussed in the previous section are advanced on the ground of promoting 'objectivity'. And perhaps that is what motivated Fisher (1926, p. 504) to recommend as follows:

> 'Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level.' (2.10.1)

This is puzzling, as Fisher liked to calculate the value of a significance level, and it is difficult to imagine him drawing such a vast distinction between a level of 0.049 and one of 0.051. The recommendation was, however, not intended to provide a forecasted Type I error rate of 0.05, as Fisher (1973, p. 47) makes entirely clear when he explains that significance tests:

> '... do not generally lead to any probability statements about the real world'.

He therefore wants to point at solitary data sets in the real world, and to employ significance tests to test, for their tenability, explanatory populations brought into the human mind. So it is difficult to interpret the rule introduced at (2.10.1) as directed at anything else than promoting objectivity. Be that as it may, we consider, by way of the following example, whether or not such a rule *would* in fact promote objectivity.

Example 2.10.1

In 1764 Arthur Young drew seven paired comparisons between the profit per acre (in pounds sterling) from wheat when broadcasting the seed (old husbandry) and drilling the seed in rows (new husbandry). His data are reproduced in Table 2.10.1. In modern terms, the

Table 2.10.1: Difference (drilled vs. broadcasted), in profit per acre (Young 1771)

Field No.	1	2	3	4	5	6	7
Difference	-0.3	-0.3	-2.3	-0.7	+1.7	-2.2	-3.0
Total: -7.1; Standard deviation: 1.607							

treatments were replicated in seven blocks, each block comprising a pair of plots with 'the soil exactly the same in both'. Let us model the treatment differences as a sample from an $N(\mu, \sigma^2)$ population ($-\infty < \mu < +\infty$, $0 < \sigma^2 < +\infty$). The datum mean difference (drilling minus broadcasting) is -1.01, whose estimated standard error is 0.6074 on 6 df. Using Student's *t* to test:

M_0: $\mu = 0$ versus M_1: $\mu \neq 0$

the test datum is $t = -1.663$ on 6 df., whose mental correlate is situated at:

(0.072, ε, 0.928) in Student's test distribution.

Consider the normative prescriptions:

If $U+\varepsilon \leq 0.075$, reject M_0 and accept the old way as better.

If $\varepsilon+V \leq 0.075$, reject M_0 and accept the new way as better.

These prescriptions would have us conclude (decide) that the old way is better, but do we believe that? Are we being 'subjective' when we suspect that the given data set is misleading? Drilling clearly intends to provide for an even dispersion of seed and for planting at a given depth. So it is difficult to believe that drilling, correctly applied, can lead to a decrease in yield. And, inasmuch as the normative prescriptions would have us ignore that, the prescriptions would be better described as promoting obstinacy rather than promoting objectivity. It would surely be better for agronomic opinion to reserve judgement till better informed by further experimentation.

2.11 Normative prescriptions as comfortable recipes

It is worth noting that the problem of normative prescriptions is aggravated in part by the attraction of cut-and-dry decision rules, i.e. it being easier to decide than to think.

2.12 Scientific data analysis in case of elimination testing

We have challenged the reader to note that in the development of elimination tests our reasoning is a seamless continuation of the reasoning that was previously employed in the development of commencement tests. So, inasmuch as Section 1.31 underlined the analytic nature of that reasoning in the context of commencement testing, the reader is challenged to note how the following example underlines the same analytic nature of the reasoning in wider context. The example also serves as a platform for a number of developments in subsequent sections.

Example 2.12.1

Snedecor (1956, p. 332) gives the gains in weight (grams) of 60 male rats in a feeding experiment for comparison of a 3×2 factorial array of rations. Each ration was fed to ten replicate rats in a completely randomised design. The rations comprised:

3 sources of protein (beef, cereal, pork)×2 levels of protein (high, low).

Tables 2.12.1 and 2.12.2 give the gains in weight and certain 'summary statistics'.

We present an appropriate analysis of these data using analytic tests of co-ordination.

To begin with we analyse the six-fold compound class characteristic by performing three different commencement tests, one for skewness, one for non-normal kurtosis and one for heterogeneity of variance, as follows.

Using a transformation described by D'Agostino (1970), the six Pearson measures of component skewness given in Table 2.12.2 are transformed to the six $N(0, 1^2)$ values denoted by Z_1 in that table. In order to test the six-fold compound class characteristic for shared skewness, the latter values are gathered into a single degree of freedom in the form of the further $N(0, 1^2)$ value given by:

$$(-1.322-0.792+0.000-1.400-0.288+0.933)\div\sqrt{6} = -1.172, \qquad (2.12.1)$$

whose mental correlate in the human mind is found to be situated at:

$$(0.12, \varepsilon, 0.88) \text{ in the test distribution.} \qquad (2.12.2)$$

The hypothesised symmetry of the compound class characteristic, thus tested, fits the given data well.

Using a table obtained via statistical simulation by D'Agostino and Tietjen (1971) the six Pearson measures of component kurtosis given in Table 2.12.2 are transformed to statistical co-ordinates and via those co-ordinates to the six $N(0, 1^2)$ values denoted by Z_2 in Table 2.12.2. In order to test the six-fold compound class characteristic for shared non-normal kurtosis, the latter values are gathered into a single degree of freedom in the form of the further $N(0, 1^2)$ value given by:

$$(-0.432+0.860+0.860+0.269-0.813+0.169)\div\sqrt{6} = +0.373,$$

whose mental correlate in the human mind is found to be situated at:

$$(0.65, \varepsilon, 0.35) \text{ in the test distribution.} \qquad (2.12.3)$$

Table 2.12.1: Weight gains of 60 rats on different rations comprising 3 × 2 factorial combinations, each ration was fed to ten replicate rats in a completely randomised design. The treatment combinations comprised three protein sources at two different levels.

Source	Level	Gains in weight (grams)										Mean
Beef	High	73	102	118	104	81	107	100	87	117	111	100.0
Cereal	High	98	74	56	111	95	88	82	77	86	92	85.9
Pork	High	94	79	96	98	102	102	108	91	120	105	99.5
Beef	Low	90	76	90	64	86	51	72	90	95	78	79.2
Cereal	Low	107	95	97	80	98	74	74	67	89	58	83.9
Pork	Low	49	82	73	86	81	97	106	70	61	82	78.7

Table 2.12.2: Moments round the mean and derived quantities for weight gains of replicate rats for 3 × 2 treatment combinations in a completely randomised design

Protein source	Beef	Cereal	Pork	Beef	Cereal	Pork
Protein level	High	High	High	Low	Low	Low
$m_2 = \frac{1}{n}\sum (x\text{-}x)^2$	206.20	203.09	107.25	173.56	222.09	246.41
$m_3 = \frac{1}{n}\sum (x\text{-}x)^3$	-1 624.80	-1 060.57	-1.50	-1 830.74	-532.27	-563.90
$m_4 = \frac{1}{n}\sum (x\text{-}x)^4$	8 9613.4	125 242.3	36 572.4	79 385.3	92 057.5	155 373.2
$\sqrt{b_1} = \frac{m_3}{m_2\sqrt{m_2}}$	-0.54874	-0.36644	-0.00135	-0.8006	-0.16082	-0.14579
Z_1	-1.3225	-0.7923	0.0000	-1.4000	-0.2876	+0.9326
$b_2 = \frac{m_4}{(m_2)_2}$	2.108	3.037	3.179	2.635	1.866	2.559
U for b_2	0.333	0.805	0.805	0.606	0.208	0.567
Z_2 for U	-0.432	+0.860	+0.860	+0.269	-0.813	+0.169

The hypothesised normal kurtosis of the compound class characteristic, thus tested, fits the given data well.

The standard test for heterogeneity of variance, introduced by Bartlett (1937), is very sensitive to non-normality (Box 1953). An alternative test, by Levene, is insensitive to non-normality (Snedecor and Cochran 1989, p. 252). Levene replaces the given data with the absolute values of the corresponding residuals and performs the analysis of variance and omnibus *F* test for between-group differences. The *F* value obtained is the test datum. Its mental counterpart, under the hypothesised model of homogeneity of variance, is approximately distributed as Snedecor's central F. In the present case the analysis in Table 2.12.3 is obtained, and the

Table 2.12.3: Analysis of variance of absolute residuals (Levene's test)

Source of variation	df.	Sum of squares	Mean square	*F* value
Between groups	6-1	171.229	34.246	0.4778
Within groups	6(10-1)	3870.440	71.675	
Total	60-1	4041.669		

mental correlate of the datum *F* in the human mind is found to be situated at:

$$(0.21, \varepsilon, 0.79) \text{ in Snedecor's central } F \text{ distribution.} \qquad (2.12.4)$$

The hypothesised homogeneity of variance of the compound class characteristic, thus tested, fits the given data well.

Note that if x_{ij} denotes the j^{th} gain from the i^{th} treatment, $\bar{x}_{i\cdot}$ denotes the mean gain from the i^{th} treatment, and s the estimated error standard deviation, then a realisation of the compound class characteristic, untrammelled by the members of the class, is given by the standardised residuals:

$$(x_{ij}-\bar{x}_{i\cdot})\div s,\ i = 1, 2, 3, ..., 6,\ j = 1, 2, 3, ..., 10. \qquad (2.12.5)$$

Inasmuch as the realised value of each one of the three test statistics used at (2.12.2), (2.12.3) and (2.12.4) can be calculated from these residuals, each of the three tests thus performed is by Definition1.15.1 a commencement test.

We now wish to consider certain elimination tests. For that purpose, beef, cereal and pork are denoted by B, C and P, respectively, high and low by H and L, respectively, and the population means and error variance as:

$$\mu_{BH}, \mu_{CH}, \mu_{PH}, \mu_{BL}, \mu_{CL}, \mu_{PL} \text{ and } \sigma^2, \text{ respectively.} \qquad (2.12.6)$$

Interest then concerns the tenable values of seven different contrasts of the form:

$$\lambda_{BH}\mu_{BH}+\lambda_{CH}\mu_{CH}+\lambda_{PH}\mu_{PH}+\lambda_{BL}\mu_{BL}+\lambda_{CL}\mu_{CL}+\lambda_{PL}\mu_{PL} \qquad (2.12.7)$$

arising when the λ's are specified as in Table 2.12.4, where for instance the contrast numbered '1' in that table arises as:

$[(+½)\ \mu_{BH}+(0)\mu_{CH}+(-½)\ \mu_{PH}]-[(+½)\ \mu_{BL}+(0)\mu_{CL}+(-½)\ \mu_{PL}]$.

Table 2.12.4: Contrasts of interest in comparison of mean weight gains of rats fed 6 different rations comprising a 3×2 factorial array of treatment combinations

Protein source	Beef	Cereal	Pork	Beef	Cereal	Pork
Protein level	High	High	High	Low	Low	Low
1 Beef vs pork × high vs low	+½	0	-½	-½	0	+½
2 Beef vs pork	+½	0	-½	+½	0	-½
3 Beef vs pork (both high)	+1	0	-1	0	0	0
4 Beef vs pork (both low)	0	0	0	+1	0	-1
5 Meat vs cereal × high vs low	+¼	-½	+¼	-¼	+½	-¼
6 Meat vs cereal (both high)	+½	-1	+½	0	0	0
7 Meat vs cereal (both low)	0	0	0	+½	-1	+½
Error variance estimate: 10.697 on 66 df.						

Each of the seven quantities thus defined is a contrast amongst the treatment effects, as the sum of the values = 0 in each case. Denote the sample means and error variance as:

$\overline{X}_{BH}, \overline{X}_{CH}, \overline{X}_{PH}, \overline{X}_{BL}, \overline{X}_{CL}, \overline{X}_{PL}$, and S^2, respectively,

where any population contrast defined at (2.12.7) is estimated by the corresponding sample contrast, i.e. by:

$$\lambda_{BH}\overline{X}_{BH}+\lambda_{CH}\overline{X}_{CH}+\lambda_{PH}\overline{X}_{PH}+\lambda_{BL}\overline{X}_{BL}+\lambda_{CL}\overline{X}_{CL}+\lambda_{PL}\overline{X}_{PL}. \qquad (2.12.8)$$

In Table 2.12.4, each of seven contrasts is expressed on single-plot basis, that is to say, as:

a mean per plot – another mean per plot,

the sum of the positive λ values and the sum of the negative λ values being equal to +1 and -1, respectively. This enables comparison of the precision of different sample contrasts considered as estimates of the corresponding population contrasts because the estimated standard error of the expression at (2.12.8), considered as an estimate of the corresponding expression at (2.12.7), is given by:

$S \times \sqrt{\frac{D}{n}}$, where

S denotes the error standard deviation,

D denotes the sum of squares of the λ's,

n denotes the number of replications (n = 10 replications in our case), and

for the seven contrasts numbered 1, 2, 3, ..., 7 in Table 2.12.4,

$$\sqrt{\frac{D}{n}} = \sqrt{\frac{1}{n}}, \sqrt{\frac{1}{n}}, \sqrt{\frac{2}{n}}, \sqrt{\frac{2}{n}}, \sqrt{\frac{3}{4n}}, \sqrt{\frac{3}{2n}}, \sqrt{\frac{3}{2n}}, \text{ respectively.} \quad (2.12.9)$$

The minimal sufficient statistic for:

$$\Gamma = \lambda_{BH}\mu_{BH}+\lambda_{CH}\mu_{CH}+\lambda_{PH}\mu_{PH}+\lambda_{BL}\mu_{BL}+\lambda_{CL}\mu_{CL}+\lambda_{PL}\mu_{PL}$$

is given by $\{C, S^2\}$, where:

$$C = \lambda_{BH}\overline{X}_{BH}+\lambda_{CH}\overline{X}_{CH}+\lambda_{PH}\overline{X}_{PH}+\lambda_{BL}\overline{X}_{BL}+\lambda_{CL}\overline{X}_{CL}+\lambda_{PL}\overline{X}_{PL}.$$

So, an appropriate elimination quantity for Γ is given by:

$$\text{Student's } t = (C\text{-}\Gamma)\div(S\times\sqrt{\frac{D}{n}}). \quad (2.12.10)$$

The resulting elimination tests can in fact not reasonably be improved upon, as they provide what statistical jargon describes as:

> 'most separating similar regions tests of co-ordination, uniformly so over all pairs of Γ values, and invariably so over all levels of elimination', where that holds for any solitary real-world data set of the form $(C, S) = (c, s)$, given the class of models.

Substantive considerations together with inspection of the means given in Table 2.12.1, then suggest that gains from beef vs. pork might be the same for high and low, and (also) might sum to zero. That is to say, it is suggested that a tenable class of models might be obtained when:

$$[+½]\ \mu_{BH}+(0)\mu_{CH}+[-½]\mu_{PH}+[-½]\mu_{BL}+(0)\mu_{CL}+[+½]\mu_{PL}$$
and also
$$[+½]\mu_{BH}+(0)\mu_{CH}+[-½]\mu_{PH}+[+½]\mu_{BL}+(0)\mu_{CL}+[-½]\mu_{PL}$$
are both hypothesised as equal to zero. (2.12.11)

Here the first hypothesised class comprises all the initial models with zero interaction of beef vs. pork × high vs. low, and the second hypothesised class comprises all the initial models with zero main effect of beef vs. pork. The communality of these two classes is the hypothesised class comprising all initial models with expected gains the same from beef as from pork. By testing the pair of values hypothesised at in (2.12.11), the corresponding pair of datum values of Student's *t*, as defined at (2.12.10), are found to be 0.000 and 0.108 both on 54 df., whose mental correlates are situated in Student's test distribution at:

$$(0.500, \varepsilon, 0.500) \text{ and } (0.543, \varepsilon, 0.457), \text{ respectively.} \quad (2.12.12)$$

Thus tested, the values hypothesised at (2.12.11) fit the given data well, where that implies that to distinguish between beef and pork is superfluous here, that is to say, *implies* that the simple effects numbered 3 and 4 in Table 2.12.4 are satisfactorily modelled as zero. This raises a question:

> Why reason in terms of the *implication?* Why not test directly whether the two simple effects are zero? (2.12.13)

We reply:

> Two such tests would be less separating than those reported at (2.12.12); this appears at once by inspection of the values listed at (2.12.9). (2.12.14)

In Section 2.14 we return to this point.

Further analysis is straightforward. Using Student's *t*, as defined at (2.12.10), to test whether or not the contrast numbered 5 in Table 2.12.4 may be modelled as equal to zero, the datum value of *t* is found to be 2.343 on 54 df., whose mental correlate is situated at:

> (0.989, ε, 0.011) in Student's test distribution. (2.12.15)

So, a tenable class of models would require non-zero interaction of meat vs. cereal × level of protein. More specifically, it would require the simple effect of meat vs. cereal at the high level of protein to be substantially larger than the corresponding simple effect at the low level of protein. In fact, using Student's *t* as defined at (2.12.10), to test whether these two simple effects (numbered 6 and 7 in Table 2.12.4) might be modelled as each being equal to zero, the datum values of *t* are found to be 2.441 and -0.872 both on 54 df., whose mental correlates are situated in Student's distribution at:

> (0.991, ε, 0.009) and (0.194, ε, 0.806), respectively. (2.12.16)

So, at first it might well seem that any tenable class of models would have a positive simple effect at the high level, and a zero simple effect at the low level. However, on second thoughts a more tenable class of models would have positive simple effects at both levels, but would have the effect at the low level of being too small to have been detected (Cf. Examples 2.9.2 and 2.9.3).

The reader will note that in the case of elimination testing the data-analytic approach is facilitated by the possibility of expressing different questions in parametric terms. In the present example for instance the elimination tests deal with questions such as: 'Is there non-zero interaction of such-and such kind?' 'If so, is there such-and-such a non-zero simple effect?' And so on. For each question in turn, the minimal sufficient statistic for the parameter involved is made to address the corresponding data. Thus, for instance, the minimal sufficient statistic for meat vs. cereal at the high protein level is given by the error variance estimate together with the contrast numbered 6 in Table 2.12.4, and is used at (2.12.16) to address the corresponding datum, *without any reference whatsoever* to the other six contrasts listed in Table 2.12.4, *and without any reference whatsoever* to the three class characteristics tested at (2.12.2), (2.12.3) and (2.12.4), respectively. In the case of a commencement test the selection of a minimally sufficient statistic for a question of interest is more subtle. Nevertheless, the principle of trying to achieve an analysis, here exemplified by the three commencement tests, is not to be abandoned. There can of course be no objection to consideration of various parametric alternatives to *assist* in the selection of commencement tests with desirable separating characteristics, and to *assist* toward analysing the class characteristics into different subsidiary characteristics – as long as we clearly understand that in actual practice we can never have a system of parameters that exhausts all possible variety of how the subject matter under investigation might have come about, i.e. as long as we clearly understand that a realistic approach will in practice necessarily involve commencement tests for which the alternatives cannot be listed exhaustively.

2.13 In the statistical laboratory

In the previous section we analysed a class of explanatory models for Snedecor's beef-cereal-pork data, on the one hand into the class characteristic, and on the other hand into the different members of the class, the latter being indexed by a vector comprising the seven parameters listed at (2.12.6). With regard to the class characteristic, a further analysis developed three subsidiary class characteristics as tested at (2.12.2), (2.12.3) and (2.12.4), respectively. With regard to the members of the class, factorial analysis then developed five subsidiary classes, each of whose members are tested either at (2.12.12), or at (2.12.15), or at (2.12.16), respectively. Each of the tests results in a *factual* statement, not a *probability* statement. In the statistical laboratory we can demonstrate those facts and other related facts as follows.

Consider the test at (2.12.4): let us calculate the datum F from the residuals indicated at (2.12.5). Then, using the hypothesised model, let us simulate 25 000 independent sets of corresponding residuals to obtain 25 000 corresponding F_0 values, whence the co-ordinates at (2.12.4) are for all practical purposes obtainable as:

U = the proportion of F_0 values < the datum F
ε = the proportion of F_0 values = the datum F
V = the proportion of F_0 values > the datum F

This conveys a fact that can be forced upon the reader's body by a graphic display of the datum F in relation to the 25 000 simulated F_0 values, at which we can then point and say: 'See for yourself how snugly the datum F is situated within the F_0 crowd.'

Again, consider the simple effect of meat vs. cereal at the high level of protein, as tested on the left at (2.12.16). The co-ordinates given there (0.991, ε, 0.009), convey a fact – a fact that can be forced upon the human body by laboratory demonstration. We can, by means of simulation, generate a population of the hypothesised kind, and by generating a sufficiently large population (numbering 25 000? 50 000? 75 000?), we can calculate, to any specified degree of precision, the co-ordinates of the mental correlate of the given datum. Thus, by pointing at a suitable display of results, we could force the human body to recognise a fact of poor fit. 'See for yourself' we would say 'that more than 99 % of the population values of Student's t are situated left of the mental correlate of the datum value of Student's t'. Moreover, by similar means we can force the human body to recognise that when hypothesising certain non-zero values for the parameter of interest, facts of good fit are obtained, thereby exhibiting, as observed physical facts, the separating characteristics of our test in the particular case, i.e. without reference to any host of other cases.

Three important points are made here.

Firstly, any facts developed by statistical data analysis are demonstrable *by laboratory procedure*. They are of precisely the same standing as any facts (concerning particular cases) as might be developed in a chemical laboratory, a microbiological laboratory, or a metallurgical laboratory. Contrary to various deeply entrenched ideas on statistical inference, investigative statistics appropriately employed is tantamount to laboratory procedure. That is not to say that its findings do not apply outside the laboratory, as in the case of chemistry, or of metallurgy, or of microbiology, but we must deny emphatically that

results of sound data analysis can at all be subject to any error. In a practical investigation it will be found that, for instance, an appropriate co-ordination test will often (even usually) not involve laboratory procedure *directly*; *indirectly,* however, it *always* relies on such procedure, because simulation always provides its ultimate line of defence.

Secondly, the facts in question are facts concerning *a single solitary datum in the real world*. True, simulated outcomes are also particular cases in the real world, but serve only to *represent* explanatory populations that are being brought into the human mind.

Thirdly, we make no attempt whatsoever to express the 'probability' of our findings as being erroneous or not. The notion of such a 'probability' is ill conceived. How does one attach a probability of error to *a physically perceived fact* – a fact that we can point out and of which we can say: 'See for yourself ...'? If we point and say: 'See for yourself that the sunset is red', does that involve the probability that the sunset is red? If we 'point' by saying: 'Feel for yourself that this water is cold', does that mean that this water is probably cold? If we 'point' by saying: 'Taste for yourself; this soy sauce is salty', what is the probability that the soy sauce might not be salty? If we point and say: 'See that horse', we *mean* that that is a horse, because if we want to say that that is *possibly* a horse, we would *say:* 'That is possibly a horse.' This point is exceedingly important, as the literature on statistical inference has largely failed to understand that statistical evidence is factual and so does not require the probabilistic notions that pervade that literature. Statistical evidence does not involve anything in the nature of taking a chance, hazarding a guess, or involving a risk. A statistical laboratory is not some sort of casino of science. Statistics is not, as a certain silly slogan would have it, 'the science of uncertainty'; the concept uncertainty belongs to psychology, not to statistics.

2.14 The principle of scientific implication

If meteorology has established that certain meteorological conditions are followed by snow and bitterly cold winds on the high grounds of the Eastern Cape, then animal husbandry must conclude that following such meteorological conditions any newly shorn sheep might perish on those high grounds if not brought into shelter. Again, if physics establishes that certain sedimentary rocks are at least six million years old, then palaeontology must conclude that a fossil imbedded in those rocks is at least six million years old. The point here is simply that scientific knowledge cannot be divorced from its implications, including those in other branches of science. This must also hold for statistical science and, as evidenced by current statistical practice, the following two examples represent cases in point in that, by tacit consent, there is universal agreement on the part of the statistical profession.

Example 2.14.1

The results of the elimination tests we performed in Section 2.13 for Snedecor's beef-cereal-pork data are summarised as follows, using the '±'-convention for initialling standard errors:

$$\begin{aligned} &\text{Meat vs. cereal (both high protein): } 99.8\text{-}85.9 = +13.9\pm5.7 \\ &\text{Meat vs. cereal (both low protein): } 79.0\text{-}83.9 = -4.9\pm5.7 \end{aligned} \qquad (2.14.1)$$

Consider also a formally similar but substantively very different data set adapted from Steel and Torrie (1980), as in Table 2.14.1. In this case there are no indications of interaction, and the elimination data are effectively summarised by the contrasts 'larboard vs. starboard' and 'decks vs. sides', as follows:

Larb. vs. Starb.: [(10.45+10.48)÷2]-[(7.76+7.85)÷2] = +2.7±1.3
Deck vs. Side: [(10.45+7.76)÷2]-[(10.48+7.85)÷2] = -0.1±1.3 (2.14.2)

Table 2.14.1: Density data measured as observed numbers of oysters per 400 cm², located on four different parts of a liberty ship artificial reef off the coast of North Carolina. Here given as mean numbers of 12 observations per location, expressed on a scale suitable for an analysis of variance (Steel and Torrie 1980)

Location	Larboard	Larboard	Starboard	Starboard
	Deck	Side	Deck	Side
Mean number	10.45	10.48	7.76	7.85

We now ask a pair of rhetorical questions.

Question 1: Consider the facts at (2.14.1) and suppose that the 30 rats assigned to high protein and the 30 rats assigned to low protein had been assigned the other way round. Would it be at all credible that instead of the facts at (2.14.1), our findings might then also have been the other way round, as below?

Meat vs. cereal (both high protein): 79.0-83.9 = -4.9±5.7.
Meat vs. cereal (both low protein): 99.8-85.9 = +13.9±5.7.

Question 2: Consider the facts at (2.14.2) and suppose that the ship had been situated the other way around. Would it be at all credible that instead of the results at (2.14.2), our findings might then also have been the other way round, as below?

Larb. vs. Starb.: [(10.45+7.76)÷2]-[(10.48+7.85)÷2] = -0.1±1.3.
Deck vs. Side: [(10.45+10.48)÷2]-[(7.76+7.85)÷2] = +2.7±1.3.

We are compelled to answer Question 1 in the negative, and Question 2 in the affirmative.

In order to answer Question 1 in the affirmative we would have to give credence to an incredible possibility: according to our own statistical tests it is simply incredible that the random assignment of 60 rats to two groups numbering 30 rats each, could so utterly fail to separate the causative agents of systematic variation (the effects of treatment) from the causative agents of statistical 'error' (the variance of experimental material). Opposed to that, we are compelled to concede that an affirmative answer to Question 2 cannot be ruled out, as we have no shred of evidence to indicate that the classification:

(larboard, starboard) × (deck, side)

identifies the causes of the observed systematic variation. What about tidal currents? What about prevailing winds? What about wave action? How was the ship situated in

relation to the passage of the sun? How was the ship situated in relation to the coast and the open sea? Did starboard, larboard, prow or stern face the open sea?

The foregoing concerns a well-known distinction between the possible interpretations of experimental data as opposed to those of survey data. We should, however, note carefully that this distinction exemplifies a broader consideration that was touched upon at (2.12.14). Consider the following question:

> Do the gains observed in Snedecor's beef-cereal-pork data point at any of the expected gains from the six treatment combinations being less than the highest?

Poor understanding of scientific reasoning might have us address this question ineptly as in Table 2.14.2 because any reasoning that appeals to Table 2.14.2 in isolation of the implications of the experimental procedures that lead to the numerical data involved, amounts to turning a blind eye to the implications of that procedure.

Table 2.14.2: Example displaying inept usage of a suite of shortfall tests

Protein source	Beef	Cereal	Pork	Beef	Cereal	Pork
Protein level	High	High	High	Low	Low	Low
Mean response	100.0	85.9	99.5	79.2	83.9	78.7
Apparent shortfall	-0.5	+14.1	+0.5	+20.8	+16.1	+21.3
Left-most co-ordinate	(•, 0.86)	(•, 0.07)	(•, 0.81)	(•, 0.01)	(•, 0.03)	(•, 0.00)

It must be firmly grasped that we statisticians have ourselves overseen the application of the treatment combinations so as to ensure against systematic confounding of those treatments with any extraneous causative agents. So, it would be self-contradictory for us to turn a blind eye to our understanding of the causative agents involved. We, by appropriate analysis, have obtained (to quote our own findings from Section 2.12) overwhelming evidence that:

> ' … a tenable class of models would require non-zero interaction of meat vs. cereal × level of protein and, more specifically, would require the simple effect of meat vs. cereal at the high level of protein to be substantially larger than the corresponding simple effect at the low level of protein'.

This finding *implies* that the highest gains (the 'best' gains) arise from the high levels of animal protein, and also *implies* that at the high level of vegetable protein the gains are lower than 'best'. So, inasmuch as Table 2.14.2 might be interpreted as casting doubt on these implications, we are compelled to choose between the factorial tests and the shortfall tests of Table 2.14.2, and our choice must favour the factorial analysis, not only because by ignoring the factorial analysis the tests in Table 2.14.2 are less efficient, but because it is wrong for scientific reasoning to turn a blind eye to known facts. Moreover, and this is the point here, those facts arise from non-statistical considerations; the tests in Table 2.14.2 are purely statistical, whereas our understanding of the causative agents that lead to the data on which those tests rely, is substantive rather than statistical. That is after all the crux of the distinction that we are always compelled to draw between an experiment and a survey.

Example 2.14.2

Consider the co-ordination tests at (2.12.12) which, *on statistical grounds,* compel us to agree that Snedecor's beef-cereal-pork data are satisfactorily modelled in terms of:

> zero interaction of beef vs. pork × high vs. low, and
> zero main effect of beef vs. pork, respectively,

and so *by implication* compel us to agree that those data are satisfactorily modelled in terms of:

> expected gains that are substantially the same from beef and pork.

Here the implication is *on substantive grounds*, as in fact we indicated at (2.12.13) and (2.12.14), where we advanced good reason for an implicational deduction rather than employing statistical tests of the simple effects numbered 3 and 4 in Table 2.12.4, i.e. good reason for reasoning on *substantive grounds* rather than *statistical grounds*.

Discussion of Examples 2.14.1 and 2.14.2

We have stated that these examples represent a universal agreement. Here there is no room for a contrary view, as our literature contains innumerable examples of factorial experiments (as opposed to surveys) where such reasoning is in evidence. Thus, any attempt to deviate from the rule given as Theorem 2.14.1 must necessarily involve self-contradiction.

Theorem 2.14.1 (The rule of scientific implication):

> It is impossible for any valid defence of a scientific conclusion not also to be a valid defence of any of its implications; and it is always wrong of an investigator to turn a blind eye to such implication. In short: scientific knowledge cannot be compartmentalised.

We have in fact already by implication committed ourselves to this rule by way of the development in Section 2.9. The reader might wonder whether it is really necessary to formulate this rule; after all, the most fundamental method of scientific research is to deduce the implications of a scientific model in order to then test the model by testing its implications against experience. However, it will be found that statistical inference often overlooks this. In Chapter 13, for instance, it will be found that Theorem 2.14.1 has devastating consequences for simultaneous statistical inference.

2.15 The notion of probability inference

Inasmuch as co-ordination testing deliberately avoids any notion of the probability of its findings being erroneous or not, such testing runs counter to the main thrust of the received theories of statistical inference. The reasons for this will be developed fully in subsequent chapters. In the meantime a brief explanation of the terrible difficulties that beset those theories will encourage the reader to reserve judgement. So, let us consider the commencement tests performed at (2.12.2), (2.12.3) and (2.12.4), where we found

that, in respect of symmetry, kurtosis and homogeneity of variance, respectively, our model of the error variation, as tested, fits the given data *acceptably* well. Thus, if that *acceptability* is tempting us into errors, those would have to be in the nature of Type II errors. And, supposing then that one could somehow account for the probabilities of such errors, i.e. ignoring (for the moment) the moot point of whether such probabilities can at all be defined meaningfully, let us denote those probabilities of skewness, non-normal kurtosis and heterogeneity of variance as β_1, β_2, and β_3, respectively. Consider then any one of the elimination tests subsequently developed in Section 2.12, for instance the test on the left at (2.12.16), which shows that any tenable model for the given data requires a simple effect of meat vs. cereal at the high level of protein that is substantially larger than zero, and let us then ask what values of that effect are tenable in respect of the given data. In order to address this question, a standard recipe in the literature of statistical inference would have us compute one or more interval estimates of that effect by bounding the pivotal quantity on the right-hand side of the equation at (2.12.10) by a pair of specified values, –g and +g (g > 0), as follows:

$$-g < (C-\Gamma)\div(S \times \sqrt{\frac{D}{n}}) < +g, \text{ where } \sqrt{\frac{D}{n}} = \sqrt{\frac{3}{20}}$$

Γ denotes the effect of interest

C, the estimator of the effect of interest, equals +13.85

S, the estimator of the error standard deviation, equals 14.65

Solving for the possible values of Γ we find the requisite intervals to be of the form:

$$+13.85+5.67(-g) < \Gamma < +13.85+5.67(+g) \text{ for appropriately specified g,}$$

where that brings us to the crux of the problem, namely how to specify g appropriately. Let:

$$(C-\Gamma)\div(S \times \sqrt{\frac{D}{n}}) \text{ denote a random variable G.}$$

Then our problem is to specify the value g of G such that:

$\Pr(-g < G < +g) = 1-\alpha(\beta_1, \beta_2, \beta_3)$ for specified $\alpha(\beta_1, \beta_2, \beta_3)$, where $\alpha(\beta_1, \beta_2, \beta_3)$ depends on the unknowns previously denoted by β_1, β_2, and β_3.

The problem is insoluble. But statistical inference refuses to admit that and tries instead to defend this, that or the other silly 'solution', depending on the brand of inference. In the most commonly occurring brand of inference, the so-called solution is to employ the following exercise in circular reasoning:

$\{\beta_1 = 0, \beta_2 = 0, \beta_3 = 0\}$ is a 'reasonable assumption', *owing to which*

G is distributed as Student's *t*, *owing to which*

$\alpha(\beta_1, \beta_2, \beta_3)$ is specifiable as $\alpha(\beta_1, \beta_2, \beta_3) = \alpha$ for any α in $0 < \alpha < 1$.

It is at this stage of our development not necessary for the reader to take a stance on the matter. For the time being all that is required is an open mind.

Chapter 3

DECISION-MAKING UNDER RISK

Populations being brought into the real world

3.1 Introduction

This book is not about decision-making under risk; it is about data analysis. However, in much of the statistical literature the distinction between these two different kinds of activity is not at all clearly drawn, with confusing consequences. So we need to come to grips with some of the ideas of decision-making under risk, at least to the extent of being able to recognise them when they try to invade the domain of data analysis. We will try to do so by presenting various examples, where the import of each example is then brought forward by way of a debate between two parties, a decision-maker and a data analyst, where the latter is a visitor in the domain of the former.

3.2 Decision-making under risk as opposed to data analysis

As the theory of hypothesis tests is the main source of the confusion we must try to clear up, the main thrust of the present chapter is to show that hypothesis tests are not properly directed at data analysis; they are directed at decision-making under risk. The following two examples of such decision-making will help explain this.

Example 3.2.1

Suppose that a decision-maker averages replicate measurements, made on specimens of amnion fluid from supposedly pregnant rabbits for sale, so as to provide assurance against phantom pregnancies. Let a historical record of such measurements show that they may be represented as realisations of independent $N(\mu , \sigma^2)$ random variables for a known value of σ^2, with μ denoting unknown numbers of foetuses present; $\mu = 0$ in the case of a phantom pregnancy. Let the inconvenience of an unplanned replacement be preferred over the embarrassment of having sold a barren animal. So, let the decision-maker average as many replicate measurements as needed to achieve:

> Type I errors (barren females erroneously sold as 'pregnant'), and
> Type II errors (pregnant females erroneously replaced as 'barren'),
> at rates controlled to be $\alpha = 0.02$, and at most $\beta = 0.20$, respectively. (3.2.1)

By solving for $\overline{X}$ and $\sigma_{\overline{X}}$ from:

$$\Pr(\overline{X}-\mu \geq +2.054\ \sigma_{\overline{X}} \mid \mu = 0) = 0.02 \text{ and } \Pr(\overline{X}-\mu \leq -0.842\ \sigma_{\overline{X}} \mid \mu \geq 1) \leq 0.20$$

the decision-maker finds that the number of replicate measurements per animal needs to be such that $\sigma_{\overline{X}} = 0.3453$, where the decision rule must be:

Classify any pregnancy as 'phantom' if and only if $\overline{X} < 0.71$.

Let a visitor to the decision-maker's workplace then witness how a population (a host of individuals) is being brought into the real world, as follows:

... $\{\overline{X} = 0.31$, 'phantom'$\}$ $\{\overline{X} = 0.92$, 'not phantom'$\}$ $\{\overline{X} = 0.70$, 'phantom'$\}$.

Visitor: 'Just a moment. Surely your last classification is wrong, is it not?'

Decision-maker: 'No. As $0.70 < 0.71$, I must classify the pregnancy as "phantom".'

Visitor: 'But when the mean, being symmetrically distributed round μ, turns out to be 0.70, much nearer to $\mu = 1$ than to $\mu = 0$, it would surely be odd, even eccentric, to be of an opinion that favours $\mu = 0$ over $\mu = 1$.'

Decision-maker: 'I agree. But my decisions are not intended to service opinions about this, that, or the other solitary individual being classified, as I am not interested in *the individuals* as such. My interest is only in *the host* that is to be comprised of them. My decisions are directed at ensuring that *the host* will meet specifications. Hence there is nothing odd or eccentric about my decisions. To you they only *seem* to be odd, *as you have mistaken decision-making under risk for data analysis*. The question here is not "How might this individual, $\overline{X} = 0.70$, *have come about*?" The question here is "How might a host of many individuals, such that $\alpha = 0.02$ and $\beta \leq 0.20$, *be brought about*?" A decision-maker proceeds from Definition 1.2.2 rather than from Definition 1.2.1.'

Example 3.2.2

Snedecor and Cochran (1989, p. 53) consider a decision-maker who wants assurance that the average content of the active ingredient in certain roots containing insecticide, will be at least eight parts per hundred, apart from a 1-per-100 chance of error. The roots are available in batches, and the decision-maker must accept, or reject, per batch, on the basis of the mean amount of the active ingredient in nine bundles of roots drawn from that batch. A historical record shows that the means can be represented satisfactorily as realisations of independent normal random variables whose expectations represent the amounts of insecticide for corresponding batches, and whose standard deviations equal $3.30 \div \sqrt{9}$, i.e. 1.10, in parts per 100. So, solving from

$$\Pr[(\overline{X}-\mu) \div 1.10 \geq 2.33 \mid \mu \geq 8] \leq 0.01,$$

the decision-maker finds that any given batch is to be classified as 'acceptable' if and only if the mean of nine bundles drawn from that batch exceeds $8 + 2.33(1.10)$, i.e. 10.56, in parts per 100. Let a visitor to the decision-maker's workplace observe how a population (a host of individuals) is being brought into the real world, as follows:

... $\{\overline{X} = 8.13$, reject$\}$ $\{\overline{X} = 10.57$, accept$\}$ $\{\overline{X} = 10.56$, reject$\}$.

Visitor: 'Just a moment. The last two terms in this array puzzle me. Surely it would be statistically naïve to be of the opinion that two means, 10.57 and 10.56, each subject to a standard error of magnitude 1.10, arose from substantially different batches?'

Decision-maker: 'I am not of such naïve opinion. You have *mistaken* two different *decisions* for two different *opinions* concerning the two individuals in question. My decisions do not express opinions about the individuals involved. My decisions are directed at fabricating a host of individuals, such that *the host* will be as specified.'

The reader is challenged to note carefully that *mistaken* reasoning need not be *wrong*. On the contrary, in each of the foregoing examples, the visitor's reasoning is entirely correct for that at which it is directed, but that at which it is directed has mistaken the decision-maker's purpose.

3.3 Decision-making under risk in case of just one population

In certain games of chance, a participant (a decision-maker under risk) will reason in terms of repetitive sampling from a singleton. Two players might for instance wager on the outcome of rolling an ordinary six-sided die. The population is then envisaged as an infinite pool of the outcomes 1, 2, 3, ..., 6 in equal proportions, with the players sampling the population over and over again and, prior to each sample, deciding how to wager. For our purposes, this kind of decision-making under risk is of little interest, as our interest is in cases involving different populations.

3.4 Decision-making under risk in case of different populations

In Example 3.2.1 the model used envisages samples drawn from *different* populations corresponding to *different* pregnancies. These different populations fall into groups of different *kinds:* $\mu = 0$, $\mu = 1$, $\mu = 2$, $\mu = 3$, ... Define, for each sample, a count of 'one error made' when a classification error is made, and when not, a count of 'zero errors made'. Let x_{ij} denote the count for the J^{th} population of the I^{th} kind. Let the counts be arrayed in different columns corresponding to the different kinds, as follows:

$$\begin{array}{cccc} x_{01} & x_{11} & x_{21} & x_{31} \ldots \\ x_{02} & x_{12} & x_{22} & x_{32} \ldots \\ x_{03} & x_{13} & x_{23} & x_{33} \ldots \\ \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \end{array} \qquad (3.4.1)$$

Let the long-run frequencies of 'one error made' in the different columns be denoted by:

$$\alpha, \beta_1, \beta_2, \beta_3, \ldots,$$

respectively. For a mathematically tractable representation of such long-run physics, we envisage the successive columns as independent outcomes of Bernoulli variables denoted by:

$X_0, X_1, X_2, X_3, \ldots$, respectively, where
$E(X_0, X_1, X_2, X_3, \ldots) = (\alpha, \beta_1, \beta_2, \beta_3, \ldots)$.

Let $f_1, f_2, f_3, \ldots (f_1+f_2+f_3+\ldots = 1)$ denote the relative frequencies with which various populations of the kinds $\mu = 1$, $\mu = 2$, $\mu = 3$, ..., respectively, happen to contribute to the corresponding columns at (3.4.1). Here we make no assumptions as to how those frequencies come about; we accept them as constants. Then the expected value of the Type II error rate is:

$$f_1\beta_1 + f_2\beta_2 + f_3\beta_3 + \ldots .$$

Henceforth we denote such an expected Type II error rate as $\overline{\beta}$. Now it might seem that the notion of such a rate does not apply in the case of the insecticide roots, as the class index in that case ranges over a continuum of values, and not, as in the case of the pregnancies, over discrete values such as 0, 1, 2, ... We note, however, that for almost all, if not all, purposes of substantive science, the range of a class index can be taken as bounded between finite (though not necessarily explicit) limits. In the case of the insecticide roots, for instance, the class index is explicitly bounded from below by zero, and it is also bounded from above by some or other definite, though not explicit, upper bound. So, by rounding the possible values of such a class index onto a finite but sufficiently fine grid of class marks, we can satisfactorily approximate any class of statistical models that would interest, for instance, the customers of Gupta and Panchapakesan (1979). For those customers, this establishes Theorem 3.4.1.

Theorem 3.4.1:

Consider an array of samples $S_1, S_2, S_3, \ldots$, drawn from an array of populations $P_1, P_2, P_3, \ldots$, respectively – just one sample per population. Let R denote a decision rule such that should R be applied in repeated sampling from any one of the populations, the Type I error rate is given by α (a specified fraction). Then, should R be applied to the array of samples, one each from the different populations, the Type I error rate is also given by α, and the expected value of the overall Type II error rate $\overline{\beta}$ is given by the mean of the Type II error rates that would arise should R be applied in repeated sampling from the individual populations.

We note that if assurance can be given that each of the constituents of $\overline{\beta}$ is bounded from above by a specified β, then $\overline{\beta} \leq \beta$. Thus we have the assurance in Example 3.2.1 that $\overline{\beta} \leq 0.20$.

3.5 Statistical thinkers versus statistical doers

We must distinguish between *intellectual* rewards and *physical* rewards. In the previous chapters we considered intellectual rewards being reaped in the conceptual world of the human mind, in the world that is 'inside of us', in the world of the thinker. In the present chapter we must consider physical rewards being reaped in the real world of the human body, in the world that is 'outside of us', in the world of the doer. A doer will of course do after forethought only, but any such forethought would be pointless without the intent to reap its rewards and, in the statistical case, that can be accomplished only through repetitive decision-making. A single, solitary decision in the real world can be at error or not at error, but it is utterly meaningless to speak of such a decision as being subject to error rates. In the previous chapters we began to lay a foundation that would enable us to grasp the foregoing

distinctions. And we did so by emphasising persistently that *repetitive sampling* takes place in the human mind only, whence it might then provide an explication of the possible origins of *a single, solitary pseudo-sample in the real world*. The present chapter develops that foundation further by emphasising persistently that, whilst repetitive sampling can provide representations of decision-making subject to error rates, it can do so *usefully*, only with reference to *a host of pseudo-samples in the real world*. The matter is crucial; so our choice of language must draw a sharp distinction between the statistical *investigator* whose thinking brings repetitive sampling into the conceptual world, and the statistical *decision-maker* whose repetitive doings bring pseudo-sampling into the real world.

3.6 The Neyman-Pearson lemma for decision-making under risk

Suppose that a certain rule for decision-making under risk would enable us to control the Type II error rate, for any specified Type I error rate, when repetitively sampling any one member of an array of singletons. Then, owing to Theorem 3.4.1, such a rule will also enable us to control the expected Type II error rate, for any specified Type I error rate, when sampling every member of the array just once. A method for finding such rules is provided by a mathematical lemma of Neyman and Pearson (1933). The lemma has several epistemologically different variants. Here we wish to develop the variant appropriate to decision-making under risk. The idea is then that we know, or we are willing to assume, that the one or the other of just two singletons – $H_{0,}$ called *the null hypothesis*, and $H_{1,}$ called *the alternative hypothesis* – satisfactorily models any member of a given array of populations. A prospective array of samples – just one sample per population – is to be drawn, and we must, with as little error as is possible, classify each population as either H_0 or H_1, the possible errors being:

A Type I error when a H_0 population is erroneously classified as 'H_1'.
A Type II error when a H_1 population is erroneously classified as 'H_0'.

Consider dividing the sample space into two disjoint regions, W^a and W^r, and using a rule of the form:

Whenever a data pattern $\in W^a$, *accept* H_0 and reject H_1.
Whenever a data pattern $\in W^r$, *reject* H_0 and accept H_1.

Customarily, and with tacit reference to the null hypothesis, W^a and W^r are called the *acceptance region* and the *rejection region*, respectively. Often too, W^r is called the *critical region*. The Type I and Type II error rates are then given by

$\alpha = \Pr(\text{a sample pattern} \in W^r | H_0)$, and
$\beta = \Pr(\text{a sample pattern} \in W^a | H_1)$,

respectively, where α is also called the *size* of the critical region, or of the test. We would wish to specify the size of the critical region. Often, however, one cannot form any region of the specified size. In order to overcome this, the sample space is divided into three disjoint regions: W^a and W^r as before, and a third region W^b, often called the *boundary region*. A test of any desired size can then be achieved by using an auxiliary random device, as follows: choose W^r to be of size $\alpha\text{-}\gamma \leq \alpha$ for γ as small as possible, and then choose W^b

such that $W^r \cup W^b$ is of size $\alpha+\eta \geq \alpha$ for η as small as possible. It is then always possible to let an auxiliary random device have 'reject H_0' and 'accept H_0' as exhaustive and mutually exclusive outcomes, with the probabilities of these outcomes in the ratio:

$$\Pr(\text{reject } H_0):\Pr(\text{accept } H_0)::\gamma:\eta.$$

A test of the desired size is then achieved by means of the rule:

Whenever a data pattern $\in W^a$, accept H_0 and reject H_1.
Whenever a data pattern $\in W^r$, reject H_0 and accept H_1.
Whenever a data pattern $\in W^b$, let the auxiliary device 'decide'.

For example, consider drawing haphazardly, and with replacement, just three chips in succession from each one of an array of bowls, the content of each bowl comprising green chips and yellow chips, either in equal proportions (H_0), or with twice as many green chips as yellow chips (H_1). If G denotes green and Y denotes yellow, the sample space comprises eight possible outcomes, as follows:

{GGG, GGY, GYG, YGG, GYY, YGY, YYG, YYY} (3.6.1)

Let the binomial model Bn(θ, n), where $\theta = \Pr(G)$ and n = 3, be deemed appropriate. Then H_0 and H_1 denote the singletons indexed by $\theta = (1/2)$ and $\theta = (2/3)$, respectively. If we choose any one of the eight descriptions at (3.6.1) as a critical region, the size of our test, α, equals $(1/2)^3$, where this is the smallest size we could achieve without using an auxiliary random device. However, by using such a device, $\alpha = (1/2)(1/2)^3$, for instance, can be achieved by choosing

W^a = any seven of the eight descriptions at (3.6.1),

W^b = the remaining one of the eight descriptions at (3.6.1), and (3.6.2)

$W^r = \Phi$ (the empty set),

and then proceeding as follows:

Whenever a data pattern is $\in W^a$, classify the bowl involved as H_0.
Whenever a data pattern is $\in W^b$, flip a coin of balanced sort and so, depending on the outcome, classify the bowl involved either as H_0 or as H_1, with equal chances.

At (3.6.1) there are eight different outcomes from which to choose a boundary at (3.6.2), and of these GGG is obviously the least like an H_0 outcome. So one can anticipate that of these eight possible tests the one with the smallest Type II error rate is obtained by choosing W^b = {GGG} at (3.6.2). With just two singletons from which to choose, as is the case here, such minimisation of the Type II error rate for any specified Type I error rate α ($0 < \alpha < 1$), is always achievable according to the variant of the Neyman-Pearson lemma that concerns us here. That variant will now be introduced using the concept of *a critical function* δ, defined below, where p denotes the disposable probability that the 'decision' of the auxiliary device is 'reject H_0', and is to be disposed of such that the test is of precisely specified size α:

$\delta(S) = 0$ if S is a sample with data pattern $\in W^a$.
$\delta(S) = p$ if S is a sample with data pattern $\in W^b$.
$\delta(S) = 1$ if S is a sample with data pattern $\in W^r$.

As the auxiliary device and the sample are statistically independent, the probability of rejecting H_0 is given by:

$$[Pr(S \in W^b) \times p] + P(S \in W^r).$$

This is called *the power of the test*. We will have chosen W^b and W^r such that

$$[Pr(S \in W^b | H_0) \times 1] + P(S \in W^r | H_0) > \alpha \leq P(S \in W^r | H_0).$$

This means that p for $0 \leq p < 1$ can always be disposed of such that

$$[Pr(S \in W^b | H_0) \times p] + P(S \in W^r | H_0) = \alpha, \text{ as specified.}$$

Having fixed the value of p, the power of the test in respect of H_1 is then:

$$[Pr(S \in W^b | H_1) \times p] + P(S \in W^r | H_1) = 1 - \beta. \qquad (3.6.3)$$

One wants this probability to be the largest possible for the specified value of α, that is to say, if given a choice between two or more tests of size α, one would choose the most powerful of those tests, i.e. the one with the smallest Type II error rate. Theorem 3.6.1 tells us how that can be accomplished in the case of just two singletons.

Theorem 3.6.1 (The Neyman-Pearson lemma for decision-making under risk):

Let an array of data sets s_1, s_2, s_3, ..., each of which can be modelled satisfactorily as a sample from one or the other of two singletons H_0, and H_1, respectively, be brought into the real world. Let the probabilities of these data sets, as modelled, be given by:

$Pr(S = s_j | H_0)$ and $Pr(S = s_j | H_1)$, for j = 1, 2, 3, ...

Then an array of hypothesis tests of the form:

$\delta(S) = 0$ if and only if $Pr(S = s_j | H_1) \div Pr(S = s_j | H_0) < c$
$\delta(S) = p$ if and only if $Pr(S = s_j | H_1) \div Pr(S = s_j | H_0) = c$
$\delta(S) = 1$ if and only if $Pr(S = s_j | H_1) \div Pr(S = s_j | H_0) > c$

c and p being chosen such that all the tests are of exact size α, are most powerful in the long run for testing the singleton H_0 against the alternative singleton H_1.

A proof of this theorem is given by Kempthorne and Folks (1971, pp. 316-320). We note that $Pr(S = s_j | H_1) \div Pr(S = s_j | H_0)$ is customarily called *the likelihood ratio*, and tests conforming to the theorem are called *likelihood ratio tests*. In brief, the theorem therefore states that for testing any singleton against any other singleton, and for any specified test size, the likelihood ratio tests are the most powerful. Consider, for instance, an array of real-world data sets that are satisfactorily modelled as binomial samples of size n = 5, with probability of failure given:

either by H_0: $\theta = 0.5$,
or by H_1: $\theta = \theta_1$, where

θ_1 denotes any one particular value such that $0.5 < \theta_1 < 1$. (3.6.4)

Then the likelihood ratio for s_j failures is given by:

$$\begin{bmatrix}5\\ s_j\end{bmatrix}(\theta_1)^{s_j}(1-\theta_1)^{5-s_j} \div \begin{bmatrix}5\\ s_j\end{bmatrix}(0.5)^{s_j}(1-0.5)^{5-s_j}$$

As $\theta_1 > 1-\theta_1$, this ratio is monotone increasing in $s_j = 0, 1, 2, 3, 4, 5$. So the most powerful tests are of the form:

$\delta(S) = 0$ if and only if $s_j < c$,
$\delta(S) = p$ if and only if $s_j = c$,
$\delta(S) = 1$ if and only if $s_j > c$,
for j = 0, 1, 2, ...

If we choose $c = 4$, the critical region $s_j > 4$, is $s_j = 5$, with size $(0.5)^5 = 0.03125$. Suppose however the specified size is $\alpha = 0.05$. Then, taking $s_j = 4$ as the boundary, the size of the boundary is $5(0.5)^4(0.5)$. We note:

$$(0.5)^5 = 5\times0.00625. \quad 0.05 = (5+3)\times0.00625. \quad 5(0.5)^4(0.5) = 25\times0.00625.$$

So, the required random device can take the form of drawing one number haphazardly from 1, 2, 3, ..., 25, where we must then proceed as follows:

Whenever $s_j = 5$ failures, reject H_0.
Whenever $s_j = 4$ failures, employ the device to draw a number.
If the number drawn is 1, 2 or 3, reject H_0.
If the number drawn is larger than 3, accept H_0.
Whenever $s_j < 4$ failures, accept H_0. (3.6.5)

From the expression at (3.6.3) the Type II error rate for this test is found to be:

$$\begin{aligned}\beta &= 1-\{[\Pr(S\in W^b|H_1)\times p]+P(S\in W^r|H_1)\}\\ &= 1-\left\{\left[5\theta_1^4(1-\theta_1)\times\left[\frac{3}{25}\right]\right]+\theta_1^5\right\}\\ &= 1-\left[\left[\frac{3}{5}\right]\theta_1^4(1-\theta_1)+\theta_1^5\right].\end{aligned} \quad (3.6.6)$$

The value denoted by θ_1 at (3.6.4) was *any* value in the interval $0.5 < \theta_1 < 1$. So we have dealt with a more general case than that indicated at (3.6.4) because if we now consider an array of real-world data sets that are modelled satisfactorily as binomial samples of size n = 5, with probability of failure given

either by H_0: $\theta = 0.5$
or by one of H_1: $\theta = \theta_1, \theta_2, \theta_3, ...$, where
$\theta_1, \theta_2, \theta_3, ...$ denote particular values such that $0.5 < \theta_1, \theta_2, \theta_3, ... < 1$, (3.6.7)

then the recipe at (3.6.5) produces an array of hypothesis tests whose *expected* Type II error rate is minimised for the specified Type I rate α, owing to Theorem 3.4.1. Such an array is called a *uniformly* most powerful array of hypothesis tests. From the expression at (3.6.6) the Type II error rate for this array of tests is now found to be

$$\bar{\beta} = 1 - \sum_{j} f_j \left[\binom{3}{5} \theta_j^4 (1-\theta_j) + \theta_j^5 \right]. \qquad (3.6.8)$$

This brings us to an important class of problems that is exemplified by the following.

Example 3.6.1

For each one of an array of consignments of a certain type of automobile component, a decision-maker subjects five components to a possibly destructive test so as to decide whether that consignment is to be accepted or rejected. On the basis of historical data, the tests have been so devised that a satisfactory component will survive the test, or be destroyed by it, with equal chances. Let the decision-maker model the situation as at (3.6.7) and so, fixing the Type I error rate at $\alpha = 0.05$, use the recipe at (3.6.5) to minimise the expected Type II error rate, β. Let a visitor to the decision-maker's workplace then observe a population being brought into the real world, as follows, where S denotes the number of failed components and X the number drawn from 1, 2, 3, ... , 25:

... {(S, X) = (1, 7), accept} {(S, X) = (4, 3), reject} {(S, X) = (4, 8), accept}.

Visitor: 'Just a moment. Is it not the case that any information concerning the possible θ values is conveyed by the S-like numbers only?'

Decision-maker: 'That is so.'

Visitor: 'So the X-like numbers convey no information whatsoever concerning those possible θ values.'

Decision-maker: 'That is so.'

Visitor: 'But then the last two terms in your array of responses do not make sense. As the only informative data with respect to θ is S = 4 in both cases, the X-like numbers being utterly uninformative in that respect, it would surely be odd, even eccentric, to draw different conclusions about the two consignments involved.'

Decision-maker: 'That is so. But I have not drawn such conclusions about the two consignments involved. You are *mistaking* two different *decisions* for two different *conclusions*. My decisions are not conclusions about corresponding consignments. My decisions are acts of fabrication that are guided by the principles of a scientific technology for bringing into the real world a host of classified consignments such that *the host* can be expected to meet certain specifications.'

The import of the foregoing kind of example is diagnostic in the following instances:

> In *the use of knowledge* (technology) it is intolerable that we should shy away from using an auxiliary random device when that would clearly help us achieve our goal, which is to bring about a prospective physical outcome (a host that would meet our specifications).
>
> In *the pursuit of knowledge* (investigation) it is intolerable that we should wish to adjoin to our data, after those data have already come about, the outcome of an auxiliary random device when that clearly cannot help us achieve our goal, which is to come to know how a given physical outcome (a data set) might have come about.

Let us recall that any data have always been recorded on a discrete and an essentially finite grid of class marks. So, in principle, randomised hypothesis tests are ubiquitous, and, given their forceful diagnostic value, examples of such tests should deliberately be brought forward and their epistemological implication underscored. But, sad to say, that is by and large not the current statistical way. Instead, randomised tests tend to be treated as something of an embarrassment, and are either not mentioned at all or exhibited briefly and then discounted quickly with words to the effect of:

> 'This is the black sheep of the statistical family. We exhibit him in all honesty as, in contrast to certain colleagues, we hold that science should not shy away from unsavoury facts. And now, having made a clean breast of his disgraceful existence, let us bundle him away, so as to consort further with only the decent members of our family.'

The issue is this: if logical reasoning leads from certain premises to a conclusion that cannot be correct, there must be a flaw in the premises. The puzzlement of our visitor displays that flaw precisely, as the puzzlement arises from statistical decision-making under risk being mistaken for investigative statistics. Conversely, the embarrassment around the existence of randomised tests is symptomatic of a premise that mistakenly takes statistical investigation to be a form of decision-making under risk.

3.7 The first of two different resolutions of the mixed sampling problem

We are now ready to take first steps toward resolving a problem introduced in Section 1.5. As pointed out in Sections 1.5 and 1.13, that problem arises when mathematical statistics formally offers substantive science a choice between different frames of reference, and it is then for *substantive science* to resolve the problem of which frame of reference is, for *its* purposes, the correct one. We now consider two examples, each of which displays the first one of two possible resolutions of the problem. The first of these examples is hardly of any practical interest, but it conveys the present resolution with great force. The second example shows how the present resolution might well be of practical interest.

Example 3.7.1

Using Instrument A, a decision-maker can classify objects for sale as 'good' or 'bad', with misclassification rate zero. Using Instrument B, he can classify those objects as 'good' or 'bad', with misclassification rate 0.10. Using Instrument A is more expensive than using B. A regulatory authority specifies rates in excess of 0.05 as unacceptable. So the decision-maker must, as cheaply as possible, achieve a rate of 0.05. That is accomplished by spinning a coin of balanced sort to determine, for each object separately, whether it is to be classified by A or by B and, without keeping records to show how particular individuals were classified, then to market only those individuals that were classified as 'good'. Let a visitor to the decision-maker's workplace witness how a population of such individuals is being brought into the real world.

Visitor: 'Surely your work would be more informative in respect of every individual if records were kept that showed, for each individual, whether its classification was by A or by B?'

Decision-maker: 'That is indeed correct.'

Visitor: 'Then I cannot understand why you keep no such records.'

Decision-maker: 'Because if I were to do so, I would obtain two identifiably different real-world populations, only one of which would satisfy the regulating authority. All the individuals in the other population would then have to be reclassified. You mistake my decisions for attempts to provide information in respect of this, that or the other individual. It is not my purpose to provide such information. It is my purpose to make a real-world population that would satisfy the regulating authority, and you can surely grasp that I am achieving that at minimum expected cost. *You are mistaking the use of knowledge for the pursuit of knowledge.*'

Example 3.7.2

The expected Type II error rate at (3.6.8) can be re-expressed as:

$$\overline{\beta} = 1 - \sum_j f_j \left[\left(\frac{3}{5} \right) \theta_j^4 + \left(\frac{2}{5} \right) \theta_j^5 \right] .$$

So, let the decision-maker mix sets of n = 4 components and sets of n = 5 components in a 3-to-2 ratio, rejecting any consignment if and only if all tested components fail, and keeping no records of how many components were tested in any particular case.

Visitor: 'Why do you prefer the present procedure to that of Example 3.6.1?'

Decision-maker: 'Because in 3-in-5 cases the present procedure calls for only n = 4 instead of n = 5 components, and yet has precisely the same operating characteristics as that of Example 3.6.1. Thus, for instance, the Type I error rate α remains fixed at

$$\left(\frac{3}{5} \right) \left(\frac{1}{2} \right)^4 + \left(\frac{2}{5} \right) \left(\frac{1}{2} \right)^5 = 0.05,$$

and both procedures are uniformly most powerful in respect of any array of the kind of alternatives that must interest us here. If you cannot follow the mathematics of the matter, I can provide proof by simulation.'

We mentioned in Section 1.5 that the mixed sampling problem is resolved in different ways depending on different purposes of substantive science. Examples 3.7.1 and 3.7.2 both involve *a decision-maker* who wishes to *use* knowledge for the purpose of gaining certain *physical* rewards. And in both examples it is the unconditional frame of reference that provides for that correctly. However, *an investigator* is someone who wishes to *pursue* knowledge for the purpose of gaining certain *intellectual* rewards. Clearly, as indicated by the visitor in Example 3.7.1, it is the conditional frame of reference that would provide for such a pursuit. Otherwise who would be so silly as to hold that a classification by means of Instrument A has a 1-in-20 chance of being a misclassification? We will return to this matter in a subsequent chapter.

3.8 The role of assumptions

It will be found that the literature proper on decision-making under risk has little, if anything, to say about commencement testing. Instead, a decision-maker is envisaged as proceeding from this, that or the other class of models, and it is taken more or less for granted that the class will be appropriate. There are two reasons for this: firstly, in practice, decision-making is often an ongoing procedure, as in the certification of plant material in nurseries, or in acceptance sampling of raw materials for manufactory. In such cases a historical record enables an appropriate class of models to be chosen. Secondly, in the absence of such a record, it is entirely justified to proceed from an informed guess, that is to say, from a reasonable assumption, as decision-making is a *necessary* activity. So we cannot criticise attempts at achieving a desired goal without suggesting a more appropriate procedure. And if such a suggestion is more appropriate, it will simply be adopted. In short, the decision-maker can always say: 'I *have* to make these decisions, and if you can think of a better procedure than the one I currently use, convince me that your procedure is better, and I will adopt it.'

3.9 When separating characteristics are irrelevant

In each of the debates we envisaged in this chapter, the visitor reasons in terms of *separating characteristics*, whereas the decision-maker reasons in terms of *operating characteristics*. In Example 3.2.1, for instance, the visitor observes (correctly so) that a particular datum $X = 0.70$, where $E(X) = \mu$, *separates* $\mu = 0$ and $\mu = 1$ as the less tenable model and the more tenable model, respectively. Whereupon the decision-maker replies (also correctly so) that the visitor's observation, though correct, is irrelevant for the purpose of realising the *operating characteristics* called for at (3.2.1). Again, in Example 3.2.2 the visitor is puzzled to find data sets, given by

$$(\overline{X}, \sigma_{\overline{X}}) = (10.57, 1.10) \text{ and } (\overline{X}, \sigma_{\overline{X}}) = (10.56, 1.10), \text{ respectively,}$$

seemingly being predicated by:

$$N(\mu, \sigma_{\overline{X}}) \text{ with } \mu \geq 8 \text{ and } \mu < 8, \text{ respectively,}$$

when the data in hand can hardly *separate* those models from each other. Whereupon the decision-maker can reply (correctly so) that the visitor has mistaken decisions for conclusions, where the decisions are those that best serve the purpose of achieving the requisite *operating characteristics*. The issue is most forcefully brought forward by Example 3.6.1 where the values of random numbers that can convey no information whatsoever for the *separation* of the alternative models involved, nevertheless cause a decision-maker to act as if they do convey such information, so as to realise specified *operating characteristics*. Once again: in Examples 3.7.1 and 3.7.2 a decision-maker deliberately suppresses data that would be of use in *separating*, for their tenability, the models being held in the human mind. This is done in order for the rump of the data to service decisions that, owing to that suppression, meet specified *operating characteristics*.

3.10 Randomisation and an old enigma of card play

Fisher (1934) describes a game called Le Her, played by two players, each of whom is dealt just one card by player A, with the cards being valued in order from aces as lowest to kings as highest. Player B may then choose to exchange cards with A, except if A holds a king. Regardless of which card A then holds, A may choose to exchange that card for one drawn from the pack, except if a king is drawn, in which case no further exchange may take place. If the two cards then held are equal, A wins; otherwise the higher of the two cards wins. It can be shown that if, as a fixed rule, B chooses to exchange any card lower than 7, it is, as a fixed rule, to A's advantage to exchange any card lower than 8, and *vice versa*. So it is to A's advantage to follow a like rule with B, and it is to B's advantage to follow an unlike rule with A. However, not knowing what rule an opponent follows makes it unclear what rule a player must follow. In the classical literature there is an unresolved disagreement on the matter. Fisher resolves it by proposing that either player, having to choose between two rules, must do so by a randomised decision, as follows: let A's rule be to change an 8 with frequency P, and B's rule to change a 7 with frequency Q. The game has 5 525 possible outcomes all told, of which the expected number of outcomes where B wins, is then given by

$$2\,828+6P+10Q-16PQ.$$

For Q greater than 3÷8, it is to A's advantage to put P = 1, and for Q less than 3÷8, it is to A's advantage to put P = 0. Hence B should choose Q = 3÷8, in which case A's policy is a matter of indifference to B. Similarly, A should choose P = 5÷8, in which case B's policy is a matter of indifference to A. To the uninformed, the players might then routinely seem to be making absurd use of random devices in order to make up their minds, but there is nothing absurd about such random decision-making. On the contrary, an informed individual who regularly plays Le Her and refuses to use Fisher's randomised decision rules, is being either obstinate or silly.

3.11 The target problem

In this section we give an example of repetitive decision-making under risk involving certain formalities that resurface in subsequent chapters where the reasons for introducing the example will become clear. Consider a series of repetitions within each of which a given number, n, of bullets are fired at a target. The target is then removed and we are allowed to measure the points of impact only. In each of the repetitions, the measurements must be used to estimate the position the target occupied in that repetition. The estimates must take the form of *confidence regions* for the unknown target centre, that is to say, sub-regions of the parameter space such that with a specified long-run frequency $1\text{-}\alpha$ $(0 < \alpha < 1)$, the centre of the target will have been within the sub-region. Let Cartesian co-ordinates (μ_{1j}, μ_{2j}) give the centre of the target in the j^{th} repetition, and let corresponding co-ordinates (x_{1ji}, x_{2ji}) for i = 1, 2, 3, ..., n give the various points of impact in the j^{th} repetition. We suppose that the (x_{1ji}, x_{2ji}) can be represented as realisations of 2n independent homoscedastic normal random variables, (X_{1ji}, X_{2ji}), with expectations

$$E(X_{1ji}, X_{2ji}) = (\mu_{1j}, \mu_{2j}), \text{ where i = 1, 2, 3, ... , n and } j = 1, 2, 3, \ldots .$$

Consider a standard 'between-within' analysis of variance in repetition *j*, where

$\overline{X}_{1j}$ and $\overline{X}_{2j}$ denote the means of the X_{1ji} and X_{2ji} groups, respectively, and

S_j^2 denotes the pooled variance estimate.

Let α denote a specified error rate $(0 < \alpha < 1)$.

Then it can be shown that:

$$\Pr\left\{[(\overline{X}_{1j}-\mu_{1j})^2+(\overline{X}_{2j}-\mu_{2j})^2]\div(2S_j^2\div n) \le F(\alpha)\right\} 1 - \alpha, \qquad (3.11.1)$$

where $F(\alpha)$ denotes the percentage point exceeded with probability α by Snedecor's F on 2 and 2(n-1) degrees of freedom. By solving for (μ_{1j}, μ_{2j}) from the equation at (3.11.1), we obtain an array of confidence regions in the form of circular disks centred at

$(\overline{x}_{1j}, \overline{x}_{2j})$, and with radii $\sqrt{(2s_j^2\div n)F(\alpha)}$, when j=1, 2, 3, ...

Consider $\alpha = 0.05$ and n = 10. Then we have $F(\alpha) = F(0.05)$ on 2 and 2(10-1) df, which equals 3.55. Thus the 95% confidence regions for the unknown target centres take the form of a series of circular disks, as exemplified in Table 3.11.1.

Table 3.11.1: Circular 95 % confidence regions for an array of target centers

For Target 1: Centered at (3.1, 7.4) with radius $\sqrt{2[(0.389)^2\div 10]3.55} = 0.33$
For Target 2: Centered at (2.7, 9.8) with radius $\sqrt{2[(0.952)^2\div 10]3.55} = 0.80$
For Target 3: Centered at (9.4, 5.3) with radius $\sqrt{2[(0.310)^2\div 10]3.55} = 0.26$
...

This allows for a different error variance from one repetition to the next, which serves the further purposes of this example best.

3.12 Decision-making under statistical risk. Is it of much practical importance?

Much of the statistical literature concerns repetitive decision-making under risk, but provides scant appropriate real-world examples. For instance, in the first 23 chapters of a book on statistical decision theory, Gupta and Panchapakesan (1979) envisage many possible problems in repetitive decision-making under statistical risk, and develop recipes for their solution. In the 24th chapter, the authors then propose to exemplify the practical value of those recipes by applying them to 20 real-world examples. But the 20 examples concern problems in data analysis, not problems in repetitive decision-making under risk. As an introductory example, for instance, they consider the results of a yield trial with seven varieties of barley as Repetition 1 in an array of such repetitions of application of a so-called subset selection procedure whose purpose it is to select, in each repetition, a subset of the seven varieties, such that 'the best variety' will be included in the subsets in a specified proportion, or more, of repetitions. However, the barley data involved were first placed in the literature by Duncan (1955) and there has never been the slightest prospect that Repetitions 2, 3, 4, ..., would appear during the ensuing half century. Similarly, in most, if not all, of the other 19 examples, we find a single, solitary real-world data set whose possible origin is to be explained in the conceptual world of the human mind. So all 20 examples are inappropriate, which compels us to ask why appropriate examples were not given. If we consult other sources, the same inappropriate exemplification is found. Gibbons, Olkin and Sobel (1977) for instance, try to complement 'theory' books like that of Gupta and Panchapakesan, with a 'methods' book. So they present many purported examples for decision-making under risk, but again we find that time and again we are presented with an example requiring data analysis, rather than decision-making under risk. Thus we are compelled to conclude that statistical decision-making under risk tends to be more of an ivory tower topic than one of real-life importance. This conclusion might of course be wrong, in which case we must plead with advocates of decision theory to develop the importance of the topic by using appropriate examples; otherwise we can but conclude that they do not have such examples.

3.13 Concluding remark on a counter-argument

It might be countered that the argument in the previous section involves an oversight in that, if in each of many different cases we were to apply different rules of decision, but such that each rule would, if conceptually applied to an array of appropriate cases, lead to a specified proportion P of correct decisions, the overall proportion of correct decisions would also equal P. We reply: That is correct, but as a counter-argument to that in Section 3.12, it fails because it leads to insurmountable difficulty. We will return to this point in section 4.29.

Chapter 4

INVESTIGATION MISTAKEN FOR DECISION-MAKING UNDER RISK

The frequentist vicious circle

4.1 Introduction

We must now take first steps toward coming to grips with what can only be described as the *idée fixe* of mathematical statistics. The idea originated perhaps with Bayes (1763) and, if not, then with Laplace (1814). Its essence envisages the development of methods that would enable investigative statistics to qualify its findings by statements of the kind ' ... and there is all of a 0.95 probability that this finding is correct', or ' ... and the probability that this finding is incorrect, is a trifling 0.01'. The idea has divided the statistical profession into different schools of thought, marred the statistical literature with controversy and, some 200 years on, has failed to produce consensus. So, to hold that the idea is a will-o'-the-wisp is not so eccentric a stance that it cannot merit serious consideration. In this chapter we show how the idea leads to 'frequentist inference', so named because such inference holds that the probabilities in question can be meaningful in the sense of long-run frequency only. Also, *and more importantly*, it holds that this is achievable by adapting the reasoning of decision-making under risk to investigative needs. The reader may anticipate that problems in investigation are thereby mistaken for problems in decision-making under risk. The reader should also note that such mistakes are the converse of those displayed in Chapter 3, where it was shown how problems in decision-making under risk might be mistaken for problems in investigation. In order for this to be clearly understood we must grasp the distinction between reasoning that is mistaken and reasoning that is simply wrong. For instance, certain South African jokes poke fun at an imaginary simpleton named Koos van der Merwe. According to one such joke Koos brings a ladder to a party because he was told the drinks will be on the house. Koos mistook 'on the house' to mean 'on the roof', rather than 'supplied by the hosts'. He *mistook* the figurative meaning for the literal one – which is not at all the same as simply being *wrong.* On the contrary, in Koos's understanding it was not at all wrong to have brought a ladder in order to be able to get to the drinks. Similarly, this chapter shows how a mathematical ingenuity called 'frequentist inference' tries to make decision-making under risk serve the purposes of data analysis without realising that its efforts are based on a mistaken understanding of such analysis.

4.2 The problem of attaining real-world repetitions

Let the reader ask: 'Is there another person identical to me? Has there ever been such a person? Will there ever be such a person?' These questions can be answered in the negative only since every individual is unique. This does not apply only to persons it applies to every single object in the real world and certainly to every data set. It must be grasped firmly that when we speak of 'W. C. Krumbein's data' as opposed to 'J.H. Edward's data', we draw a distinction that is vastly more complicated than a distinction between two different sets of numbers. So we introduce Definition 4.2.1.

Definition 4.2.1:

The term *data set* refers to a real-world individual and must be understood to embrace all the circumstantial details that constitute that individual's uniqueness, and are thereby limited to that individual's proprietary.

It might be objected that: 'My brother and I share our mother, who is therefore not limited to his proprietary'. Then we must reply: 'You and your brother share the same *sort* of relationship to your mother, but your *particular* relationship to her is limited to *your* proprietary'.

Definition 4.2.1 raises a problem that *must* be solved if the ideas of frequentist inference are to have any hope at all of practical implementation. We refer to it as *the problem of attaining real-world repetitions*. In order to come to grips with it, let us re-visit the waiting times for the eruptions of Vesuvius (Tables 1.15.2 and 1.15.3). Frequentists would have an investigator reason in terms of populations of repetitions being brought into the real world, as the idea is to exercise *real-world* control over the frequencies of Type I errors and Type II errors. So, for the Vesuvial data, consider a population of repetitions arising as follows:

> Repetition 1: The 1st universe originates. The 1st earth cools, and Vesuvius the 1st appears. In A1stD 79 Vesuvius the 1st erupts. A 1st series of waiting times follow.
>
> Repetition 2: The 2nd universe originates. The 2nd earth cools, and Vesuvius the 2nd appears. In A2ndD 79 Vesuvius the 2nd erupts. A 2nd series of waiting times follow.
>
> Repetition 3: The 3rd universe originates. The 3rd earth cools, and Vesuvius the 3rd appears. In A3rdD 79 Vesuvius the 3rd erupts. A 3rd series of waiting times follow.
>
> ... . (4.2.1)

As a matter of real-world experience such nonsense will clearly not do. So frequentist inference is faced with the problem of how substantive investigators are supposed to attain the proposed populations of real-world repetitions. Do not think that this problem is peculiar to certain types of data only. Consider trying to persuade an agronomist to repeat a replicated yield trial with oats. *Repeat* it? Yes, say about a thousand times. *A thousand times!* Well, how about a hundred times? *Don't be daft!*

4.3 Stereotypic arrays

Inasmuch as frequentists must achieve real-world error rates, Neyman (1952) clearly realised that imaginary nonsense repetitions such as those considered at (4.2.1) simply will not do. Instead he proposed essentially as follows: let each one of the D-like symbols below denote a decision to accept or reject a null hypothesis, where each successive array in turn represents conceptual repetitions of a given hypothesis test.

> Test 1: D_{11}, D_{12}, D_{13}, ..., of which a proportion, precisely α, are Type I errors.
> Test 2: D_{21}, D_{22}, D_{23}, ..., of which a proportion, precisely α, are Type I errors.
> Test 3: D_{31}, D_{32}, D_{33}, ..., of which a proportion, precisely α, are Type I errors.
> .... . (4.3.1)

Now form an array of conceptual decisions by randomly selecting just one decision from each of the arrays at (4.3.1). Then we obtain

> $D_{1.}$, $D_{2.}$, $D_{3.}$, ..., of which a proportion, precisely α, are Type I errors. (4.3.2)

Choosing $\alpha = 0.05$, the decisions in this last array might then be represented as being brought into the real world, for instance as follows:

> $D_{1.}$: A χ^2 test for random dispersion leads to $\chi^2 = 15.50$ on 8 df. The 0.05 critical region for this test is $15.51 \leq \chi^2 < \infty$. Therefore H_0 is accepted, i.e. our decision is that the dispersion is random.
>
> $D_{2.}$: A Raleigh test for isotropic directions leads to $q^2 = 18.13$ for $n = 10$ directions. The 0.05 critical region for this test is $29.95 \leq q^2 < \infty$. Therefore H_0 is accepted, i.e. our decision is that the directions are isotropic.'
>
> $D_{3.}$: A Cramer-Von Mises test for exponential waiting times leads to CvM = 0.225. The 0.05 critical region for this test is $0.224 \leq \text{CvM} < \infty$. Therefore H_0 is rejected, i.e. our decision is that the waiting times are not exponential.
>
> ... . (4.3.3)

It will be found appropriate to call such an array '*a stereotypic array*'. We challenge the reader to note that inasmuch as the idea of such arrays was introduced by Neyman himself, he established beyond reasonable contest that his dual theories of hypothesis tests and confidence regions are directed at *trying to create real-world populations to specification*. The introduction of such arrays would otherwise be pointless.

4.4 A dilemma

We have seen stereotypic arrays originating from certain mathematical thought trying to come to grips with the interface between statistics and substantive science. We will now prove by dilemma that, in so doing, such mathematical thought has tried to foist a creature of its own invention on the discourse of substantive science. To begin with, we

note that Neyman would have us envisage an incongruous real-world population arising from such diverse data sets as:

> A data set on the frequencies of clockwise rather than anti-clockwise dust devils.
> A data set on the ability of various rootstocks to tolerate waterlogged conditions.
> A data set on the numbers of babies born with harelip in Birmingham, England.
> A data set on occupancy behaviour in a certain sort of beetle.
> A data set on the waiting times for the eruptions of Vesuvius.
> A data set on the responses of *Octopus* in a learning trial.
> .... .

The incongruous nature of such a population lands frequentist inference on the horns of a dilemma, as follows: consider (and this leads to the first horn of the dilemma) whether or not the foregoing population, *as such*, makes *substantive* sense:

> 'What possible bearing,' hortology will ask, 'can the proportions of clockwise and anti-clockwise dust devils have on the survival of waterlogged root stocks?'
>
> 'What conceivable bearing,' entomology will ask, 'can the occurrence of babies with harelip have on the occupancy behaviour of beetles?'
>
> 'How on earth,' ethology will ask, 'can the eruptions of Vesuvius be related to the learning abilities of *Octopus*?'
>
> ... .

This forces us to recognise that the incongruous population has not arisen as a concept of substantive science; it is a creature of mathematical statistics. Now consider (and this leads to the second horn of the dilemma) how the incongruity might be removed. There is only one way. We will somehow have to so restrict the stereotypic arrays that they cannot be held to be incongruous. Consider a yield trial with oats. If one tries to devise an appropriate stereotypic array for the given trial by limiting the further terms of the array to yield trials with oats, the substantive investigator will say:

> 'But the two fertiliser trials I am conducting in the Swartland are not repetitions of the cultivar trial I am conducting in the Overberg. Moreover, the fertiliser trials are at Riebeeck Kasteel and Klawer, respectively, and are not repetitions of one another. Should these trials not be discernibly different in subject matter of interest to me, I would have treated them as constituents of a single trial.'

Again, an eruption of Vesuvius is not a repetition of an eruption of Krakatoa, and neither of the two is a repetition of an eruption of Paricutin. The reader will note that this argument has driven us back into trying to defend the silly populations envisaged in Section 4.2. Hence, on each horn of the dilemma, we are compelled to recognise that any populations arising from the stereotypic arrays are artefacts of mathematical reasoning. That shows that such reasoning is trying to solve a problem of its own invention, and not one arising from the needs of substantive science. This concludes our proof, by dilemma, of Theorem 4.4.1.

Theorem 4.4.1:

Apart from decisions arising from the accept-reject rules involved, different terms in a stereotypic array cannot *as such* be referred to in evidential terms. They can, *as such*, be identified in terms of non-evidential labels only.

Thus, the data sets considered in Examples 1.21.4 and 1.21.5 may be distinguished by being labelled as W.C. Krumbein's data and J.H. Edward's data. And, if imbedded in a stereotypic array of tests, say of size $\alpha = 0.001$, we may state that in each term a hypothesis test of size $\alpha = 0.001$ using Raleigh's test statistic, rejects the hypothesised model. However, any further evidential facts, such as those discussed in Section 1.21, may not be brought into consideration, as that would have us stray from the normative prescriptions required for the specification of a stereotypic array. And *that*, after all, is precisely why frequentists always have to insist that the specification of a Type I error rate must be *without reference to the data*.

4.5 Stereotypic arrays destroy statistical evidence

In Section 1.14 we met certain normative prescriptions. Section 4.4 enabled us to grasp that the prescriptions originate from reasoning that mistakes investigation for decision-making under risk. In Sections 1.14 and 2.9 we displayed the destructive effect of such reasoning, both on substantive evidence and on statistical evidence. An especially revealing way of grasping the destruction of statistical evidence is provided by shortfall testing, as the following example shows.

Example 4.5.1

Snedecor and Cochran (1989, p. 382) use covariance adjustments of the yields of six maize cultivars to account for differences in stand. Table 4.5.1 displays the adjusted means and a suite of shortfall tests for the elimination of cultivars whose yields might seem lower than best.

Table 4.5.1: A suite of five shortfall tests using Dunnett's many-one-*t* statistic.

Cultivar	Adjusted mean	Shortfall	Many-one-*t*	Leftmost co-ordinates
D	219.32			
F	213.66	5.66	0.781	(0.478, ε, 0.522)
C	193.16	26.16	3.610	(0.994, ε, 0.006)
A	191.80	27.52	3.797	(0.996, ε, 0.004)
B	190.98	28.34	3.910	(0.997, ε, 0.003)
E	189.58	29.74	4.104	(0.998, ε, 0.002)
Estimated standard error of a difference between two means = 7.247 on 14 df.				

The estimated standard error shown in the table is obtained by using the approximation of Finney (1946). An empirical investigation by Sadie (1996) indicates that such approximation is reliable in the case of shortfall testing, provided that the covariate treatment means are statistically homogenous. For the case in hand, such homogeneity is indicated by $F = 1.21$ on 5 and 15 df., whose mental correlate is to be found at (0.65, ε, 0.35) in Snedecor's test distribution. The results given in Table 4.5.1 are compelling: there is no evidence that D or F could be lower than best, as D gave the highest mean, and the mental correlate of the standardised shortfall for F is situated almost exactly on the median of Dunnett's test distribution. As for C, A, B and E, there is compelling evidence that all four are lower than best, as the mental correlates of their standardised shortfall values are all situated far down in the right-hand tail of Dunnett's test distribution. But just look how destructive of this evidence the stereotypic reasoning would be. If $\alpha = 0.01$, poor performances by C, A, B and E are revealed, but an entirely satisfactory performance by F is concealed. If $\alpha = 0.50$, an entirely satisfactory performance by F is revealed, but poor performances by C, A, B and E are concealed. Here any specified value of α destroys important evidence.

4.6 Stereotypic arrays destroy substantive evidence

This hardly requires proof. Consider the array at (4.3.2) possibly arising as follows:

> $D_{1.}$ denotes a decision taken by a botanist investigating plant dispersion.
> $D_{2.}$ denotes a decision taken by a zoologist investigating bird navigation.
> $D_{3.}$ denotes a decision taken by a geologist investigating volcanic eruptions.
> … .

Here possible evidential contributions by botany, zoology, geology, etc. are ignored, as they are peculiar to individual terms. As we emphasised by way of Theorem 4.4.1, stereotypic reasoning cannot recognise that the substantive source of a given term will differ from the substantive source of every other term. Suppose, for instance, that the next term in the present array concerns an animal nutritionist investigating the effect on gained weight, of a supplemented ration for broiler chickens. Let the mean weight from the supplemented ration minus the mean weight from the control ration be equal to 16, whose estimated standard error = 10.0 on 10 df. The statistical co-ordinates directing us to where in Student's distribution the mental correlate of $(16\text{-}0)\div 10.0 = 1.6$ is to be found, are (0.93, ε, 0.07). The nutritionist now might well want to conclude as follows:

> The supplement seems to increase weight gain because, although the co-ordination, considered in itself, is not *very* extreme, observations not of a statistical nature, *as described in the previous section of the report*, indicate that the supplement had a beneficial effect on the metabolism of the animals involved. On physiological grounds, such benefits are expected to promote weight gain.

However, the airing of such opinions is forbidden by the normative prescriptions for constructing a stereotypic array.

> 'No! No! No!' the referees will say. 'It has long been the policy of our journal to commit to $\alpha = 0.05$ Type I error rates. You must report that the supplement was found to have no effect on weight gain, as the value of Student's t was not in the appropriate critical region.'

This is not an unrealistic scenario. It is a sad fact that many referees have been brainwashed by the ideas of frequentist inference into adopting just such silly positions. In an example recently brought to this writer's attention, the external examiner of a PhD thesis insisted on the normative prescriptions 'being the scientific method'.

4.7 Stereotypic arrays involve silly embarrassments

Frequentist inference is related to wagering. For instance, we cannot *retrospectively* wager on a certain horse to be the winner of a given race, unless we can do so *without reference to the outcome of the race*. Similarly, for given data, frequentist inference will allow us to specify the Type I error rate *retrospectively*, but only if we can do so *without reference to the data*. In the following example we show that this cannot be done without courting silly embarrassment.

Example 4.7.1

In the aftermath of World War II, food relief agencies supplied powdered cow milk to supplement the diet of infants in certain communities in Asia. It subsequently turned out that in a high proportion of cases this had been extremely harmful, even mortally so, because of allergic reactions to cow milk (People in Asia use predominantly buffalo milk). In view of this tragic *faux pas* by the food relief agencies it is entirely realistic to consider an investigator who firmly believes that, as supplement in a food ration for ducks, a certain mineral by-product of a given industrial process can only be beneficial or harmless. The investigator will therefore be advised by much of the statistical literature that if δ denotes the effect of the supplement, a hypothesis test of H_0: $\delta = 0$ versus H_1: $\delta > 0$ would be appropriate. Let the investigator specify a Type I error rate of $\alpha = 0.01$ for a hypothesis test using Student's t on 11 df., where that rejects H_0 in favour of H_1 if, and only if, the observed value of t exceeds its upper 0.01 percentage point ($t = +2.72$). Let the observed value of t turn out to be $t = -3.06$, which falls short of the *lower* 0.01 percentage point of t ($t = -2.72$). So the investigator is embarrassed to find that the observed value of t is situated in the lower tail of the test distribution, which tail, as Kendall and Stuart (1961, p.182) put it, has 'turned out to be the wrong one.' In a repetition of the trial, the investigator might then heed the advice of Kendall and Stuart, and so, as a 'common-sense precaution', and without reference to the new data, use an 'unbiased' two-tailed test of H_0: $\delta = 0$ versus H_1: $\delta \neq 0$, at the specified Type I error rate of $\alpha = 0.01$. Let $|t|$ on 11 df. for the new data then turn out to be $|t| = |-3.07|$, which falls short of the 0.01 percentage point of $|t|$ ($|t| = 3.12$). Thus a *doubly embarrassed* investigator twice over obtains forceful evidence that the supplement has a harmful effect. However, by the normative rules of frequentist inference he must pretend to be unaware of that; on the contrary, he must report that the new trial has *confirmed* the finding of the previous trial, to wit: '*The supplement has no effect.*' The source of this nonsense is *inter alia* the mistaken idea that data analysis requires to be governed by controlled error rates.

4.8 Stereotypic arrays cannot avoid circular reasoning

We are now ready to take first steps toward coming to grips with what is by far the most serious defect in the notion of statistical inference. Such inference (and this is the case for *all* the received theories of statistical inference) simply cannot avoid a certain source of circular reasoning without violating some fundamental principle of scientific reasoning. In the case of frequentist inference that circularity is simply as follows: Neyman's stereotypic arrays would have us view any hypothesis test as just one term in a progression of terms, each term taking the form of a decision made at the risk of a Type I error, and subject to the same specified Type I rate from term to term in that array. This involves an oversight because in the case of an elimination test, at least one of the previous terms would involve *acceptance of the class characteristic*. The reasoning thus cannot avoid being circular, because it cannot recognise that such *acceptance* might be subject to a Type II error. The following example displays this form of circular reasoning.

Example 4.8.1

Let an investigator count the numbers of plants of species A in each of n = 4 quadrats and find 0, 2 and 3 of the plants in 2, 1 and 1 of the quadrats, respectively. Let the investigator also count the numbers of plants of species B in each of n = 8 quadrats and find 0, 1 and 2 of the plants in 4, 1 and 3 of the quadrats, respectively. For each species in turn, let μ denote the population mean number of plants per quadrate, and let the investigator wish to test

$$H_0: \mu = 0.5 \text{ versus } H_1: \mu > 0.5. \tag{4.8.1}$$

(Other pairs of alternatives would serve equally well for the present purposes, except that the foregoing makes for tidy arithmetic.) Let the investigator assume that the data sets can be represented as Poisson samples. If X denotes the sample total, then

$$\Pr(X \geq x) = \Pr\left[\chi^2_{2x+2} < 2n\mu\right], \tag{4.8.2}$$

where χ^2_{2x+2} denotes a central chi-square random variable on 2x+2 df.

Using this formula we find that for the alternatives at (4.8.1) and for hypothesis tests of size $\alpha = 0.05$, the critical regions are given, in terms of the sample total, x, by

$$\{4 \leq x < \infty\} \text{ for species A, and } \{7 \leq x < \infty\} \text{ for species B.} \tag{4.8.3}$$

In both cases the observed number of plants belongs to the critical region. So in both cases $H_0: \mu = 0.5$ is rejected in favour of $H_1: \mu > 0.5$, subject to a Type I error rate of $\alpha = 0.05$. Owing to different sample sizes, the Type II error rates differ. For instance, with respect to $\mu = 2.0$ as alternative, the form at (4.8.2) gives these rates as follows:

$$\text{For species A: 1-Pr}\left[\chi^2_{2(4)+2} < 2(4\text{x}2.0)\right] = 0.10, \text{ denoted by } \beta_A \text{ say .} \tag{4.8.4}$$

$$\text{For species B: 1-Pr}\left[\chi^2_{2(7)+2} < 2(8\text{x}2.0)\right] = 0.01, \text{ denoted by } \beta_B \text{ say .} \tag{4.8.5}$$

All this *assumes* that the data can indeed be represented as two Poisson samples of sizes n = 4 and n = 8, respectively. Can the assumption be tested? If so, the tests will have to be commencement tests. So, consider a pair of co-ordination tests based, for convenience, on the ordering proposed by Fisher (1950). (Any effective ordering would do for the present purposes. Fisher's ordering just happens to be convenient.) For species A, the test distribution is given in Table 4.8.1, where $3^{[1]}2^{[1]}0^{[2]}$ describes the test datum whose mental correlate is thus found to be situated at

(0.82, 0.12, 0.06*) in the test distribution.

Table 4.8.1: Test distribution for a Poisson class characteristic

Pattern	Probability	Order	Pattern	Probability	Order
$5^{[1]}0^{[3]}$	$4/4^5$	O_5	$3^{[1]}1^{[2]}0^{[1]} \cup 2^{[1]}1^{[3]}$	$2(240)/4^5$	O_2
$4^{[1]}1^{[1]}0^{[2]}$	$60/4^5$	O_4	$2^{[2]}1^{[1]}0^{[1]}$	$360/4^5$	O_1
$3^{[1]}2^{[1]}0^{[2]}$	$120/4^5$	O_3			

For species B, the test distribution is given in Table 1.7.2, where $2^{[3]}1^{[1]}0^{[4]}$ describes the test datum whose mental correlate is thus found to be situated at

(0.82, 0.08, 0.10*) in the test distribution.

Instead of these co-ordination tests, a frequentist will perform hypothesis tests (Tallis 1988). So, let us consider the corresponding hypothesis tests of size 0.05, using the significance level as test statistic. Then we find:

For species A: SL = (0.12+0.06) = 0.18 > 0.05. The Poisson model is accepted.
For species B: SL = (0.08+0.10) = 0.18 > 0.05. The Poisson model is accepted.

It cannot reasonably be denied that these two acceptances might be subject to Type II errors. So, let us denote the error rates in respect of some or other possible alternative (any reasonable possibility will do for the present purposes) as:

β^*_A for species A, and β^*_B for species B.

At (4.8.4) and (4.8.5) we saw that for the same kind of alternative, and owing to the species A sample size being only half the species B sample size,

$\beta_A > \beta_B$, as α was fixed at the same level for the two elimination tests.

Similarly, for the same kind of alternative, and owing to the species A sample size being only half the species B sample size, we must have that

$\beta^*_A > \beta^*_B$, as α was fixed at the same level for the two commencement tests.

In fact, we must grant that:

$\beta^*_A > \beta^*_B > 0$, otherwise what purpose do the commencement tests have?

But the elimination tests originating at (4.8.1) denied this implicitly. In fact, they tacitly *assumed that*

$$\beta^*_A = \beta^*_B = 0, \tag{4.8.6}$$

so that 'error rates' for those elimination tests can purportedly be deduced as done at (4.8.3), (4.8.4) and (4.8.5) – and *that* is circular reasoning. The assumption at (4.8.6) begs the question.

The influence of the Neyman-Pearson theory makes this kind of circularity ubiquitous in present-day statistical practise. Consider 'testing "the assumption of normality" in ANOVA. What is actually tested is whether the errors *look* approximately normal, where that tacitly admits the possibility of Type II errors with non-zero 'error rates', whereas what is actually assumed is that the errors *are* normal, where that tacitly takes any Type II 'error rates' to be zero. The actual assumption cannot possibly be tested.

4.9 Stereotyping in general

Who would have us reason that: 'Nelson Mandela is an African politician, and African politicians are corrupt, hence Nelson Mandela is corrupt'? Who would have us reason that: 'Beyers Naude was an Afrikaner, and Afrikaners are racists, hence Beyers Naude was a racist'? Who would have us reason that: 'Emily Hobhouse was British, and the British were imperialists, hence Emily Hobhouse was an imperialist'? It reflects badly on our profession that so much of the statistical literature would have us reason in terms of stereotypes, and *foists* such reasoning on us as being a *necessary* doctrine. In effect, we are *brainwashed* into believing that such reasoning is 'the scientific method'.

4.10 Different usage of the term 'paradox'

Our usage of the term 'paradox' in subsequent sections refers to *a self-contradictory assertion arising from mistaken reasoning*. There is another, slightly different, usage of the same term. Figure 4.10.1 depicts the outcome of placing two identical coins next to each other and then rolling the coin on the left along half the circumference of the other coin, as depicted by the arrow. The depiction seems to be wrong, as it would seem that the coin ending up on the right, having been revolved through only half its circumference and having started from the upright position, should then end up in the upside down position. Yet (paradoxically) the depiction is correct; the reasoning is wrong. However, in the various statistical examples that follow mistaken reasoning, rather than wrong reasoning, produces results that are incorrect.

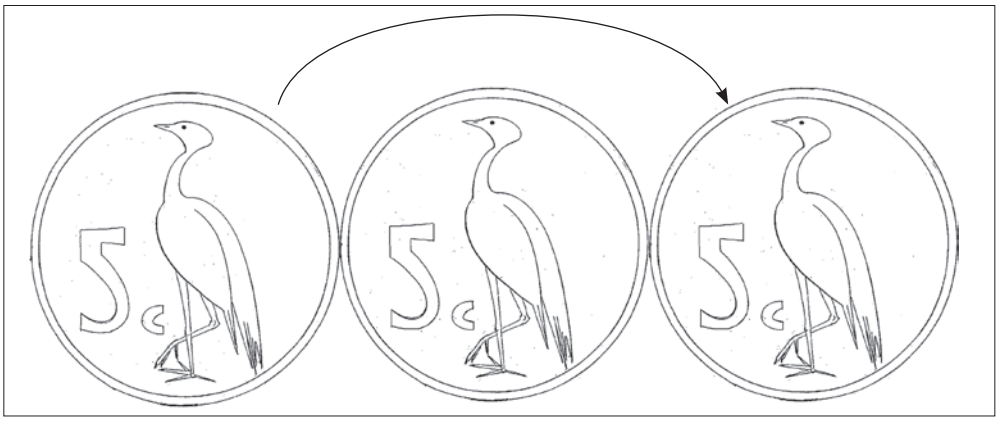


Figure 4.10.1: Example of a paradox

4.11 Pratt's paradox

The mistaken reasoning of present interest leads to a class of paradoxes that have not been widely understood at all and are consequently often misinterpreted. We use the term 'paradox' in the sense of a self-contradictory proposition that arises from mistaken reasoning. One such paradox, introduced by Pratt (1962), arises as follows: suppose that an Instrument A has been used to measure the percentages of a certain ingredient in a number of items under investigation. The investigator consults a frequentist who learns that, although none of the measurements exceeded 80%, if it had happened, Instrument A would have indicated no more than only that 80% had been exceeded. So the frequentist holds that the data must be represented in terms of censored sampling. The investigator recalls, however, that Instrument B, usually avoided as being inconvenient to use, could have been used to make any measurements in excess of 80%, had that been required. The frequentist, thus reassured, then holds that the data must be represented in terms of uncensored sampling. Subsequently, however, the investigator discovers that B is in disrepair, upon which the frequentist, thus informed, reverts to representation in terms of censored sampling. Here the paradox is that the appropriate model seems to depend on the state of repair of Instrument B that had nothing to do with making the measurements in hand. The contradiction is *apparent* only, as it falls away when the different possibilities are sorted out into just three cases, as below.

Case 1
Suppose that, for certain reasons of the kind considered in Chapters 1 and 2, a population of samples is to be brought into *the conceptual world of the human mind*, there to serve as possible explanation of how a solitary, *censored* real-world data set might have come about. In such a case the explanatory model must *predict* (must bring into the human mind) a population of many samples, some of which are censored. Otherwise, the attempt at modelling how the particular data came about would be at fault by omission of a relevant fact – one that is explanatory in respect of how *those particular data* came about.

Case 2
Suppose that, for certain reasons of the kind considered in Chapters 1 and 2, a population of samples is to be brought into *the conceptual world of the human mind*, where it is to serve as possible explanation of how a solitary, *uncensored* real-world data set might have come about. In such a case the explanatory model must *predict* (bring into the human mind) a population of many samples, none censored. Otherwise, the attempt at modelling how the particular data came about would be at fault by the inclusion of an irrelevant fact – one that is not explanatory in respect of how *those particular data* came about.

Case 3
Suppose that, for certain reasons of the kind considered in Chapter 3, repetitive use of Instrument A must bring a population of many data sets into *the real world of the human body*. In such a case the model employed must *forecast* (envisage in the future world) how some of those many data sets might well turn out to be censored. Otherwise, the model would be at fault by failure to envisage *what possibly might come about*.

Pratt's paradox is a creature of mistaken reasoning: reasoning that mistakes statistical investigation (Cases 1 and 2) for statistical technology (Case 3). Yet, in the statistical literature, such reasoning is ubiquitously promoted by asking investigators to reason in terms of 'such data as might be obtained if the investigation were done over and over again'. We must rebuff such nonsense; no data set can be done over again. When investigation has brought a population comprising many samples into *the conceptual world of human mind*, its purpose is not to forecast what 'doing it over again' might bring into *the real world*. Its purpose is to judge whether or not the data pattern in hand is similar to the patterns that typically arise under specified physical circumstances.

4.12 Berkson's paradox

Another paradox arising from mistaken reasoning of the kind that interests us here is due to Berkson (1938). Consider any two singletons that are investigated as alternative models of how given data might have come about. Taking either one of them (either one will do) as the null hypothesis, Berkson considers a specified Type I error rate (any non-zero rate will do) and notes that the Type II error rate approaches zero with increasing sample size. So he asks why we bother to test the null hypothesis for a fixed sample size, knowing that with a sufficiently large sample size the model chosen to be the null hypothesis will almost surely be rejected. Here the mistake is the idea of a specified Type I error rate. With increasing sample size, a co-ordination test will either place the mental correlate of the test datum snugly within just one of the two alternative crowds, or else will place it on the outskirts of both.

4.13 The vacuous-interval paradox

In general, frequentist inference (mistakenly) tries to make some of the properties of a collective bear upon an individual held to be one of its members whereas, as we have seen in previous sections, that courts paradoxical consequences. A ubiquitous form of the mistake tries to make confidence intervals serve the purposes of data analysis. The following three examples show how one then courts a certain paradox that may be described as *the vacuous interval paradox*. The reader is challenged to note that these examples draw on precisely the same mathematical and numerical formalities, but employ those formalities in very different ways.

Example 4.13.1

Consider an array of $N(\mu_J, 1)$ random variables and corresponding hypothesis tests of

$$H_0: \mu_J = 0 \text{ versus } H_1: \mu_J > 0, \text{ for } J = 1, 2, 3, \ldots .$$

Suppose that $0 \leq \mu_J < \infty$ for all J. Then a uniformly most powerful array of hypothesis tests of size $\alpha = 0.05$ is indicated by

$$\Pr(X_J - \mu_J > 1.645) = 0.05, \text{ leading to the decision rule:}$$
$$\text{Accept } H_1 \text{ whenever } X_J > 1.645, \text{ for } J = 1, 2, 3, \ldots . \qquad (4.13.1)$$

Recall that *a uniformly most powerful array of hypothesis tests* is such that for every false μ value the probability of rejecting it is the largest possible. The corresponding property for confidence intervals is that of being *a uniformly most accurate array of confidence intervals*, where for every false μ value the probability of accepting it is the smallest possible. As the array defined at (4.13.1) is uniformly most powerful, a corresponding array of uniformly most accurate confidence intervals is indicated by

$$\Pr(X_J - \mu_J \leq 1.645) = 0.95, \text{ leading to the decision rule:}$$
$$\text{Accept, as 'possible', any } \mu_J \text{ value} \geq X_J - 1.645, \text{ for } J = 1, 2, 3, \ldots . \qquad (4.13.2)$$

Let x denote any positive number. Inasmuch as $0 \leq \mu_J < \infty$, it then follows that:

$$-x \leq \mu_J < \infty \text{ implies that } 0 \leq \mu_J < \infty, \text{ for } J = 1, 2, 3, \ldots .$$

So, a typical 0.95 array of confidence intervals arising at (4.13.2) might be as follows:

$$X_1 = 1.66, \text{ so the corresponding confidence interval is } +0.02 \leq \mu_1 < \infty.$$
$$X_2 = 1.73, \text{ so the corresponding confidence interval is } +0.08 \leq \mu_2 < \infty.$$
$$X_3 = 1.56, \text{ so the corresponding confidence interval is } \quad 0 \leq \mu_3 < \infty.$$
$$\ldots . \qquad (4.13.3)$$

Here, precisely 0.95 of the true μ_J values will (given the assumptions made) be bounded correctly by the intervals in such an array, where that is a property of the array as a collective. It is not a property of any one term. The true μ_J value for any one term is either contained in the interval, or not contained in the interval. Here, however, that is no problem. We can envisage the use of such an array by a sampling inspector monitoring a commercial product to ensure a discernible presence ($0 < \mu$) of a certain ingredient

measurable as a N(μ, 1) random variable X. In order for the J^{th} submission to be classified according to specification, the decision rule must classify $E(X_J)$ as 'positive'. So, for the array at (4.13.3) the inspector must decide as follows:

> according to spec, according to spec, not according to spec, ... , respectively.

However, an investigator of just one, solitary real-world data set *to be explained*, is not an inspector of an array of many real-world data sets *to be decided upon*. So let us in our next example consider how the particular datum X = 1.56 given at (4.13.3), might reasonably be employed to develop an informed opinion about what possible E(X) values might describe its particular source.

Example 4.13.2

In order to make it clear that, whereas an array comprising many x values was being brought *into the real world* at (4.13.3), there is now just one single real-world x value, we denote X = 1.56 as $X = x_{obs}$ (observed x). Our reasoning now must concern the case where circumstantial details and commencement tests have brought *into the human mind* a class of models where the mental correlate of that single observed x has a corresponding value taken on by a N(μ, 1) random variable X, and the question to be addressed is: how does the quality of fit of the members of this class vary with the various possible values of μ ($0 \leq \mu < \infty$)? So, by tracing the situation of the mental correlate of x_{obs}, when that correlate is transported from model to model to model in the human mind, we find the correlate being situated

> at (•, ε, 0.06), (•, ε, 0.07), (•, ε, 0.09), ..., when μ = 0.0, 0.1, 0.2, ..., respectively,
> at (0.50, ε, 0.50), when μ = 1.56, and
> at (0.07, ε, •), (0.06, ε, •), (0.05, ε, •), ..., when μ = 3.0, 3.1, 3.2, ..., respectively.

It can hardly be thought that this is 'uninformative data analysis' because it tells us a great deal about the quality of fit of many different models in respect of the data in hand.

Example 4.13.3

A frequentist who must deal with the problem in Example 4.13.2 will mistakenly try to do so by means of the reasoning in Example 4.13.1. As in Example 4.13.2, the test datum is X = 1.56. And as in Example 4.13.1, the confidence coefficient $1\text{-}\alpha = 0.95$ is chosen without reference to the data. So, the frequentist must report that:

> $0 \leq \mu < \infty$ is a 95% confidence interval for μ.

It can hardly be thought that this is 'informative data analysis', as it tells us absolutely nothing about μ that was not already known before the data in hand were obtained.

Discussion

The value of a confidence coefficient is a property of an entire array of intervals, and not of any one particular interval. So we court a paradox when we try – as frequentist inference would have us do – to make a confidence interval address a single, solitary real-world data set. This is shown in Example 4.13.3, where we must refer to $0 \leq \mu < \infty$ as 'a 95% confidence interval for μ', whilst we are so to speak '100% confident' that its value

is contained in the interval. However, we note that, just as in the previous sections, the paradox arises from misused physical reasoning.

4.14 A CLASS OF FALLACIOUS ARGUMENTS

A paradox is often used to bring specific principles of reason into disrepute, and so persuade us to accept other principles. This practice involves a pitfall in that we may be persuaded to replace the principles under attack with fallacious principles. So a paradox is not simply to be seen as an excuse for introducing alternative reasoning. Unless the *source* of the paradox is *understood,* we are in danger of introducing principles that are inferior to those we think to rectify. This has often not been understood, so that paradoxes of the kind we met in Sections 4.11, 4.12 and 4.13, have been used to attack frequency physics and thus motivate the introduction of metaphysical ideas on statistical inference. One must not be taken in by those arguments, as they cast doubt on the reasoning of physics by *misusing* it. An example to this effect is given by Hogg and Craig (1970, pp. 207-208): let $\{X_1, X_2\}$ denote a random sample of size n = 2, drawn from a population of values uniformly distributed on the interval θ-½ to θ+½ ($\theta > 0$). Let the smaller and larger of the sample values be denoted by $X_{(1)}$ and $X_{(2)}$, respectively. Then $(X_{(1)}, X_{(2)})$ is a 50% confidence interval for θ, as the probability of sample values straddling θ is given by:

$$\Pr(X_1 < \theta, X_2 > \theta)+\Pr(X_1 > \theta, X_2 < \theta) = [(½)\times(½)]+[(½)\times(½)], \text{ i.e. } ½.$$

However, the sample range $R = X_{(2)}-X_{(1)}$ will often be > ½, in which cases we know *for sure* that the value of θ is contained in $(X_{(1)}, X_{(2)})$. In fact, as the density function of R for n ≥ 2 is given by:

$$f(r) = n(n-1)\, r^{n-2}\,(1-r) \text{ for } 0 < r < 1, \qquad (4.14.1)$$

the probability of R being larger than ½ when n = 2, is given by:

$$\int_{½}^{1} 2(1-r)dr = ¼$$

So, 25% of the time we will then refer to an interval as 'a 50% confidence interval for θ' when we are so to speak 100% sure that the interval contains θ. Here the reader will recognise the vacuous-interval paradox. As in Section 4.13, it arose because of *forecasting* instead of *pointing*. The present example is revealing in yet another way as it involves an ancillary statistic, as follows: (X_1, X_2) is minimally sufficient for θ, and the transformation to the mean and range, i.e. the transformation:

$$\overline{X} = (X_1+X_2)\div 2 \text{ and } R = X_{(2)}-X_{(1)}$$

is one to one. So, the mean and range are minimally sufficient for θ. However, the distribution of the range will obviously be independent of θ, as indeed shown at (4.14.1). R is therefore ancillary and labels alternative frames of reference. Substantive science must then select the frame of reference that is appropriate for the present purpose. In

the present chapter, that purpose is *investigative*, i.e. we should be pointing and asking 'How might *this particular individual* have come about?' Now since

$$X_{(1)} = \overline{X}\text{-}½\ R \text{ and } X_{(2)} = \overline{X}\text{+}½\ R, \text{ where } \theta\text{-}½ \le X_{(1)} \text{ and } X_{(2)} \le \theta\text{+}½,$$

it follows that

$$\overline{X}\text{-}½\ (1\text{-}R) \le \theta \le \overline{X}\text{+}½\ (1\text{-}R). \qquad (4.14.2)$$

Thus if $(\overline{X}, R)=(\overline{x}, \text{r})$ *for the particular individual* being investigated, the value of θ cannot fall outside the range:

$$\overline{x}\text{-}½\ (1\text{-}r) \le \theta \le \overline{x}\text{+}½\ (1\text{-}r) \text{ for } \textit{that particular individual}. \qquad (4.14.3)$$

For instance:

$$\text{If r = 0.2 for that individual, we know that } \overline{x}\text{-}0.4 \le \theta \le \overline{x}\text{+}0.4. \qquad (4.14.4)$$

$$\text{If r = 0.6 for that individual, we know that } \overline{x}\text{-}0.2 \le \theta \le \overline{x}\text{+}0.2. \qquad (4.14.5)$$

So, if our data corresponded to the case at (4.14.4) we must consider as possible that θ might deviate from $\bar{x}$ by as much as 0.3. But if our data corresponded to the case at (4.14.5), we would have to be stupid to consider as possible that θ might deviate from $\bar{x}$ by as much as 0.3. Clearly then, in order to investigate which possible values of θ might serve to explain how the particular data might have come about, the distribution of $\overline{X}$ conditionally on R = r is the appropriate frame of reference. That distribution is a uniform distribution whose density function is given by:

$$f(\bar{x} \mid \text{r}) = (1\text{-r})^{-1} \text{ with } \bar{x} \text{ ranging over } \theta\text{-}½\ (1\text{-}r) < \bar{x} < \theta\text{+}½\ (1\text{-}r).$$

A test for comparing any pair of θ values from the range of possible values indicated at (4.14.3) is given by the co-ordinates of the mental correlate of the datum $\bar{x}$ within this distribution. These are readily obtained by noting that the $\bar{x}$ value partitions the distribution into a pair of 'rectangles' whose areas, computed as height × width, are:

$$\text{U} = (1\text{-r})^{-1}\times\{\overline{x}\text{-}[\theta\text{-}½(1\text{-}r)]\} \text{ and } \text{V} = (1\text{-r})^{-1}\times\{[\theta\text{+}½(1\text{-}r)]\text{-}\overline{x}\}, \text{ respectively.}$$

Hence for the particular individual involved, the trace of the mental correlate of $\bar{x}$, when transported from model to model in the human mind, is given by:

$$(\text{U}, \varepsilon, \text{V}) = \left[½\text{-}\left(\frac{\theta\text{-}\overline{x}}{1\text{-}r}\right), \varepsilon, ½\text{+}\left(\frac{\theta\text{-}\overline{x}}{1\text{-}r}\right)\right],$$

$$\text{where the only possible } \theta \text{ is such that } \overline{x}\text{-}½(1\text{-}r) \le \theta \le \overline{x}\text{+}½(1\text{-}r). \qquad (4.14.6)$$

Approximating ε as zero, this trace is represented by a straight line between the points:

$$[\theta, (\text{U}, \text{V})] = [\overline{x}\text{-}½\ (1\text{-}r), (1, 0)] \text{ and } [\theta, (\text{U}, \text{V})] = [\overline{x}\text{+}½(1\text{-}r), (0, 1)],$$

as shown in Figure 4.14.1, and there is nothing anomalous about it. This resolves Hogg's and Craig's example of the vacuous-interval paradox. Like the example we developed in Section 4.12, the example of Hogg and Craig originates in a mistaken attempt at *prediction*

followed by forecasting instead of *prediction followed by pointing*. However, their example is especially instructive in more ways, as will be shown in the next four sections.

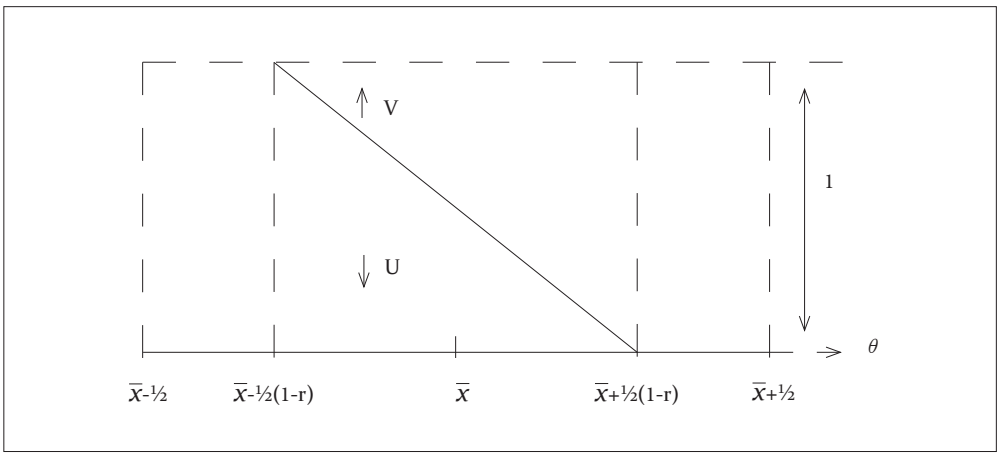


Figure 4.14.1: Values of $\theta < \bar{x}$-½(1-r) or of $\theta > \bar{x}$+½(1-r) are ruled out deterministically. For $\theta < \bar{x}$ the pointing co-ordinate is V; for $\theta > \bar{x}$ it is U. A trace is invariably data dependent, as is here the case, because the trace is a description of how the human mind envisages the test datum as coming about

4.15 When operating characteristics are irrelevant

For the problem in the previous section, the lower and upper bounds for a conceptual system of symmetric (1-α) confidence intervals for θ are given by those values of θ such that the trace at (4.14.6) reaches

$[½\ \alpha, \varepsilon, 1\text{-}½\ \alpha]$ and $[1\text{-}½\ \alpha, \varepsilon, ½\ \alpha]$, respectively.

The resulting system comprises intervals of the form

$$\bar{X}\text{-}½(1\text{-}\alpha)(1\text{-}r) \le \theta \le \bar{X}\text{+}½(1\text{-}\alpha)(1\text{-}r) \text{ for the various } r. \qquad (4.15.1)$$

The form at (4.15.1) arises within the conditional frame of reference labelled by the particular value of r. However, frequentist inference will look upon it as a recipe for repetitive decision-making under risk, whose operating characteristics are those that refer to *the real world*, where the decision-maker does not draw X_1 and X_2 such that $X_{(2)}\text{-}X_{(1)}$ takes a given value, *r*. In order to grasp the implications of this, let us, for the moment, reason in discrete terms by envisaging how an array of real-world *r* values would arise as

$r', r'', r''', \ldots$, with frequencies $f', f'', f''', \ldots$, respectively, $f'+f''+f'''+\ldots = 1$.

For each *r* value, the form at (4.15.1) involves a *conceptual* Type I error rate equal to α, thus resulting in *the real-world* Type I error rate also equal to α, as follows:

$$(f'\times\alpha)+(f''\times\alpha)+(f'''\times\alpha)+\ldots = \alpha \qquad (4.15.3)$$

And, for any given alternative, the form at (4.15.1) involves *different* conceptual Type II error rates,

> β', β'', β''', ..., corresponding to r', r'', r''', ..., respectively,

thus resulting in *the real-world* Type II error rate, as follows:

> $(f'\times\beta')+(f''\times\beta'')+(f'''\times\beta''')$ ... $=\overline{\beta}$ in the notation in Section 3.4. (4.15.4)

Welch (1939), referring to R as labelling 'samples of the same configuration', and to r as labelling 'the actual configuration observed', expresses this:

> '... in actual sampling from a population, we derive samples with all configurations', so that, when testing against any given alternative, 'the real power of the test will therefore be measured' as 'the weighted mean' of those different powers arising 'within the separate configurations, the weights being the probabilities ... of the configurations'. (4.15.5)

Thus Welch (quite correctly) measures 'the real power of the test', when simplified as at (4.15.4), as

> $[f'\times(1-\beta')]+[f''\times(1-\beta'')]+[f'''\times(1-\beta''')]+\ldots = 1-\overline{\beta}$. (4.15.6)

Welch arrives at the procedure above for interval estimation via an interpretation of a principle proposed by R.A. Fisher, whom he quotes as saying that:

> 'in interpreting our estimate ... [we] ... may take as its sampling distribution that appropriate to only those samples which have the actual configuration observed' (Fisher 1936). (4.15.7)

In Chapter 5 we will find that Welch's reasoning proceeds from a misinterpretation of Fisher's proposal, but that need not concern us here. What concerns us now is the nature of frequentist inference, of which the fundamental premise is that any findings of statistical investigation are decisions made under risk, which risk is to be accounted for by way of the *real-world* error rates when making such decisions repetitively. An inexorable consequence of that premise is that investigators are required, in one way or another, to ignore certain informative facts about the particular individual under investigation on the grounds that any individual must be viewed only in terms of the facts that apply to a host of such individuals. This has been exemplified in a variety of ways in previous sections. Further development of Welch's reasoning exemplifies this in yet another way, as follows: the equation at (4.15.3) tells us we fix the overall Type I error rate at α by fixing at α the Type I error rate within each of the subsidiary frames of reference. Obviously, however, different Type I rates α', α'', α''', ... within the same subsidiaries can also yield an overall Type I error rate equal to the specified α, the equation at (4.15.3) then being replaced by:

> $(f'\times\alpha')+(f''\times\alpha'')+(f'''\times\alpha''')+\ldots = \alpha$, for appropriate α', α'', α''', ...

So, the system of symmetric confidence intervals arising from Welch's interpretation of the form at (4.15.1) is not the only such system available. So Welch then considers its *accuracy*, i.e. its probability of excluding false values of θ, and he shows that the following system is *uniformly more accurate* in respect of such values:

$$\left[\overline{X}\text{-}½(1\text{-}R),\ \overline{X}\text{+}½(1\text{-}R)\right] \text{ whenever } R \geq \sqrt{\frac{\alpha}{2}}, \text{ and}$$

$$\left[\overline{X}\text{-}½(1\text{+}R)\text{+}\sqrt{\frac{\alpha}{2}},\ \overline{X}\text{+}½(1\text{+}R) - \sqrt{\frac{\alpha}{2}}\right] \text{ whenever } R \leq \sqrt{\frac{\alpha}{2}}\ . \qquad (4.15.7)$$

At (4.14.3) it was found that for almost any particular individual under investigation, certain values of θ are utterly impossible, yet the recipe at (4.15.7) asks us to ignore that information where that would commit us to defective epistemology, as displayed by the following example.

Example 4.15.1

Let the data be $\overline{X} = 3.6$ and $R = 0.6$. The investigator is required to specify a value of α, perhaps in advance, to ensure that it was done *without reference to the data*, and must ignore the fact that for the particular individual under investigation, any θ value not in the interval $3.4 < \theta < 3.8$ is utterly impossible. The reader will find that

> if $\alpha = 0.01$, the confidence interval arising at (4.15.7) is $3.4 < \theta < 3.8$,
> if $\alpha = 0.02$, the confidence interval arising at (4.15.7) is $3.4 < \theta < 3.8$,
> if $\alpha = 0.03$, the confidence interval arising at (4.15.7) is $3.4 < \theta < 3.8$,
> ...
> if $\alpha = 0.72$, the confidence interval arising at (4.15.7) is $3.4 < \theta < 3.8$. (4.15.8)

Compare this to the development in Example 3.7.1, which also mixes a determinate case with a statistical case. The defective epistemology displayed at (4.15.8) compels us to agree with Kempthorne and Folks (1971, p. 377) when they say that such a system is:

> 'ineffective as an ordering of tenability of values for θ for *the given set of data*' (their italics). (4.15.9)

We take the position that a data analyst must reject Welch's reasoning because in the discourse on the pursuit of knowledge, the *operating characteristics* of procedures for repetitive decision-making in respect of a host of real-world individuals are irrelevant. The data analyst must be concerned instead with the *separating characteristics* of any method for ordering the tenability of alternative models a single real-world individual has brought into the human mind as explanations of how that solitary individual might have come about. So, in the next section, we re-consider the present problem from that point of view.

4.16 The data-analytic approach to Welch's problem

A data analyst must take account of the taxonomy of a proposed class of models, and be wary of oversimplifications that conceal such taxonomy. The issues raised by Welch's problem in the context of elimination testing are simplified without loss of principle by considering samples of size n = 2. But in order to account for taxonomy involved by commencement testing, we have to consider samples of size n > 2. Then, as $(X_{(1)}, X_{(n)})$ is minimally sufficient for θ, the class characteristic is given by:

the probability of the sample, conditional on $(X_{(1)}, X_{(n)}) = (x_{(1)}, x_{(n)})$,

which amounts to taking the joint distribution of $X_{(2)}, X_{(3)}, X_{(4)}, ..., X_{(n-1)}$, to be that of a random sample of n-2 values from a population of values uniformly distributed on the interval ranging from $x_{(1)}$ to $x_{(n)}$. An attractive commencement test statistic is:

S^2 = the conditional variance of $X_{(2)}, X_{(3)}, X_{(4)}, ..., X_{(n-1)}$, (4.16.1)

as its realised value tends to separate the hypothesised model (of uniform dispersion), from incipient alternatives involving over-dispersion, and from incipient alternatives involving under-dispersion. Should the mental correlate of the datum conditional variance then be found situated snugly within the hypothesised crowd, we would be encouraged to continue with the hypothesised class of models. Note, however, (and this is crucial) that such continuation would commit us to discourse that abandons the *idée fixe*. To believe, for instance, that such continuation can be subject to 'controlled error rates' is utter nonsense, for if the continuation were at error, such an error would be in the nature of a Type II error, whose rate cannot possibly be accounted for. That is so because the 'correct alternative' would be utterly unknown. Therefore, to continue by ignoring the implication of our inability to account for any such Type II rate, would amount to assuming that the purported Type II rate is zero, and would thus amount to our falling victim to a vicious circle of the kind displayed in Section 4.8. The issue here is this:

> Scientific investigation, in general, has just one instrument for the evaluation of given data, and that is to judge the quality of fit of alternative models brought into the human mind, as alternative explanations of how those given data might have come about. (4.16.2)

So let us consider Welch's problem from that point of view. We begin by noting that his formulation envisages one of those situations where an investigator can declare: 'The possible explanations of how these data might have come about can now be limited to those listed here.' So, in that case, we must envisage how in the investigator's opinion – as tested against circumstantial details, prior experience and theoretical reasoning – the possible explanations can be limited to those where the data are represented as:

> A sample from a population described by one of the list of densities
>
> $f(x) = 1$, on $\theta - \frac{1}{2} \leq x \leq \theta + \frac{1}{2}$, indexed by θ for $0 < \theta < \infty$. (4.16.3)

As a further test of the tenability of this opinion, the investigator might well employ a co-ordination test, possibly using the test statistic defined at (4.16.1), should sample size allow it. As the investigator would have been reluctant to persist with the class of models given at (4.16.3) if the hypothesised co-ordinates of the mental correlate of the datum variance had turned out to be extreme, we must, for Welch's purpose, presuppose that such was not the case. Let us therefore envisage a commencement test with the following outcome:

> The mental correlate of the datum of conditional variance was found to be situated well within the hypothesised crowd at (0.18, ε, 0.82) in the test distribution. So the class characteristic, thus tested, fits the given data well. (4.16.4)

As explained above, it is utterly impossible to qualify this finding by way of an error rate, yet it is entirely reasonable to present it as evidence favouring the hypothesised class, as it exemplifies the universal method of scientific investigation described at (4.16.2). Since the case of $n > 2$ has now served its purpose, let us revert to Welch's simplification by considering two investigations involving similar materials, similar circumstances and similar methods of measurement, such that a commencement test in one case with $n > 2$, provides acceptable evidence for the other case where $n = 2$. Consider obtaining, now with $n = 2$, the data $\bar{X} = 3.6$ and $R = 0.6$. Many of the models listed at (4.16.3) are eliminated as being those models under which it would be impossible for the given data to have come about. The remaining models are then all those for which

$$\bar{x}\text{-}½(1\text{-}r) \le \theta \le \bar{x}\text{+}½(1\text{-}r), \text{ i.e., } 3.4 \le \theta \le 3.8.$$

The method indicated at (4.16.2) thus enables the investigator to narrow down further the list of models that might explain how the given data came about, these being those in which the data are represented as:

> A sample from a population described by one of the list of densities:
>
> $f(x) = 1$, on $\theta\text{-}½ \le x \le \theta\text{+}½$, indexed by θ for $3.4 \le \theta \le 3.8$. (4.16.5)

It might be thought that one may describe the reasoning that has advanced us from the system at (4.16.3) to the system at (4.16.5) as having a zero error rate, but that would be a bad idea, as it would introduce a concept that is not needed. This must be thoroughly understood.

> Science, in all its tasks, must always strive to maintain a minimal sufficiency of constituents. Never, ever introduce something that is not needed. (4.16.6)

All the models listed at (4.16.5) provide possible explanations of how the data in hand might have come about. Some explanations fit the data well, some awkwardly, and some poorly, as shown by the following abbreviated trace obtained from the form at (4.14.6):

> $(U, \varepsilon, V) = (\bullet, \varepsilon, 0.05), (\bullet, \varepsilon, 0.10), (\bullet, \varepsilon, 0.15)$ at $\theta = 3.42, 3.44, 3.46$, respectively,
> $(U, \varepsilon, V) = (0.50, \varepsilon, 0.50)$ at $\theta = 3.60$, and
> $(U, \varepsilon, V) = (0.15, \varepsilon, \bullet), (0.10, \varepsilon, \bullet), (0.05, \varepsilon, \bullet)$ at $\theta = 3.74, 3.76, 3.78$, respectively.

A trace is a list of facts, not forecasts; facts are not qualified by probabilities. Here all the models labelled by $\theta < 3.42$ or $\theta > 3.78$ fit the data poorly (observed fact) and all of the models labelled by θ such that $3.46 < \theta < 3.74$ fit the data well (observed fact). Moreover (and the argument that now follows is devastating) these facts cannot be invalidated, because in order to counter any attempt to do so, we need only ask:

> Has an invalid method been used at commencement? (4.16.7)

An answer in the negative entails self-contradiction, as *the same method* has produced the trace. An answer in the affirmative is not possible, because, as we have explained above, when searching for an *acceptable* class of models that method is essentially the *only one* available. Clearly then, frequentist inference cannot survive this argument. Moreover (and this is where the devastation will surface), none of the received theories of statistical inference

is able to survive this argument. We will therefore be returning to the matter from time to time. In order to provide an aid to memory, let us call it *the problem of achieving non-circular elimination*. Its essence is twofold. Firstly, at commencement we must rely on co-ordination tests, or on some or other superficially different but essentially equivalent approach, as that constitutes the only scientifically valid approach. Consequently, we are unable to disparage as invalid any result from a continuation of that approach for elimination testing. Secondly, any other approach for elimination testing is incapable of defence if it requires the commencement to provide a foundation that no commencement can provide. In the case of frequentist inference above, we have shown that it requires commencement tests capable of zero Type II error rates, which no commencement test can provide.

4.17 A REVEALING PERIPHERAL CASE

When, in Welch's problem, $R = 1$ apart from negligible rounding, we know for a fact that $\theta = \bar{x}$ apart from negligible rounding. Yet, if we use the system given at (4.15.7), having specified *without reference to the data*, $\alpha = 0.25$ say, and where it turns out that $\bar{x} = 7$ say, we would in effect have to report that:

> The value of θ is 7 apart from negligible rounding. But mind you, this result was obtained by a method that is erroneous in 25% of cases. (4.17.1)

Opposed to this, any substantive investigator who has managed to avoid statistical indoctrination would report (correctly) that:

> The value of θ is 7 apart from negligible rounding. (4.17.2)

The second sentence at (4.17.1) was determined entirely by an arbitrary value chosen *without reference to the data*, and a corresponding sentence is conspicuously absent at (4.17.2). This clearly reveals that the system at (4.15.7) springs from a doctrine at variance with the *investigative* method of the substantive sciences. This is not so because the doctrine involves mathematical or scientific error, but because it involves epistemological error. It has mistaken an investigative problem (one in the pursuit of knowledge) for a technological problem (one in the use of knowledge).

We note that this section deals with a peripheral case in the sense that $R = 1$ positions us at the interface between statistics and substantive science. Hence mutual understanding is tested, in that a statistical test should not disagree with the finding of substantive science, as expressed at (4.17.2). It is thus of considerable interest that co-ordination tests do not disagree with the finding at (4.17.2). A suite of co-ordination tests, precisely performed, would generate a trace here, where owing to rounding, the co-ordinates of the mental correlate of the test datum vary slightly, perhaps

> from (0.990, 0.010, Ø) when $\theta = 6.999$,
> through (0.495, 0.010, 0.495) when $\theta = 7.000$,
> up to (Ø, 0.010, 0.990) when $\theta = 7.001$,

and where the finding at (4.17.2) agrees with this by way of the phrase 'apart from negligible rounding'.

4.18 A fundamental rule on ancillary statistics

Owing to the developments in Sections 3.7, 4.13, 4.14, 4.15 and 4.16, it is now clear that the problems raised by ancillary statistics are not problems in forecasting. As we saw in Section 3.7, the frame of reference when forecasting is simply that which has been specified by the forecaster. However, this fact has been obscured by the ideas of frequentist inference in terms of which investigative problems are (mistakenly) interpreted as forecasting problems. When such mistakes are ruled out, the issue is clarified to a large extent. Moreover, it is easy for such mistakes to be ruled out simply by recognising that the investigation of a data set must, as an invariable scientific rule, always take the form of asking: 'How might *this particular* data set have come about?' In a later chapter on ancillary statistics we will return to this point.

4.19 Neyman's theory of statistical inference

Frequentists tend to be unaware of the stereotypic nature of their reasoning. This is because they tend not to perceive their stereotypic arrays in substantive terms, but in terms of the statistics of the prescribed behaviour instead. Thus for instance, a typical text on hypothesis tests would point out that Type I and Type II error rates do not apply to just one particular instance of a hypothesis test. In one such test, the text will explain, the decision made is either at error ('rate' = 1), or not at error ('rate' = 0). The typical text explains that these rates apply to the method used, which means that the properties of the method used become the properties of the user's behaviour. That behaviour is then held forth as committing Type I errors in trifling proportions of cases, whilst minimising the rates at which any Type II errors are committed. Thus Neyman (1952, p. 209), in explaining his theory of confidence intervals, recommends that:

> ' ... as an act of will, not reasoning', we are ' *... to behave as if it were known for certain that the true value of* θ *lies between the lower and the upper confidence limits computed from actual observations*' (original italics). (4.19.1)

He argues that:

> 'The motivation behind this rule of behavior is simple: taking into account the operational interpretation of confidence intervals, we know that the long-run relative frequency of cases where our actions will be adjusted correctly, is equal to the (confidence coefficient) which we have selected ourselves.' (4.19.2)

Neyman is well aware that the long-run relative frequency in question is 'deduced from specified assumptions', but is evidently unaware that such reasoning courts the vicious circle dealt with in Section 4.8. Nor has he been alone in this. On the contrary, such reasoning has turned out to be exceedingly persuasive, thus distracting attention from the scientific method of data analysis and beguiling statisticians instead with the idea that statistical inference equals 'behaviour subject to low error rates'. This has not only entrenched the shortcomings discussed in previous sections, but has also lead to the further idea that error rates *accumulate* owing to 'simultaneous statistical inference'. However, this need not concern us at present. For now we wish to make it clear that

the theory of hypothesis tests and its dual theory of confidence intervals are directed at the attainment of real-world error rates. For instance, Neyman (1952, pp. 210 onward), describes a sampling experiment that may be conducted to illustrate the operational interpretation of confidence intervals for an unknown parameter θ, and points out that in a hundred repetitions of the experiment:

> ' ... you will find it instructive to select for your sampling experiment a set of, say, 100 different values (quite arbitrary) of θ'. (4.19.3)

Subsequently he remarks that one will of course realise quickly that the resulting data are independent of those '100 different values', and so one may as well have used the single most convenient value of θ. But that is beside the point here. The point here is simply this: the statements at (4.19.1), (4.19.2) and (4.19.3) clearly show Neyman as committed to the idea of trying to achieve real-world error rates.

4.20 The incoherence dilemma of frequentist inference

We return to the *idée fixe*. It falls under the general notion of *probability inference*, which is conveyed by Definition 4.20.1.

Definition 4.20.1:

The term *probability inference* refers to the idea of developing a system of inductive logic whose fundamental concept, corresponding to that of deductive implication, would be that of probabilistic implication. It refers, in other words, to the idea of developing a concept that can represent the degree of rational credibility ('probability' in some sense) that a given body of evidence confers on a particular scientific hypothesis.

If there is to be any hope of developing a satisfactory theory of probability inference it would have to involve some or other form of statistical inference as a special case. The concept of a 'degree of rational credibility' as in Definition 4.20.1 would, in the statistical case, then have to be supplied by a statistical error rate, or a confidence coefficient, or a fiduciary probability of some sort. In this section we show that in trying to develop such a concept, frequentist inference falls into incoherence that constitutes one horn of a dilemma; the other horn being the vicious circle displayed in Section 4.8. It will in fact become obvious that *any* attempt to develop such a concept in the statistical case is bound to land on the horns of that dilemma. Theorem 4.20.1 states this fact in terms more general than frequentist terms.

Theorem 4.20.1 (The incoherence dilemma of statistical inference).

It is not possible to unify commencement tests and elimination tests by coherent probability calculus. Any attempt to overcome this difficulty either fails to achieve coherence (the first horn of the dilemma), or else falls into circular reasoning (the second horn of the dilemma).

No satisfactory treatment of the notion of 'statistical inference' could possibly evade this dilemma. So, though Section 4.8 has in effect already provided its proof, the issue is

of such overriding importance that we must, at the risk of belabouring the matter, make sure that the limitation it imposes upon us is firmly grasped. So we now give two further examples to show why coherent probability calculus in the sense referred to in Theorem 4.20.1 is generally unattainable by frequentist inference. In each example, the data in hand are first analysed using co-ordination tests, after which the incoherence arising from the introduction of frequentist ideas is shown. In each example the main thrust of the argument calls for a probability statement of the sort envisaged by Definition 4.20.1, a statement regarding the possible values of a particular parameter, denoted by θ.

Example 4.20.1

This example involves three different data sets and a problem to be solved, as follows: despite the presence of hundreds of sharks in the coastal waters around Cape Town, many between three and six metres in length, they rarely attack humans, their staple food being the Cape fur seal. During the 15 years 1990 to 2004 inclusive, the numbers of fatal shark attacks on humans in Cape waters were as follows (in sequential order):

1, 0, 0, 0, 1, 0, 0, 2, 1, 1, 0, 0, 0, 1, 2. (Data Set 1) (4.20.1)

Thus the total numbers of fatal attacks over the five consecutive three-year periods were as follows (in sequential order):

1, 1, 3, 1, 3. (Data Set 2) (4.20.2)

The numbers of *all* shark attacks over the 15 years were as follows (in sequential order):

4, 2, 1, 4, 9, 3, 3, 5, 15, 8, 4, 3, 3, 4, 3. (Data Set 3) (4.20.3)

The problem to be solved. We must try as best we can (this is to be the main thrust of the development) to form an opinion about the possible values of

θ = Pr(more than 2 fatal attacks within the same calendar year),

where that requires of circumstantial reasoning and commencement tests to bring into the human mind the tenable members of a class of models indexed by θ. (4.20.4)

A solution using co-ordination tests: To commence with, we ask, for each of the three data sets, whether or not the numbers of attacks might be modelled satisfactorily as a Poisson sample. Using the dispersion index test as explained in Examples 1.11.1 and 1.11.2, we find:

For Set 1: The datum chi-square, 12.7 on 14 df, is found to be situated at approximately (0.4, 0.1, 0.5*) in the test distribution. So the Poisson class characteristic, as tested, fits the given data well.

For Set 2: The datum chi-square, 2.67 on 4 df, is found to be situated at approximately (0.4, 0.2, 0.4*) in the test distribution. So the Poisson class characteristic, as tested, fits the given data well.

For Set 3: The datum chi-square, 36.5 on 14 df, is found to be situated at approximately (0.9990, 0.0001, 0.0009*) in the test distribution. So, in this case, the Poisson class characteristic, as tested, fits the given data very poorly.

Set 3 is bound to involve sources of heterogeneity that would be partly removed by restricting the analysis to fatal attacks only because such attacks involve sharks of less variable size and species. Also, the term 'non-fatal attack' is not as well defined as the term 'fatal attack'. Furthermore, a fatal attack is sure to be reported whereas some non-fatal attacks would go unreported. However, one cannot assume that Set 1 is completely homogeneous; the alternative possibility of some undetected heterogeneity must be conceded. Nevertheless, the numbers of fatal attacks is approximated satisfactorily by the Poisson class characteristic, thus providing for a tenable solution to the problem stated at (4.20.4). The same solution then holds for both Set 1 and Set 2 since, for each of these, we must proceed from an analysis of the form given at (1.8.1), where the first factor and the datum it addresses are precisely the same for sets 1 and 2, as follows: for sampling from any member of our class models, the sample total X is distributed as a Poisson random variable and is minimally sufficient for the index, which may be taken as $E(X \div 15) = \lambda$ say. Making X address the corresponding real-world datum, X = 9 fatal attacks, we invariably obtain over all levels of co-ordination a suite of uniformly most separating co-ordination tests for comparing whatever pairs of alternative values of λ might be of interest, $0 < \lambda < \infty$. In order to address the problem formulated at (4.20.4) we then note that:

$\theta = 1\text{-}e^{-\lambda}(1+\lambda+\frac{1}{2}\lambda^2)$ is a one-to-one transformation.

It follows from Theorem 2.5.1 that the trace of the mental correlate of our test datum, X = 9 fatal attacks, expressed in terms of the probability of interest, θ, is given by:

(U, ε, V) = (0.025, ε, •), (0.050, ε, •), (0.100, ε, •), when θ = 0.004, 0.006, 0.008,
(U, ε, V) = (0.5, ε, 0.5), when θ = 0.028, and
(U, ε, V) = (•, ε, 0.100), (•, ε, 0.050), (•, ε, 0.025), when θ = 0.071, 0.089, 0.108,

respectively. In the absence of any contrary evidence, a reasonable opinion might thus hold that the probability of interest is not much less than 1%, and not much more than 7%. We note, however, that this is not at all prescriptive; facts derived by statistical data analysis must leave room for any other facts that might bear upon the matter.

Introducing frequentist ideas: In order to give a standard two-sided confidence interval for the probability of interest, we must specify, without reference to the data, the value of the confidence coefficient as say $1\text{-}\alpha = 0.95$. And (as can be derived from the co-ordination tests given above) this specification would have us assert:

> The interval bounded by θ = 0.004 and θ = 0.108 has been obtained by a method that would cover the true value of θ in 95% of cases. (4.20.5)

But if we try to defend this assertion we land on the horns of a dilemma: for if (and this leads to the first horn) we simply assume that X may be taken to be a Poisson random variable, we land in the vicious circle displayed in Section 4.8. If (and this leads to the second horn) we try to defend the assertion by citing commencement tests, we fall into incoherence, as follows: each of the two dispersion index co-ordination tests enables us to perform a corresponding hypothesis test, and we must, in each case and without reference to the data, specify a Type I error rate. Let us specify say $\alpha = 0.05$ in each case. Then, expressing each test in terms of the significance level as test statistic, we find that:

For Set 1: SL = 0.1+0.5 = 0.6 > α; so, the Poisson model is accepted.
For Set 2: SL = 0.2+0.4 = 0.6 > α; so, the Poisson model is accepted. (4.20.6)

The theory of hypothesis tests would in each case therefore have us assert that:

The decision has been reached by a method that would in 5% of cases lead to false rejection of the null hypothesis. (4.20.7)

But the assertion at (4.20.7) cannot provide for a coherent probability calculation that would culminate in the assertion at (4.20.5). That is because in the case of Set 2 the dispersion index test involves just four of the fourteen degrees of freedom that are involved by the corresponding test in the case of Set 1. So for the two different tests performed at (4.20.6) the probabilities of having incurred corresponding Type II errors cannot be equal. Coherent probability calculations would thus either have to lead to an interval estimate of θ for Set 2 that is wider than that for Set 1, or else would have to lead to a coverage probability for Set 2 that is less than that for Set 1.

Remark: One cannot simply accept Poisson sampling as obviously the correct model in the present case. A statistician who adopts a class of models for the representation of given data without as much as a cursory examination of those data does not amount to much. One who does do even the most cursory examination has thereby performed an informal commencement test. Literature that fails to come to grips with this is fundamentally defective – unfortunately that is the case in almost all currant literature on statistical inference.

Example 4.20.2

Again we consider given data and a problem to be solved, in this case as follows:

The data: Bliss (1967, p. 212) considers the healing times of skin wounds on the backs of 40 rats – 20 treated with medication, and 20 unmatched controls. He uses Fisher's g_1 and g_2 to test for skewness and kurtosis, and Snedecor's F to test for heterogeneity of variance. That motivates an analysis of healing rate, the reciprocal of healing time, and the use of Student's t for interval estimation of an additive treatment effect.

The problem to be solved: We must try as best we can (this is to be the main thrust of the development) to form an opinion about the possible values of:

θ = the treatment effect, as expressed on a suitable scale,

where that requires of circumstantial reasoning and commencement tests to bring into the human mind the tenable members of a class of models indexed by θ (4.20.8)

A solution using co-ordination tests: The statistical co-ordinates that direct us to the situations of the hypothesised mental correlates of the observed g_1, g_2 and F for this example are given in Table 4.20.1.

Table 4.20.1: Hypothesised co-ordinates for an analysis of data on skin wounds on the backs of 20 rats treated with medication, and of 20 untreated control rats. Response was measured as healing time, or healing rate, the reciprocal of time.

	Healing time		Healing rate	
	Treated rat	Control rats	Treated rats	Control rats
Skewness	(•, ε, 0.162)	(•, ε, 0.008)	(0.336, ε, •)	(0.114, ε, •)
Kurtosis	(0.321, ε, •)	(•, ε, 0.058)	(0.309, ε, •)	(0.180, ε, •)
Variance ratio		(•, ε, 0.005)		(•, ε, 0.376)

The transformation from times to rates, reduced skewness, kurtosis and heterogeneity of variance, all of which were evidently due to lingering wounds, resulting in skewness to the right, platykurtosis and inflated error variance, mainly amongst the controls. Denoting the additive treatment effect on healing rate as θ, its estimate as $\bar{d}$ and the estimated standard error of $\bar{d}$ as $s_{\bar{d}}$,

$$(\bar{d}\text{-}\theta)\div s_{\bar{d}} = \text{Student's central } t \text{ on } (20\text{-}1)+(20\text{-}1) = 38 \text{ degrees of freedom,}$$

where $\bar{d} = 1.2185$ and $s_{\bar{d}} = 0.1446$ for the given data set. (4.20.9)

This elimination pivot generates the following abbreviated trace for various situations of the mental correlate of the corresponding test datum, when variation in the value of θ transports the correlate from model to model in the human mind:

(0.025, ε, •) and (•, ε, 0.025) are attained at θ = 0.93 and 1.51, respectively,
(0.050, ε, •) and (•, ε, 0.050) are attained at θ = 0.97 and 1.46, respectively, and
(0.100, ε, •) and (•, ε, 0.100) are attained at θ = 1.02 and 1.41, respectively.

So, as tested, the members of our class of models that fit the given data satisfactorily are those for which θ is not much more than 0.05 less than 1.02, and not much more than 0.05 more than 1.41. This is a fact; not simply an opinion. The investigator might also be of that opinion with regard to the substantive interpretation of θ, but we must distinguish the facts that help inform such an opinion from that opinion itself. If we do not, we risk falling victim to the idea that statistical facts require a knowing subject who must specify an error rate without reference to the data, and as an act of will refuse to admit that the case in hand might be statistically unusual. We will return to this point.

Introducing frequentist ideas: To test for normality, Bliss uses the recipe given at (1.31.2), thus performing 'simultaneous statistical inference' with regard to the four alternative possibilities, skewness to the left, skewness to the right, leptokurtosis and mesokurtosis. In respect of healing times this test leads to:

chi-square on 2 df = 1.19 for the treated rats but 8.31 for the controls,
where the latter exceeds the 0.025 critical value, 7.38. (4.20.10)

As a further test for normality of healing times Bliss uses Snedecor's *F* to compare the variance of the healing times of treated rats to that of controls. This test leads to:

F on 19 and 19 df = 3.45,
where that exceeds the 0.025 critical value, 2.53. (4.20.11)

Such tests lead Bliss to reject the hypothesised normality of healing times, and instead to consider healing rates. Instead of the results at (4.20.10) one then finds:

chi-square on 2 df = 1.16 for the treated rats and 1.45 for the controls,
where neither of these exceeds the 0.025 critical value, 5.02. (4.20.12)

And, instead of the result at (4.20.11) one then finds:

F on 19 and 19 df = 1.17,
where that does not exceed the 0.025 critical value, 2.53. (4.20.13)

Such tests prompt Bliss to accept the hypothesised normality of healing rates. Thus re-interpretation of the test at (4.20.13) as one for homogeneity of variance given normality, is taken to justify the use of Student's t for interval estimation of an additive treatment effect on healing rate. Hence Bliss finds that:

$0.9257 < \theta < 1.5113$ is a 0.95 confidence interval for θ. (4.20.14)

The procedures that have culminated in the result at (4.20.14) are clearly inspired by the idea of 'probability inference' as expressed by Definition 4.19.1. At (4.20.10) for instance, the Type I error rates incurred in respect of healing time, say α_1 and α_2, are functions of constituent rates, say α_{11j}, α_{12j}, α_{13j} and α_{14j}, for j = 1, 2 respectively, in respect of which we can show by coherent probability calculus that for the recipe used by Bliss:

$$\alpha_{11j}, \alpha_{12j}, \alpha_{13j}, \alpha_{14j} \approx \frac{-1+\sqrt{1+\alpha_{1j}}}{6}, \text{ for } j = 1,2. \quad (4.20.15)$$

Bliss specifies $\alpha_j = 0.05$, and thus by implication specifies:

$$\alpha_{11j}, \alpha_{12j}, \alpha_{13j}, \alpha_{14j} \approx 0.0125, \text{ for } j = 1,2. \quad (4.20.16)$$

A moot point here is whether or not simultaneous statistical inference would require, instead of the two separate tests as at (4.20.10), just a single test based on the sum of the two chi-square values. The result would be chi-square = 9.40 on four degrees of freedom, which fails to exceed its 0.05 critical value, 9.49, and would thus force Bliss to adopt, as an act of will, a model that with good reason he considers inappropriate. However, whatever view one might take, it is nevertheless impossible to avoid incoherence. This becomes apparent immediately when one notes that the Type II error rates incurred at (6.19.12) in respect of healing rate, say β_1 and β_2, are functions of constituent rates, say β_{11j}, β_{12j}, β_{13j} and β_{14j}, for j = 1, 2 respectively, but where these rates are utterly unknown to the extent of not even being at all capable of some or other mathematical expressions corresponding to those at (4.20.15) and (4.20.16). Again, at (4.20.11) the Type I error rate incurred could be specified, but consider as follows the Type II rate thus incurred at (4.20.13).

Denote the variances of the data on healing rates by

s_1^2 for the treated rats, and s_2^2 for the control rats,

the corresponding random variables that come into the human mind by

$$S\,\frac{2}{1} \text{ and } S\,\frac{2}{2}, \text{ respectively, with } \mathrm{E}(S\,\frac{2}{1}) = \sigma\,\frac{2}{1}, \text{ and } \mathrm{E}(S\,\frac{2}{2}) = \sigma\,\frac{2}{2},$$

and the 0.0125 critical value of Snedecor's F on 19 and 19 df by $F_{19,19}$ (0.0125). Then the Type II error rate incurred at (4.20.13) is capable of mathematical expression as

$$1\text{-}\Pr[F_{19,19}(0.0125) \geq (S_1^2 \div S_2^2) \div (\sigma_1^2 \div \sigma_2^2)] \geq F_{19,19}^{-1}\ (0.0125)],$$

where $\sigma_1^2 \div \sigma_2^2$, however, is an unknown value. (4.20.17)

Clearly then frequentist inference involves non-trivial probabilities by way of Type II error rates that cannot, in the sense of Definition 4.19.1, be brought into account by coherent probability calculus, i.e. by calculus able to provide 'the probability', in the frequentist sense, that given data confer on a scientific hypothesis about how those data might have come about.

The issue is this: Neyman's theory of statistical inference tells us (correctly so) that the interval estimate at (4.20.14) was obtained by a method that, under certain hypothetical circumstances, would bracket the true value of interest in 95% of cases, but cannot confer that 95% probability onto the actual circumstances, and so cannot provide a 'probability' in the sense required by Definition 4.20.1. Its attempt to do so is incoherently eclectic. That its reasoning is eclectic is exemplified by the moot point raised in connection with the tests at (4.20.10). That its reasoning is incoherent is very simple to prove, as follows:

> Specify 0.050 instead of 0.025 as the Type I error rate at (4.20.10) and (4.20.12).
> Specify 0.050 instead of 0.025 as the Type I error rate at (4.20.11) and (4.20.13).

The reasoning would then lead to precisely the same confidence interval at (4.20.14). Bearing in mind that by altering a Type I error rate we alter any corresponding Type II error rate, it follows beyond any reasonable contest that the confidence coefficient at (4.20.14) does not arise from a coherent evaluation of the given body of evidence.

We have developed Theorem 4.20.1 in the context of frequentist inference. However, it is obvious that Definition 4.20.1 cannot possibly be satisfied by statistical inference of whatever kind for the simple reason that an elimination test is always preceded by commencement testing of some or other kind so as to provide for the class of models then to be employed by the elimination test. That inexorably involves the possibility of Type II errors, the 'probabilities' of which cannot be provided.

4.21 The scientific status of proof by dilemma

Present-day statistical practice rests on deeply entrenched habits of thought in support of the notion that statistics requires its own special form of 'inference'. These habits of thought are not aware of the extent to which they sustain that notion by incoherent and circular reasoning, and may therefore be expected to cast doubt on the scientific validity

of our proofs by dilemma. We must not allow that. So, let us note that a proof by dilemma is unexceptionable in science. This is exemplified by the scientific view on astrology. Astrology purports to forecast the course of every individual human life on the basis of the positions, at the moment of inception of that life, of earth's planets, and of the appropriate position of 12 astrological constellations named Aries, Taurus, Gemini, etc., comprising the 12 parts of the zodiac. But the question: 'Which of the 12 constellations is appropriate to a given human life?' lands astrology on the horns of a dilemma. There are two different schools of thought, respectively called 'tropical' and 'sidereal'. The tropical school adheres to the concept of the zodiac as formulated by ancient astrologers thousands of years ago referring to charts that no longer reflect the positions of the planets or stars accurately. This is so because of what is known as 'the precession' of the earth, as a result of which the constellations have since shifted such that the zodiac has undergone a displacement equal to one whole constellation. For instance, if you are a Taurus according to the tropical school, you are actually an Aries according to the sidereal school, since the latter school takes precession into account. This brings us to the dilemma: if you are a Taurus according to the tropical school (this leads to the first horn of the dilemma), your horoscope is not governed by the positions of the celestial bodies, but by an ancient authority. And if you are a Taurus according to the sidereal school (this leads to the second horn of the dilemma), your horoscope is governed by a classification principle that for thousands of years failed to observe that it was producing fallacious horoscopes. Thus we have a proof by dilemma that astrology is not a science. That is the case because science, on the one hand, does not employ the method of authority. And, on the other hand, science must look askance at reasoning that, albeit in terms of correcting for the precession of the earth, persists in an ancient theory of forecasting that refuses to employ the scientific method of systematically examining whether or not its forecasts are realised. It must be grasped firmly that the method of proof that here holds for astrology also holds for statistics, as science cannot tolerate any reasoning that is incoherent, or that proceeds from self-serving assumptions incapable of empirical defence. There can be no merit in habits of statistical thought that obstinately refuse to abandon ideas and convictions that our proofs by dilemma have shown to amount to a meme, that is to say, to a self-replicating element of culture, passed on by imitation (Dawkins 2003).

4.22 The randomised test paradox

Amongst the paradoxes of present interest, the randomised test paradox is perhaps the most revealing. For a paradigmatic example we suppose that an investigator wishes to establish whether or not a certain taster can discriminate by taste between two wines. So in each of seven separate replicates the taster is served three specimens of wine, one from a randomly chosen one of the wines, and two from the other wine; one of the latter two is labelled 'duplicate'. In each replicate the taster must try by taste to identify, or failing that to guess, which specimen is odd. Let the number of successful identifications be modelled as a binomial random variable X. Let θ denote the probability of a success ($0.5 \leq \theta < 1$). Let the given data be such that $X = 6$. The investigative question is:

How might these particular data have come about? (4.22.1)

In the following we consider first how a co-ordination tester would have us deal with the question, and then how a frequentist would have us do so, given agreement that all of the possible ways in which X = x might have come about, are exhausted by the binomial models

$$[x, \Pr(X = x|\theta)] = \left[x, \left(\frac{7!}{x!(7-x)!} \right) \theta^{x}(1-\theta)^{7-x} \right] \text{ for } x = 0, 1, 2, ..., 7. \qquad (4.22.2)$$

Our problem is to judge which of these models then provide tenable explanations of how the given datum, X = 6, might have come about.

Example 4.22.1: A co-ordination tester's approach

X is minimally sufficient for θ, and ordering on the value of X we obtain a suite of co-ordination tests that are most separating, uniformly so over all possible pairs of index values, and invariably so at every attainable level of co-ordination. Consider, for instance, the pair of models indexed by θ = 1÷2 and θ = 3÷4. Expressing the model indexed by θ = 1÷2 as:

$$2^7\Pr(X = x \mid \theta = 1\div2) = 1, 7, 21, 35, 35, 21, 7, 1,$$
$$\text{for } x = 0, 1, 2, 3, 4, 5, 6, 7, \text{ respectively,} \qquad (4.22.3)$$

the mental correlate of the given datum, X = 6, is found rather awkwardly far down at:

$$(U, \varepsilon, V) = (0.94, 0.05, 0.01) \text{ in the right-hand tail of that model.} \qquad (4.22.4)$$

Expressing the alternative model indexed by θ = 3 ÷ 4 as:

$$4^7\Pr(X = x|\theta = 3\div4) = 1\times3^0, 7\times3^1, 21\times3^2, 35\times3^3, 35\times3^4, 21\times3^5, 7\times3^6, 1\times3^7,$$
$$\text{for } x = 0, 1, 2, 3, 4, 5, 6, 7, \text{ respectively,} \qquad (4.22.5)$$

the mental correlate of the given datum, X = 6, is found snugly within the crowd at:

$$(U, \varepsilon, V) = (0.56, 0.31, 0.13) \text{ in that model.} \qquad (4.22.6)$$

Such comparisons of all possible pairs of index values are made available by the trace of the mental correlate of the given datum, which trace is given by:

$$[U(\theta), \varepsilon(\theta), V(\theta)] = [1-7\theta^6(1-\theta)-\theta^7, 7\theta^6(1-\theta), \theta^7], \text{ for } 1\div2 \leq \theta < 1. \qquad (4.22.7)$$

A co-ordination tester would hold that for X = 6, this trace comprises the whole of the available evidence that one might bring to bear on the question at (4.22.1) and would thus hold that in respect of that question, there is nothing more to consider. In order to understand this, we note that there are infinitely many one to one transforms of the minimal sufficient statistic, examples of such transforms being ln(X+1), CX for C any non-zero constant, 1÷(X+1), ... The distributions of different transforms have different shapes; so those differences in shape cannot possibly be part of the evidence to be considered. The whole of the evidence to be considered is therefore all that and just that which, in respect of the given datum, is invariant under one to one transformation, and where that invariant, and just that invariant, is conveyed by the statistical co-ordinates of the mental

correlate of the given datum. The point here is simply that X is merely a label; so, its co-ordinates convey the whole of the evidence in a form that measures fit directly.

Example 4.22.2: A hypothesis tester's approach

The most direct frequentist approach to the problem would be to limit interest to just two alternative hypotheses that are substantively expressed as

H_0: no discriminative ability, versus H_1: some discriminative ability,

and mathematically expressed as

H_0: $\theta = 0$ versus H_1: $\theta > 0$.

A frequentist interprets the problem to be one in decision-making under risk, the risk being that of erroneously rejecting one or the other of the two alternative hypotheses. The problem is then taken to be that of controlling those risks, insofar as possible, by specifying small probabilities of erroneous decision, and of otherwise minimising the probabilities of error. The choice of the alternatives H_0: $\theta = 0$, H_1: $\theta > 0$ has to some extent been governed by such a view of the problem, because of the Neyman-Pearson lemma for repetitive decision-making under risk – the lemma tells us how the probability of erroneously rejecting H_1 can be uniformly minimised for any specified probability (α for $0 < \alpha < 1$) of erroneously rejecting H_0. Specification of α must then be without any reference to the data, as say $\alpha = 0.05$ (a widely favoured value). It then turns out that for the data in hand

the critical region is $X = 7$, where $\Pr(X = 7 \mid H_0) = 1 \div 2^7$,
the boundary is $X = 6$, where $\Pr(X = 6 \mid H_0) = 7 \div 2^7$, and
the acceptance region is $X < 6$, where $\Pr(X < 6 \mid H_0) .= 120 \div 2^7$. (4.22.8)

As $0.05 = (1 \div 2^7) + (27 \div 35)(7 \div 2^7)$, the Neyman-Pearson lemma for repetitive decision-making under risk thus tells us that, in order to obtain a uniformly most powerful test of size 0.05, we must whenever $X = 6$, draw Y, a haphazard one of the numbers 1, 2, 3, ..., 35, and then employ the following decision rule:

If $Y > 27$, accept H_0: $\theta = 0$, thereby rejecting H_1: $\theta > 0$.
If $Y \leq 27$, reject H_0: $\theta = 0$, thereby accepting H_1: $\theta > 0$. (4.22.9)

Here is a paradox, as follows: we have been asked to develop scientific answers to the question: 'How might these data have come about?' We respond by drawing a chip from a bowl containing eight chips labelled 'H_0 is the correct explanation', and 27 chips labelled 'H_1 is the correct explanation'. Then, following Neyman as quoted at (4.18.1), we, in an 'act of will', must 'behave as if it were known for certain' that the explanation offered by the label on that chip is the true explanation. This is a paradox precisely as defined in the second sentence of Section 4.10: firstly because we have 'a self-contradictory proposition' as there defined, in that Y cannot possibly provide any information toward answering the question at (4.22.1), and secondly because, and again as defined in Section 4.10, the proposition arises from mistaken reasoning. In fact, it arises, as do all those paradoxes we dealt with previously in this chapter, from reasoning that has mistaken a problem in the investigation of one particular case, for a problem in repetitive decision-making in a host of such cases.

Discussion

An overwhelming majority of frequentists are somehow with limited understanding agreed upon the objectionable nature of a randomised hypothesis test when used for the analysis of particular data and so would agree with Agresti (2002, p. 27) to persecute the notion of such use. They would also agree with him when, by way of giving his reason for that, he vehemently (and wrongly) declares of the auxiliary random number that:

' ... it is absurd to let this random number influence a decision.' (4.22.10)

There is nothing 'absurd' about a scientific technology achieving its stated objectives, and, as we have seen in Chapter 3, randomised decision rules are unexceptionable. On the contrary, it would be truly absurd for any informed individual to refuse using such technology in the case of R.A. Fisher's example of repetitive decision-making under risk as described in Section 3.10. However, as Agresti's book is obviously directed at data analysis, he is quite correct in refusing to countenance the use of randomised decision rules for such analyses. Where he goes wrong, along with other frequentists, is when he thinks that all that needs to be done about it, is to avoid using that particular kind of decision rule for the analysis of given data, thus failing to realise that any decision rule is inappropriate for such analysis.

4.23 An irrelevant disagreement

We note in passing that much of the literature on statistical data analysis would have the investigator specify the Type I error rate. Other such literature would instead have the investigator report the corresponding realised significance level, so that the reader of the report may specify his/her preferred Type I error rate. The issue is irrelevant, as in data analysis no such specification should be made at all. A rate can be realised in a host of cases; it cannot be realised in just one particular case.

4.24 A liaison between 'knowledge' and its 'knower'

At (4.22.J, J = 2, 3, 4, ... 7) probability concepts give mathematical expression to a variety of predictive models, and to the predication of given data by those models. All the predications involved are facts of the form 'such data could come about by way of ..., and so these data might have come about thus', — are facts that can by simulation be forced upon the human body, — are facts beyond reasonable contest. Instead of that, the idée fixe proposes to introduce further probability concepts to give expression to the idea that statistical inference produces not facts, but uncertain knowledge. And, as that uncertainty cannot possibly spring from the substantive subject matter under investigation, it must spring from the knower of the purported knowledge. Various received theories of statistical inference thus involve a knower whose uncertainty is then, in some or other sense, expressed by 'probability'. In the case of frequentist inference the knower must, by acts of will, treat particular decisions made under risk as knowledge of uncertain kind. That uncertainty is then to be expressed by way of the error rates of a purported host of such decisions.

It will be found that each of various received theories of statistical inference has its own device for expressing 'the uncertainty of the knower of the uncertain knowledge', and the result is invariably a liaison of the purported 'knowledge' and its 'knower'. In the case of frequentist inference, the liaison is revealed with special clarity by the randomised test paradox, as any such test is a liaison of three constituents, as follows in the case of Example 4.22.2.

> Constituent 1: The display at (4.22.8) comprises facts brought forward from the trace displayed at (4.22.7). The knower is not involved; only the data are.
>
> Constituent 2: The Type I error rate $\alpha = 0.05$ is specified without reference to the data in hand. The knower, being the specifier, is involved.
>
> Constituent 3: The random number Y is drawn to achieve the specified error rate: $0.05 = (1\div 2^7)+(27\div 35)(7\div 2^7)$. The knower, being the drawer of Y, is involved.

We have remarked that an overwhelming majority of frequentists refuse to use a randomised hypothesis test for the analysis of given data. Yet, such tests enable one to achieve the stated objective of frequentist inference, which is to minimise Type II error rates whilst achieving specified Type I error rates, where it would then be silly to rule out any attainable error rate considered desirable. So we have to ask: 'Why then such refusal?' The only discernible answer is:

> Y is adjoined to the given data after those data have already come about, and thus cannot possibly partake of how those data might have come about. (4.24.1)

This reasoning applies not only to Y, but also to α. So it compels us to recognise that Constituents 2 and 3 must both be expelled from data analysis. The outcome is then to remove all those constituents arising from the knower only, and thus to dissolve the liaison between the knowledge and its knower. This takes us back to Constituent 1 (the trace at (4.22.7)) as conveying all that we can learn by making the models agreed to at (4.22.2) address the given data. This underscores a fundamental principle of scientific reasoning, a principle we met at (4.16.6):

> Never ever introduce a constituent that is not needed. (4.24.2)

As indicated already, various theories of statistical inference produce their own forms of liaison between the knowledge and its knower. We meet the liaisons in subsequent chapters, and find that each one springs from some or other defective epistemology. So we must call for their dissolution. It is worth noting that Popper (1979) arrives at a similar view, but on wider grounds. He calls for 'an epistemology without a knowing subject', which he holds, is required to attain 'objective knowledge'. In a statistical context we must, however, be leery of the terms objective and subjective, because these terms have come to be associated with a silly dispute between frequentists and Bayesians. So, let us rather describe developments of the kind in Example 4.22.1 as exemplifying the attainment of impersonal knowledge, of knowledge without a knower. Or, better still, let us note that knowledge without a knower is indicated whenever we speak of facts. After all, the trace at (4.22.7) simply conveys facts, does it not? Facts that can be forced upon the human body, is that not so? It cannot be said of co-ordination tests that they produce uncertain knowledge; it must rather be said of such tests that they produce factual knowledge.

4.25 A PERSUASIVE DIVERSION

Attempts at explaining how these data might have come about are by introduction of the knower of frequentist inference, persuasively diverted into attempts at keeping track of 'the knower's error rates'. That, in turn, leads to ideas of simultaneous statistical inference. In order to come to grips with this, we now develop, instead of the co-ordination tests of Table 4.20.1, and in respect of healing time, corresponding hypothesis tests imbedded in a stereotypic array. In order to display the reasoning of simultaneous statistical inference in its most accessible form, the version of that reasoning used here differs slightly from the version we introduced in Example 1.31.1 and used in Example 4.20.2.

Example 4.25.1

Let the Type I error rate of the prospective stereotypic array be specified as $\alpha = 0.01$, especially low error rates being widely recommended in cases of medical interest. We begin with the treated rats. Consider the test for skewness in Table 4.20.1. It employs a test statistic Z_1 whose hypothesised distribution is:

$$N(\theta, 1) \text{ with } \theta = 0, \tag{4.25.1}$$

and whose distribution might, for argument's sake, alternatively be:

$$N(\theta, 1) \text{ with } \theta = -3.608 \text{ say, in case of skewness to the left, or}$$
$$N(\theta, 1) \text{ with } \theta = +3.608 \text{ say, in case of skewness to the right.} \tag{4.25.2}$$

If $\theta = +3.608$ would be the only possible alternative, the appropriate decision rule for an array of one-sided tests of hypothesis of size $\alpha = 0.01$ would be:

$$\text{Reject } \theta = 0 \text{ when and only when } Z_1 \geq +2.327. \tag{4.25.3}$$

If $\theta = -3.608$ would be the only possible alternative, the appropriate decision rule for an array of one-sided tests of hypothesis of size $\alpha = 0.01$ would be:

$$\text{Reject } \theta = 0 \text{ when and only when } Z_1 \leq -2.327. \tag{4.25.4}$$

Now note (and this is crucial) that it is utterly impossible for a solitary data set in the real world to come about with $\theta = -3.608$ and also with $\theta = +3.608$. However, frequentist inference would then reason that, if either the one or the other of the two alternatives is possible, and if the specification $\alpha = 0.01$ is to be maintained by the knower, the rules at (4.25.3) and (4.25.4) must be replaced by:

$$\text{Reject } \theta = 0 \text{ when and only when either } Z_1 \geq +2.576 \text{ or } Z_1 \leq -2.576. \tag{4.25.5}$$

It appears at once that such inference has now replaced the problem of how this one particular data set might have come about with a different problem involving many different data sets, such that it might be that $\theta = -3.608$ in some sets, and $\theta = +3.608$ in other sets. The different problem now is how to construct a stereotypic array of decisions by the knower in respect of many different data sets, such that the Type I error rate of those decisions meets the specification. The issue is crucial because, on the one hand, it explains why simultaneous statistical inference has no place at all in the theory of co-ordination tests, and on the other hand, it also explains why simultaneous statistical

inference is unavoidably entailed by frequentist inference. Again: on the one hand, two contradictorily different explanations of how a particular real-world data set might have come about can serve as alternative explanations, but they cannot possibly simultaneously serve as an explanation. On the other hand, different terms in an array of real-world data sets might arise in contrary ways, and so, even though each term can arise in one way only, the properties of the array as a whole will reflect those contrary ways. Once again: a frequentist will reason that when Popper's 'knowing subject' uses the rule at (4.25.3) to test for skewness to the right, and then separately uses the rule at (4.25.4) also to test for skewness to the left, that knowing subject's Type I error rate equals 2×0.01. So a frequentist will reason that, in order to maintain the knower's overall Type I error rate at 0.01, the rates of each of the two tests must be specified separately to be $\alpha = 0.005$. So, at (4.25.5) the value 2.576 replaces the values 2.327 at (4.25.3) and (4.25.4).

There is more to come.

Table 4.20.1 proposes not only a test for skewness, but also a test for kurtosis using a separate N(0, 1) test statistic, say Z_2. So, the reasoning that has lead us to the decision rule at (4.25.5), when independently applied to Z_2, leads to the same rule, but with Z_2 in place of Z_1. Now, as indicated in Example 1.31.1, under the hypothesised model, Z_1 and Z_2 are approximately distributed as statistically independent random variables. So if the knower would use the rule at (4.25.5) to test for skewness, and then use the corresponding rule with Z_2 separately to test for kurtosis, the knower's Type I error rate would be 2×0.01. Thus, in order to maintain the overall Type I error rate at 0.01 when simultaneously testing for skewness to the left, skewness to the right, leptokurtosis and mesokurtosis, the rule at (4.25.5) and the counterpart rule involving Z_2, must be replaced by rules such that:

$(1-2\alpha)(1-2\alpha) = 0.99$, i.e. such that $\alpha = 0.0025$ per subsidiary rule.

The requisite subsidiary rule in respect of the model hypothesised at (4.25.1) and in respect of the second alternative at (4.25.2) is then:

Reject $\theta = 0$ when and only when $Z_1 \geq +2.810$. (4.25.6)

There is more to come.

So far we have considered the treated rats only, where the foregoing reasoning is repeated in respect of the control rats only. So, the knower's specified Type I error rate, 0.01, must then be maintained when a simultaneous test for skewness to the left, skewness to the right, leptokurtosis and mesokurtosis, is performed for both groups of rats. The rule at (4.25.6) and its counterpart involving the control rats must then be replaced, and the replacement must be such that α on the subsidiary rule satisfies

$(1-2\alpha)(1-2\alpha)(1-2\alpha)(1-2\alpha) = 0.99$, i.e. such that $\alpha = 0.00125$. (4.25.7)

The requisite subsidiary rule in respect of the model hypothesised at (4.22.1), and in respect of the second alternative at (4.22.2), is then:

Reject $\theta = 0$ when and only when $Z_1 \geq +3.025$. (4.25.8)

There is still more to come.

Table 4.20.1 also proposes using Snedecor's *F* as an additional commencement test, and under the hypothesised model *F* is distributed independently of Z_1 and Z_2. So, the specified Type I error rate of 0.01 must account for simultaneously testing, not only for skewness to the left, skewness to the right, leptokurtosis and mesokurtosis in both groups of rats, but also for homogeneity of variance against the two alternative cases of heterogeneity. The recipe at (4.25.7) must therefore be replaced such that α in the subsidiary rule satisfies

$$(1\text{-}2\alpha)(1\text{-}2\alpha)(1\text{-}2\alpha)(1\text{-}2\alpha)(1\text{-}2\alpha) = 0.99, \text{ i.e. such that } \alpha = 0.001. \qquad (4.25.9)$$

The requisite subsidiary rule in respect of the model hypothesised at (4.22.1) and in respect of the second alternative at (4.22.2) is then:

$$\text{Reject } \theta = 0 \text{ when and only when } Z_1 \geq +3.080. \qquad (4.25.9)$$

The outcome of such simultaneous statistical inference in respect of the hypothesis, i.e. that the two sets of healing times be modelled as two independent homoscedastic normal samples, is then as follows:

> As the smallest of the relevant significance levels in Table 4.20.1 is SL = 0.005, which exceeds the specification of α = 0.001 at (4.25.9), the simultaneous tests would have us accept the hypothesis. (4.25.10)

This lands us on the horns of a dilemma.

Source of the first horn: a knower advancing toward 'more conservative' policies

The reasoning in Example 4.25.1 steadily advances toward more conservative policies. At (4.25.3) and (4.25.4) we must turn a blind eye to normal deviates less than 2.327 standard error units as being not significant. However, in order to conserve the overall Type I error rate at 0.01, that instruction is first replaced with 2.576 units as 'not significant', then with 2.810 units, then with 3.025 units, and finally with 3.080 units. As a result the Type I error rate for any of the 2×2×2×2×2 constituent tests is reduced

> from 0.01, to 0.005, to 0.0025, to 0.00125, and finally to 0.001. (4.25.11)

Source of the second horn: a knower retreating to 'less conservative' policies

To the unwary, the advance toward more conservative policies might well seem fine. After all, did we not begin with the idea that, as medical research requires special 'protection' from committing 'errors', we must tolerate only satisfactorily low Type I error rates (such as 0.01 or even less?) This reasoning is grossly misleading because it diverts our attention from Type II error rates. For instance, at (4.25.2) we considered the possibility that θ = +3.608 owing to skewness to the right. We noted that if that were to be the only possible alternative in respect of that constituent amongst 2×2×2×2×2 constituent tests, the appropriate rule for an array of one-sided hypothesis tests of size 0.01 would be the one given at (4.25.3). In that case the Type II error rate for that particular constituent would be given by

$$\Pr(Z_1 \geq +2.327 | \theta = +3.608) = 0.1. \qquad (4.25.12)$$

However, when the rule at (4.24.3) is replaced by the rule at (4.25.5), this increases to

$$\Pr(Z_1 \geq +2.576 | \theta = +3.608) = 0.15. \qquad (4.25.13)$$

Proceeding further, it will thus be found that by reducing the Type I error rate for each of the 2×2×2×2×2 constituent tests

from 0.01, to 0.005, to 0.0025, to 0.00125, and finally to 0.001,

as indicated at (4.25.11), the Type II error rate for the particular constituent test under consideration at (4.25.12) and (4.25.13), has been increased

from 0.1, to 0.15, to 0.21, to 0.28, and finally to 0.30, respectively.

Such increases accrue for each of the various 2×2×2×2×2 constituent tests, with a silly consequence made glaringly obvious by Table 4.20.1, as follows; it is inconceivable that any competent statistician would defy that table in order, as an act of will, to conclude that the model, 'two independent homoscedastic normal samples', fits healing times better than it fits healing rates. Yet at (4.25.10), simultaneous statistical inference would have us draw precisely that conclusion. So, in order to avoid such silly conclusions, a frequentist is compelled to retreat to some less conservative policy. Note, for instance, that the moot point raised in connection with the tests at (4.20.10) arises because Bliss preferred such a less conservative policy.

The dilemma

If the observed value of Z_1 would be for argument's sake $Z_1 = 3.080$, the co-ordinates of its mental correlate would be:

(0.999, ε, 0.001) in the hypothesised distribution, and
(0.299, ε, 0.701) in the right skew alternative distribution considered at (2.25.2).

Yet the rule at (4.25.10) would have us conclude that the hypothesised model is more tenable than any right skew alternative. In general, simultaneous statistical inference cannot avoid such difficulties because whatever device is used to conserve the overall Type I error rate at its specified value, will reduce the Type I error rates of subsidiary constituent tests, thus enlarging the Type II rates of those constituents. The larger the number of constituents, the more pronounced this effect. So the literature on simultaneous statistical inference is forever wrestling with a self-inflicted dilemma: must one (this is the first horn of the dilemma) pursue more conservative policies, so as to conserve a specified Type I rate? Or must one (the second horn of the dilemma) retreat toward less conservative policies, so as to avoid the more obvious of the silly consequences of such policies?

4.26 The knower is not needed

The dilemma in the previous section cannot be resolved sensibly because, in defiance of the principle underscored at (4.24.2), frequentist inference introduces, by way of the knower, a constituent that is not needed. For that reason precisely, the 'correct extent' of the influence of that constituent cannot be established. That is the case because any

redundancy is absolute; one redundancy cannot be more redundant, or less redundant, than another redundancy. Also, whilst avoiding unneeded constituents is a universal requirement of science, in certain sciences this is not easily accomplished. In ethology it is difficult to avoid the introduction of anthropomorphic constituents; in developing the theory of biological evolution, it is difficult to avoid the introduction of teleological constituents. And we have now seen that in statistical data analysis it is difficult to avoid the introduction of a knowing subject. So, having established that the knower of Neyman's theories is redundant, nothing but an intruder in the domain of data analysis, the knower must be banished from that domain.

4.27 Getting it wrong from the outset in the case of frequentist inference

As the various theories of statistical inference spring from great mathematical talent, it is not to be expected that they involve mathematical errors. They are fundamentally defective owing, instead, to epistemological errors arising at the very outset by way of defective understanding of the investigative method of substantive science. It is of the utmost importance to grasp this, because ours is a mathematically minded profession, and those who are so minded are notorious for losing track of reality. So it must be firmly grasped that the circular reasoning displayed in Section 4.8 is the crux of the matter. Section 4.8 shows us how the mathematically minded, having caught sight of a class of mathematically formulated models indexed by a parameter, are so taken with the idea of getting on with the mathematics that the question: 'How can such a class of models be arrived at?' is overlooked. The result is an epistemology that gets it wrong from the very outset; that takes the first wrong step by proceeding from the idea that decision-making under risk, which concerns the use of knowledge, can be adapted for purposes of data analysis, which concerns the pursuit of knowledge. And so, as we found in Section 4.8, that epistemology falls at once into circularity for which there cannot be a subsequent correction. Beyond Section 4.8, this chapter simply concerned the consequences of that circularity. The stereotyping, the paradoxes, the distortion of evidence as in Example 4.25.1, the need for 'acts of will' to override common sense, the silly idea that any 'knowledge' has to partake of its 'knower', and so on – all these defects are consequences of that first wrong step.

4.28 Neyman's 'knowing subject'

Measurements of the magnetism of certain dated rocks provide a data set that can be used to form an opinion on the possible positions of the earth's magnetic pole at that date. The data set can be represented as a cluster of points on the surface of a sphere representing the earth, and Fisher (1953) developed a class of models indexed by the possible positions of the pole. By ignoring the slight curvature of the earth's surface, a class of models more suitable for our purposes is obtained in the form introduced for the target problem in Section 3.11, except that we must then suppress the j-like count of real-world repetitions, as there is now just one single data set in the real world. The unknown position of the pole is then represented by Cartesian co-ordinates (μ_1, μ_2), and

the minimal sufficient statistic for (μ_1, μ_2) is given by the following triplet of statistically independent random variables:

$\overline{X}_1$, which is an N(μ_1, $\sigma^2 \div n$) random variable,
$\overline{X}_2$, which is an N(μ_2, $\sigma^2 \div n$ random variable, and
2(n-1) S^2, which is a $\sigma^2 x^2$ random variable on 2(n-1) degrees of freedom.

Reasoning similar to that in section 3.11 here leads to a (1-α) confidence region in the form of a disk centred at:

$(\overline{X}_1, \overline{X}_2)$ with radius $\sqrt{2(S^2 \div n)F(\alpha)}$,

where, as before, $F(\alpha)$ is the value that, with specified probability α, is exceeded by Snedecor's F on 2 and 2(n-1) df. But what, in this example, does '(1-α) confidence' mean? We can answer only in terms of a population of repetitions. But that cannot be!

Earth 1 in repetition 1? Earth 2 in repetition 2? Earth 3 in repetition 3? ...

So, for frequentist inference to defend its model of a forecasted probability of success in '(1-α) real-world instances', those instances have to be modelled as instances of inferential behaviour by 'a knowing subject'. We must not agree to such a model, as substantive investigation asked for geological models of how these given data might have come about. It is then perfectly possible that such models may *inter alia* require statistical constituents; in the present case they clearly do require such constituents. Clearly also, in the matter of how these given data might have come about, Neyman's knowing subject played no role, none whatsoever.

4.29 Stereotypic arrays and decision-making under risk

We return to Section 3.13 where the idea was raised that in decision-making under statistical risk, not to be mistaken for data analysis, reasoning in terms of a stereotypic array is justified. Consider, for instance, the manager of a factory for the manufacture of say lawn mowers, who will continually have to deal with matters such as:

> Running sampling inspections to reach a decision on whether or not to buy certain raw materials.
> Running tests to reach a decision on which of five new welding plants to purchase.
> Running experiments to reach a decision on whether a new method is better than the old one. (4.29.1)

Does it not make sense, so the counter-argument of Section 3.13 would try to persuade us, to imbed the decisions in a stereotypic array so that the manager's Type I error rate can be kept to, say, a trifling 1%? We must answer in the negative, because this argument is essentially circular, the principle having been pleaded by the phrase 'not to be mistaken for data analysis', in the first sentence of this section. This is so because the array envisaged at (4.29.1) does not comprise homogenous terms, but comprises instead precisely the motley circumstances that lead to the dilemma we developed in Section

4.4. And so, willy-nilly, the manager is a data analyst who would otherwise be prevented from substantive thought.

4.30 How is frequentist inference sustained despite its obvious defects?

In the light of the foregoing this question would seem inescapable. The answer is that it survives by way of brain-washing, as follows: a typical present-day introduction to statistics starts its development of statistical inference by arguing persuasively that such inferences (note the plural) might, by chance, be erroneous, i.e. are subject to statistical risk. Typically this idea is then developed via confidence intervals 'because they are easiest for the students to understand'. Then hypothesis tests are developed as the dual of confidence intervals 'because that is easiest for the students to understand'. Next, there follows a process of entrenchment by way of developing concepts such as the power of a test, most powerful tests, one-sided, two-sided and unbiased tests, monotone likelihood ratios, and so on, along with a wealth of illustrative examples. Then comes the point at which, having started with the idea of a confidence interval, the development ends with what can only be described as 'an inadvertent confidence trick' in the form of a section showing how the 'assumptions' (sampling having been 'assumed' to be from binomial, Poisson, normal, etc. populations) can be 'tested' by (and this is the trick) hypothesis testing presented in such a way that only Type I errors are accounted for. The students have by then been dazzled by pretty mathematics and carefully conceived examples to the extent of not noticing that the 'assumptions' would, for their justification, require accounting for Type II errors. In the process, the lecturer himself/herself, having previously (as a student) undergone such brain-washing, becomes even more deeply entrenched in belief in the veracity of the matter. This curiously back-to-front development is reinforced by psychological reluctance to admit that acceptable assumptions for frequentist inference are subject to Type II errors (acceptance errors) of unknown frequency, otherwise the elaborate construct called frequentist inference must, like Humpty Dumpty, undergo a great fall. To see this, one only has to reverse the back-to-front reasoning, in which case it becomes apparent at once that commencement tests do not justify the 'assumptions', because those assumptions are tantamount to assuming that commencement tests can, for specified Type I error rates, deliver Type II error rates that are equal to zero.

Chapter 5

SIGNIFICANCE TESTS

R.A. Fisher's method for avoiding the frequentist vicious circle

5.1 Introduction

In the previous chapter we saw that when elimination testing is construed as a form of hypothesis testing, we fall into a vicious circle. This is because we cannot claim to achieve elimination tests at specified Type I error rates without then having to assume that we can discern the appropriate class characteristics *without any errors of Type II*. The purpose of the present chapter is to show that if elimination tests are construed as significance tests, the circularity is avoided, or so R.A. Fisher evidently believed, and in this chapter we, for explanatory purposes, do not challenge that belief.

5.2 Significance levels as measurements of quality of fit

We consider below three problems in data analysis. In each case we first present an analysis that arrives at certain findings by means of co-ordination tests, and then we show how the same findings can be expressed by way of significance tests.

Example 5.2.1

Suppose that in an experiment with laying hens, each of three breeds was represented by 100 hens individually caged in a completely randomised design, and that a record was kept of the numbers of eggs laid per hen per day – the more the better. Table 5.2.1 shows hypothetical data and a suite of shortfall tests using Dunnett's many-one-*t* statistic for the elimination of breeds whose egg production can be tenably modelled as lower than best only.

Table 5.2.1: A suite of shortfall tests for eliminating breeds as 'lower than best'

Entry	Mean number of eggs laid per hen per day	Shortfall	Many-one-*t*	Left-most hypothesised co-ordinates
Breed A	0.9975	-0.0816		
Breed B	0.9159	+0.0816	+1.865	(0.944, ε, 0.056)
Breed C	0.8938	+0.1037	+2.371	(0.983, ε, 0.017)
Estimated standard error of an entry mean: 0.03093 on 3(100 – 1) df.				

The model employed represents the raw data as three samples originating from homoscedastic normal populations. So, a data analyst might well use Pearson's chi-square to

test for non-normality on say 30 df., and Bartlett's chi-square to test for heterogeneity of variance on 2 df. Let the mental correlates of the resulting test data be found to be situated in the appropriate test distributions at (0.63, ε, 0.37) and at (0.24, ε, 0.76), respectively. In such a case the data analyst might report as follows:

> Normal errors, by the Pearson test, fit the given data well, as chi-square = 31.94 on 30 df. is situated at (0.63, 0.37) in the test distribution. Homoscedastic errors, by the Bartlett test, also fit the data well, as chi-square = 0.549 on 2df. is situated at (0.24, 0.76) in the test distribution. By the shortfall tests, both B and C seem lower than best, as the corresponding many-one-*t* values must be co-ordinated to the right of (0.944, 0.056) and (0.983, 0.017), respectively, in Dunnett's test distribution. So, by the tests performed, A appears to be the sole best entry. (5.2.1)

Now note (and this is crucial) that this report involves no assumptions whatsoever, as four distinctly different tests were used to test four distinctly different subsidiary models against four distinctly different subsidiary data sets. As explained in Example 2.3.2, we make no assumption when we point at a spoor saying: 'The *shape* of this spoor *is unlike* that of an aardvark, but if it *were* to be that of an aardvark, *the size* of the spoor *is like that* of a juvenile.' Instead of making assumptions, we thus test a model of shape against a datum of shape and, *quite apart from that*, we test a model of size against a datum of size. Similarly, at (5.2.1), instead of making assumptions, a model of distributional shape is tested against a datum of distributional shape, and apart from that a model of comparative variability is tested against a datum of comparative variability, and apart from that ... .

A significance tester would issue essentially the same report as the one at (5.2.1), except that the four sets of statistical co-ordinates would be replaced by the corresponding significance levels,

$$\text{SL} = \varepsilon{+}0.37,\ \text{SL} = \varepsilon{+}0.76,\ \text{SL} = \varepsilon{+}0.056,\ \text{SL} = \varepsilon{+}0.017, \qquad (5.2.2)$$

respectively. A significance tester looks upon such significance levels as *measures of fit*, and holds that any reasonable interpretation of the term 'error rate' would make it utterly impossible for the four SL values at (5.2.2) to provide anything in the nature of an error rate for the conclusion drawn. In order to understand this view, we note that in respect of the substantive subject matter of the investigation, the whole conclusion is simply this:

> Breed A appears, by the tests performed, to be the sole best entry (5.2.3)

Thus if the error rate in question is to be substantively relevant, it must apply to the conclusion at (5.2.3). But if we try to appraise the contribution of Bartlett's test to the requisite rate, we must consider an acceptance error (a Type II error) whose rate is a function of unknown variance parameters. Worse: if we try to appraise the contribution of Pearson's test to that requisite rate, we must provide for acceptance errors for which we cannot even provide useful mathematical expressions, as we would then be testing against various alternatives that are only broadly envisaged possibilities. We cannot even know in respect of which of those various alternative possibilities the appraisal would have to be made – unless we were clairvoyant.

Example 5.2.2

Let us revisit Table 1.1.1, which gives the results of an experiment to measure the effect of carbaryl treatment on the half-lives of the fruit of certain trees. Shapiro-Wilk tests for normality performed in Example 1.32.1 result in a test datum whose mental correlate co-ordinates at (0.18, ε, 0.82) in the test distribution. The variances of the observed half-lives are given by:

$S_1^2 = 1.126$ for the treated trees, and $S_2^2 = 0.803$ for the control trees.

Using Snedecor's *F* to test for homogeneity of variance, we find:

$S_1^2 \div S_2^2 = 1.405$, where Pr(Central *F* on 4 and 4 df. > 1.405) = 0.38,

showing that the datum *F* is situated at (0.62, ε, 0.38) in Snedecor's test distribution. So, using the pooled error estimate (1.126+0.803)÷2 and Student's *t*, we find that at the (0.05, ε, 0.95) level of co-ordination, or to the left of that, the effect of carbaryl is modelled as reducing the half-life of the fruit by 1.7 days or more.

A significance tester, taking $\varepsilon \approx 0$, would report essentially the same findings, as follows:

> As tested, there is no significant non-normality (SL = 0.18) and no heterogeneity of variance (SL = 0.38). With Student's *t* held at any level of significance ≤ 0.05, the effect of carbaryl must be modelled as hastening ripening time by 1.7 days or more.

Here again, the significance tester cites the various significance levels as *measures of fit*, and holds that in any reasonable interpretation of the term 'error rate' those levels cannot supply an error rate for the conclusion drawn. In order to grasp this view, we note that from the point of view of substantive science, the thrust of the conclusion is:

> Carbaryl appears, by the various tests performed, to shorten the half-life of such fruit by at least 1.7 days. (5.2.4)

Thus, if the error rate in question has to be at all substantively relevant, it must apply to the conclusion at (5.2.4). But in order to appraise the contribution of the test based on Snedecor's *F* to the requisite rate, we find we must consider an acceptance error (a Type II error) whose rate would have to be given by:

$$\Pr\left[\,(\textit{Central } F \textit{ on } 4 \textit{ and } 4 \textit{ df.})\times\left(\frac{\sigma_2^2}{\sigma_1^2}\right) \geq 1.405\right] \text{ . However, } \left(\frac{\sigma_2^2}{\sigma_1^2}\right) \text{ is unknown.}$$

Worse: in order to appraise the contribution of the Shapiro-Wilk tests to the requisite rate, we have to consider an acceptance error (a Type II error) for whose rate we are not even able to give a useful mathematical expression, as the test is against alternatives that are at best incipient. And in any case, as those alternatives cannot even be known to be of just one particular kind, we would not know in respect of which of them that contribution would have to be appraised.

Example 5.2.3

In Example 1.15.2 we found that the volcanic action of Vesuvius, from AD 79 up to and including AD 1944, falls into two distinctly different eras. For each era, the waiting times between eruptions, as judged by the Cramer-Von Mises test, can be represented satisfactorily as a random sample from an exponential population. For the ancient era, the mental correlate of the test datum is situated at (0.85, ε, 0.15*) in the Cramer-Von Mises test distribution, and for the modern era the mental correlate of the tests datum is situated at (0.87, ε, 0.13*) in the Cramer-Von Mises test distribution. Let the sample means and their expected values be denoted as follows:

For the $n_1 = 10$ waiting times of the ancient era: $\overline{X}_1$ with $E(\overline{X}_1) = \theta_1$.
For the $n_2 = 20$ waiting times of the modern era: $\overline{X}_2$ with $E(\overline{X}_2) = \theta_2$.

With exponential waiting times an elimination pivot for θ_J is given by $2n_J\overline{X}_J \div \theta_J$, which is distributed as central chi-square on $2n_J$ df ($J = 1, 2$). Hence, by specifying

U = 0.025, 0.050, 0.075 as elimination levels,

the corresponding upper elimination bounds are found to be:

θ_1 = 324, 286, 267 years, respectively, and
θ_2 = 25.6, 23.6, 22.5 years, respectively.

So, expressing these findings in terms of models that, as tested, are consonant with the data, a significance tester would report:

> For each of the eras separately, a model of exponential waiting times, as tested, is consonant with the given data (SL = 0.15 for the ancient era, and SL = 0.13 for the modern era). For the ancient era, $\theta_1 \leq$ 324, 286, 267 years, as tested, are consonant with those data at significance levels ≥ 0.025, 0.050, 0.075, respectively. For the modern era, $\theta_2 \leq$ 25.6, 23.6, 22.5 years, as tested, are consonant with those data at those same levels, respectively.

We note in passing that in respect of the theory of significance tests, rather than that of hypothesis tests, a dual theory of consonance intervals, rather than confidence intervals, is indicated (Kempthorne and Folks 1971, Section 13.2).

In terms of the concept 'confidence' rather than 'consonance', a subscriber to the dual theories of hypothesis tests and confidence intervals has to specify, *without reference to the data*, a single Type I error rate, say α = 0.05, so as to be able to claim that:

$0 < \theta_1 <$ 324 years, is 'a 0.95 confidence interval for θ_1', and
$0 < \theta_2 <$ 23.6 years, is 'a 0.95 confidence interval for θ_2'. (5.2.5)

Here the expression, 'a 0.95 confidence interval for θ', *intentionally* claims to have obtained each of the two intervals in question

> by a method for which it can be forecasted that in 0.95 of cases, the corresponding true value of will be contained in the interval. (5.2.6)

In order to defend the forecast made at (5.2.6) one must be able to provide reasons to believe that each of the two sets of waiting times somehow belongs to a host of cases that can be represented as exponential samples. But if we try to cite the results of the Cramer-Von Mises tests as reasons for such belief, we find ourselves forced to admit that those tests are open to Type II errors, otherwise we are unable to explain why we performed those tests in the first place. Moreover, should those Cramer-Von Mises tests have involved any Type II errors, we could not possibly have known that to have been the case. We could not even have known what alternatives would be involved. So we cannot possibly know what the values of the corresponding Type II error rates would be. Nevertheless, by simulating what would happen with this, that or the other choice of reasonably possible alternatives, we can show empirically that if, in respect of such a choice, β_1 and β_2 denote the Type II error rates for the ancient era and the modern era, respectively

$0 < \beta_1 < \beta_2$, owing to the difference in sample size ($n_1 = 10$ versus $n_2 = 20$).

However, in order to defend the forecasts made at (5.2.5) and (5.2.6), we are forced to *assume that*

$0 = \beta_1 = \beta_2$,

so that it can be reasoned that the class characteristic is exponential,
so that it can be reasoned that the method used has the property claimed at (5.2.6), *and that* is circular reasoning.

5.3 Concluding remark

The use of significance tests, as outlined in the foregoing examples, might seem to avoid the frequentist vicious circle. However, we must ask: 'Can it do so without violating any fundamental principle of science?' In the following chapter we will show that this question can only be answered in the negative.

Chapter 6

AN INADVERTENT CONFOUNDING ON THE PART OF R. A. FISHER

The seminal source of 'simultaneous statistical inference'

6.1 Introduction

We are now ready to come to grips with a fundamental and irreconcilable distinction between significance tests and co-ordination tests. We will develop the distinction by uncovering a deeply hidden flaw in R. A. Fisher's theory of significance tests. As we explained in the previous chapter, he proceeds from a population in the human mind, which is considered as a model of how a given data set in the real world might, or might not, have come about. Often a model can be analysed into several subsidiary models. Different significance tests can then be used to measure the quality of fit of the different subsidiary models. The resulting measurements are called 'significance levels'. How are such levels defined? In its treatment of this question the statistical literature has been astoundingly careless, where, as pointed out by Freund and Perles (1993), either one or the other of two non-equivalent definitions is often used. Kendall and Stuart (1961) for instance, introduce one of these definitions for hypothesis tests, and then tacitly rely on the other one for confidence intervals (vol. 2, ch. 20 and 22, respectively). The two definitions are developed in the following section, and in the section after that we identify one of them as Fisher's definition. In subsequent sections we exhibit the flaw in Fisher's theory, and we show that removal of the flaw has devastating consequences.

6.2 A pair of non-equivalent definitions

Let a measurement made on the amnion fluid of a pregnant rabbit be modelled as the value taken on by a random variable, X, whose expectation, μ, represents the number of foetuses present ($\mu = 1, 2, 3, \ldots < \infty$). Write $Z = X-\mu$. Let a historical record, thus involving known values of μ, provide $z_1, z_2, z_3, \ldots, z_n$, capable of being modelled as a random sample of Z values. In order to test whether or not $z_1, z_2, z_3, \ldots, z_n$, could be modelled as a sample from a *normal* population, let an investigator use an appropriate chi-square test, obtaining chi-square equal to say 10.5 on 8 degrees of freedom. Here 10.5 is a class mark, i.e. an approximation for some or other hypothesised sample value between say 10.25 and 10.75, which value is otherwise indeterminate. Thus the chi-square test divides the infinite population of hypothesised chi-square values into separate parts measuring say U, ε and V, respectively, as depicted in Figure 6.2.1., and where

$(U, \varepsilon, V) = (0.75, 0.03, 0.22)$ in the present case.

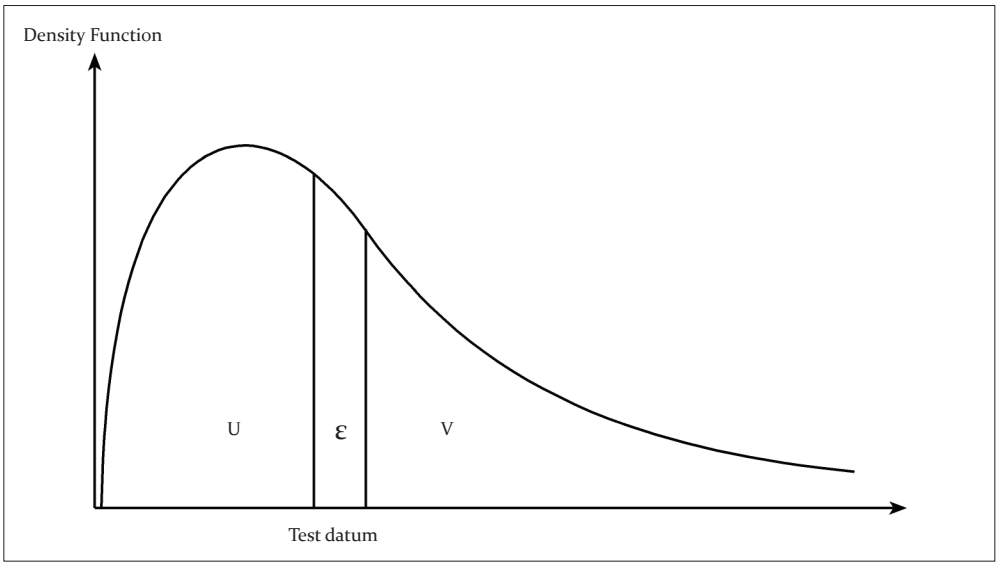


Figure 6.2.1: A depiction of how, in the human mind, a chi-square test divides an infinite population of hypothesised chi-square values into parts labelled U, ε and V, respectively

The diagram in Figure 6.2.1 depicts a good fit, thereby showing how the normal class characteristic, as tested, would fit the given data well, where such facts can of course be forced upon the human body, if needs be by simulation. The question now to be considered, and indeed to be considered throughout this chapter, is simply this:

> How can we best report such facts by way of numbers? (6.2.1)

The correct answer to this question can only be:

> Such a report must be fully informative; it must supply whatever information we require for a reconstruction of the diagram. (6.2.2)

So, at least two of the numbers U, ε and V, must be reported. For instance, identifying U and V as left- and right-statistical co-ordinates, respectively, we could report the co-ordinates of the mental correlate of the given chi-square value as (0.75, 0.22). Given this report, the facts depicted in Figure 6.2.1 can be recovered, as the value of ε, the statistical rounding, is recovered by calculating 1-0.75-0.22. If χ^2 denotes the random variable involved, the statistical co-ordinates in question are given by:

$$(U, V) = [\Pr(\chi^2 < \text{the observed chi-square}), \Pr(\chi^2 > \text{the observed chi-square})].$$

In practice the statistical rounding is often (even usually) very small. So, one can adopt a convention that tacitly understands a rounding to be negligibly small whenever only one of the two co-ordinates, say the right co-ordinate, is reported. In the present case the only value we would then report would be

$$\Pr(\chi^2 > \text{the observed chi-square}) = 0.22. \quad (6.2.3)$$

The expression at (6.2.3) corresponds to Definition 2 of Freund and Perles (1993), except that their definition is unrealistic with regard to rounding, as they presuppose a means of exact measurement of any real number values, where such a means cannot exist. The other of their two definitions, their Definition 1, corresponds in the present case to

$$\Pr(\chi^2 \geq \text{the observed chi-square}) = 0.03+0.22. \qquad (6.2.4)$$

From the point of view of significance testing, this definition is the more attractive of the two, as a decidedly poor quality of fit is attained if and only if both the rounding and the pointing co-ordinate are small. (In a subsequent development it will be found that, from a different point of view, their Definition 1 is the more attractive. So let us not dismiss it out of hand.) The expressions at (6.2.3) and (6.2.4) might have arisen from an ordering of the form

$$O_1 = \{0 \leq \chi^2 < 0.5+0.25\} \text{ and } O_t = \{0.5t-0.25 \leq \chi^2 < 0.5t+0.25\},\ t = 2, 3, 4, \ldots,$$

where $t = 21$ for the given datum, $0.5(21) = 10.5$. In the present case this ordering is right-sensitive with regard to the usual alternatives that come to mind, in which case the inverse ordering is of course left-sensitive to those alternatives. In order to avoid unnecessary complication, the following development uses right-sensitive orderings, unless stated otherwise. It will be found convenient to refer to a significance level as a 'P value', it having been made entirely clear in the previous chapter that in Fisher's theory such values are not *specifications*; they are *observations*. They do not belong to the discourse of *forecasting*; they belong to the discourse of *pointing*. This being clearly understood, the Definitions arising at 6.2.4 and 6.2.3 can respectively be stated very simply, as follows:

Definition 6.2.1:

The P value in any given case is the hypothesised complement of the left statistical co-ordinate in that case. (We will call this *the inclusive definition* since it 'includes' the rounding.)

Definition 6.2.2:

The P value in any given case is the hypothesised right statistical co-ordinate in that case. (We will call this *the exclusive definition* Since it 'excludes' the rounding.)

We will also require the following definition:

Definition 6.2.3:

The attainable P values in any given case are those P values that might actually be observed under either Definition 6.2.1 or Definition 6.2.2.

Example 6.2.1

Consider just N independent attempts to discriminate by taste between two items. Let the number of successful attempts be modelled as a value taken by a binomial random variable, X. The appropriate ordering for testing

M_0: Pr(success) = 0.5 versus M_1: Pr(success) > 0.5,

is given by $O_T = O_{X+1}$ for X = 0, 1, 2, ..., N. Using Pascal's triangle, we then find that with say N = 3, the attainable P values are given by

8/8, 7/8, 4/8, 1/8, 0/8, where
8/8 is attainable under Definition 6.2.1 only,
0/8 is attainable under Definition 6.2.2 only,
and the other four P values are attainable under either of the two definitions.

We note in passing that for this example the exclusive definition is clearly defective, as the hypothesised right co-ordinate is evidentially vacuous whenever X = N.

6.3 R.A. Fisher's usage of the term 'significance level'

In R.A. Fisher's writings small P values point at alternatives. His paper on Mendel's data is a rare exception to this rule. However, he does not always make his choice of definition entirely clear. For instance, in *Statistical methods for research workers* (1970, p.79) he informally introduces the notion of a P value as if by the exclusive definition. In effect he then uses such a value in the subsequent development of a concrete example (p. 95) when, referring to 'a normal deviate 3.61 times its standard error' he says: 'The probability of exceeding such a deviation in the right direction is about 1 in 6 500.' However, in a subsequent section entitled *The exact treatment of* 2×2 *tables*, he recognises the statistical rounding and its right co-ordinate when he refers to 'the probabilities of the set of frequencies observed, and the two possible more extreme sets of frequencies which might have been observed' (p. 97). He then carefully introduces the inclusive definition by calculating the significance level as the sum of the three probabilities. At the bottom of p. 100, however, he reverts to the exclusive definition by referring to 'the probability of χ^2 exceeding 11.417'. In *The design of experiments* (1966) he develops the inclusive definition in a closely reasoned section entitled *The test of significance* (pp. 13-15). Subsequently, however, he again expresses himself in a way that amounts to the use of the exclusive definition (e.g. on p. 37). These, and other of his writings, compel us to make the following conclusions:

For closely reasoned explanations he favours examples involving a discrete test statistic, and then always uses the inclusive definition, and calculates the P value as an observed measure of quality of fit.

When using continuous approximation he often abbreviates 'equal to or exceeding' to 'exceeding', thereby using the exclusive definition, but clearly not intentionally so, as he uses such an abbreviation only when the statistical rounding is negligibly small. So Fisher's *intentional* definition is the *inclusive* one.

> In all his writings, the term 'significance level' refers to just one number, which can be arrived at exactly only by calculation from given data, but which is nevertheless intended by him to be interpreted as a hypothetical probability.

We are compelled to the third conclusion when we ask: 'Why does Fisher insist on the replacement of the three numbers U, ε and V with just one number, P?' We will find that the only answer capable of defence is that he wishes to make a formal probability statement, where a probability can only take the form of just one number, not of two or three numbers. Nevertheless, as our first conclusion states, he wants the probability to serve as a calculated measure of the tenability of a conceptual singleton under test to explain how a given real-world data set might or might not have come about. All this is made abundantly clear by Fisher himself when he says (1973, p.47):

> 'In general tests of significance are based on *hypothetical* probabilities calculated from the null hypotheses. They do not generally lead to any probability statements about the real world, but to a rational and well-defined measure of the reluctance to the acceptance of the hypotheses they test.' (original italics) (6.3.1)

Furthermore (and this is a crucial point) Fisher uses a behavioural interpretation when considering the importance of a calculated significance level. This point has been carefully explained by Cox and Hinkley (1974) and Cox (1977) and, as we will show in subsequent development, their explanation is correct beyond reasonable contest. They employ the notations T and $t = t_{obs}$ (observed t) to underscore that T denotes a test statistic in the human mind, whose range includes a value denoted by t_{obs}, which value is also that of a test datum computed from a given real-world data set. Ordering is on the magnitude of T, and the corresponding 'observed significance level' is then calculated in terms of the inclusive definition as

$$p_{obs} = \Pr(T \geq t_{obs} \mid M_0), \text{ where } M_0 \text{ denotes the hypothesised model.} \quad (6.3.2)$$

We note in passing that this equation transforms T into an equivalent, but left- sensitive, test statistic P, for which the corresponding test datum is p_{obs}. Thus also

$$p_{obs} = \Pr(P \leq p_{obs} \mid M_0), \text{ where } M_0 \text{ denotes the hypothesised model.} \quad (6.3.3)$$

Fisher, as quoted at (6.3.1), would have us look upon the calculated significance level as a measure of our reluctance to accept the hypothesised model it tests. Moreover, we will find that Fisher attaches to this measure a well-defined physical meaning, which meaning is behavioural, and which meaning Cox and Hinkley (1974, p. 66), using the notation H_0 instead of M_0, have carefully spelled out in terms of the inclusive definition (of Fisher), as follows:

> 'Suppose that we were to accept the available data as evidence against H_0. Then we would be bound to accept all data with a larger value of t as even stronger evidence. Hence p_{obs} is the probability that we would mistakenly declare there to be evidence against H_0, were we to regard the data under analysis as just decisive against H_0.' (6.3.4)

Here we must of course understand that a small P value is vacuous unless pointing at a substantively conceivable alternative. Fisher (1970, p. 95) reminds us of this when

referring to a test datum as deviating 'in the right direction'. Note also that at (6.3.4) the expressions: 'suppose we were to', 'then we would be bound to' and 'were we to regard', serve to underscore that the behaviour referred to and the probability attached to it, are *purely hypothetical*. Stated otherwise, p_{obs} does not denote 'an error rate in the real world', but rather 'an error rate in a conceptual world *that might have been*'. For this reason Cox and Hinkley (p. 66) hasten to explain further, that as a general rule:

> '... we are not especially interested whether p_{obs} exceeds some preassigned value, like 0.05. That is, we do not draw a rigid borderline between data for which $p_{obs} > 0.05$ and data for which $p_{obs} \leq 0.05$.' (6.3.5)

So p_{obs} is given a two-fold meaning:

> Firstly, as explained at (6.3.4), p_{obs} is the probability that certain behaviour would, under given hypothetical circumstances, lead to false rejection of the hypothesised model. Secondly, as explained at (6.3.5), p_{obs} must also serve as a measurement of fit on a scale higher than dichotomous. (6.3.6)

We propose to prove, in subsequent sections, that this attempt at a two-fold meaning inadvertently confounds two irreconcilably different concepts. But before proceeding to that, we must first show, in this section, that the explanations at (6.3.4) and (6.3.5) correctly render R. A. Fisher's idea of 'a significance test'. So, to begin with, we note that, as eluded to by the first sentence at (6.3.1), significance levels resemble Type I error rates that might have been, rather than Type II error rates that might have been. This is so because in Fisher's usage, significance levels must be *calculable*, as they must serve as *measurements*. The incipient alternatives of commencement testing, for instance, cannot provide for calculations. So, as a general principle, significance levels are calculated from hypothesised models only. In the theory of significance testing as such, this principle also applies to elimination testing. Also, apart from calculability, a significance level must satisfy two further requirements:

> Firstly, it must convey a physical meaning; one that is capable of being forced onto the human body, if needs be by simulation. Secondly, the meaning conveyed must show in some sense whether the hypothesised model is, on the measurement made, more appropriate, or less appropriate, than the alternative one has in mind.(6.3.7)

This much is recognised by Cox (1977, p.50) when he states that a statistical test must produce an evidential concept with a physical meaning that is appropriate to the use of the test, and the explanations given at (6.3.4) and (6.3.5) are clearly aimed at these requirements. It then remains only to show that Fisher would have to agree with those explanations. In order to do so, consider how we might try to understand Fisher's book, *Statistical methods for research workers* (1970), if we knew all about co-ordination tests, but nothing at all about significance tests. We might find the book easy reading up to say p. 60 where it might then seem that in the case of a very small rounding, the result of a test is to be summarised by reporting the value of the pointing co-ordinate only. This interpretation might be accepted till we reach p. 97 where Fisher computes a 'significance level' as the sum of the rounding and the pointing co-ordinate. Giving this some thought, we might find it to be an acceptable further recipe for summarising the results of a test where, for instance, we need not bother to distinguish between say

(0.96, 0.3, 0.1*) and (0.96, 0.1, 0.3*),

even though the co-ordination on the left *does* allow for slightly more extreme placing of the modelled datum than does the co-ordination on the right. Next, on p. 120, 'the Table for *t*' in fact refers to a table for $|t|$, where we would then think of that as a very odd way of tabulating *t* co-ordinates. Why tabulate 2V rather than V, and then have to say: 'If it is proposed to consider the chance of exceeding the given values of *t*, in a positive (or negative) direction only, then the values of P should be halved?' Has the table he provides, so we might wonder, also some other purpose? Subsequently (pp. 121-122) he analyses the results of an experiment with ten patients on the efficacy of two different supposedly soporific drugs in producing sleep. For Student's *t* from ten paired comparisons, he obtains

$$t_{obs} = 4.06 \text{ in favour of drug B over drug A,} \qquad (6.3.9)$$

and he declares that, on 9 degrees of freedom,

> 'only one value in a hundred will exceed 3.250 by chance, so that the difference between the results is clearly significant'. (6.3.10)

At first there might seem to be an error. He forgets, we might think, that V equals half of P; the correct co-ordinate equals one in *two* hundred. And when he then says, 'will exceed', he also forgets, we might think, to say '*in the right direction*' (as, we might recall, he *did* remember to say on p. 95). However, it then promptly appears that such is not at all the case, when Fisher proceeds to remark that by a sign test applied to the nine non-zero differences obtained from the selfsame ten paired comparisons in question,

> 'we should, in this case, have been led to the same conclusion with almost equal certainty; for if the two drugs had been equally effective, positive and negative signs would occur with equal frequency. Of the 9 values other than zero, however, all are positive, and it appears from the binomial distribution,
>
> $(\frac{1}{2}+\frac{1}{2})^9$,
>
> that all will be of the same sign, by chance, only twice in 512 trials.'

Moreover, this is immediately followed by a further remark telling us that the *t* test is here to be preferred to the sign test. This tells us that Fisher is explaining his principle of testing, rather than his choice of test statistic. So it will then dawn on us that Fisher did not forget that 'V equals half of P', or forget to ask whether or not the deviation is 'in the right direction', but that he deliberately used a $|t|$ ordering rather than a *t* ordering. The reader should note that our discussion leads to this point via a pedestrian route, because we wish to underscore that Fisher introduces two-tailed tests *without explanation*, both in *Statistical methods for research workers* and in *The design of experiments*. In the latter book, two-tailed tests are introduced on p. 38, and it is again precipitately done, without explanation. Yet both books otherwise provide careful and detailed explanations. We are thus compelled to recognise that the explanation for Fisher using $|t|$ as test statistic at (6.3.10) is the direct and obvious one of extending the behavioural reasoning at (6.3.4) in the following manner, where H_0 now denotes a normal population of differences with mean equal to zero:

Suppose that we were to accept t_{obs} as evidence against H_0. Then we would be bound to accept $-t_{obs}$ as equally strong evidence against H_0. So we would be bound to accept any larger value of $|t|_{obs}$ as even stronger evidence against H_0. Hence, the probability that we would mistakenly declare there to be evidence against H_0 were we to regard the data under analysis as just decisive against H_0, is given by

$$p_{obs} = \Pr(|T| \geq |t|_{obs} \,|\, H_0). \qquad (6.3.11)$$

That this correctly interprets Fisher's notion of 'a two-tailed test', and that he thereby introduced the notion of 'simultaneous statistical inference', is beyond any reasonable contest.

6.4 Proof of an inadvertent confounding

Often the index of a class of statistical models is a real-valued parameter whose range includes, and is bounded by, one of the values that parameter might take, for instance if θ represents the probability of success when trying to discriminate by taste between two items, $0.5 \leq \theta$. Again, a mineral supplement in a feed ration for dairy cows might be harmless or beneficial, so if δ represents its effect on milk production, $0 \leq \delta$. Once again, if μ represents the number of foetuses borne by a pregnant rabbit, $1 \leq \mu$. In this section we use such an example to prove that R. A. Fisher's idea of a significance test involves an inadvertent confounding of two incompatibly different concepts. Consider for that purpose, the counts given in Table 1.15.1. Recall that the angular transform in degrees, of a binomial count out of $\underline{n}$, has variance

$820.7 \div n$, where $n = 5$ syllables per response in Table 1.15.1.

However, counts such as those in Table 1.15.1 often involve some variation in excess of the binomial. So, consider, as a class of models for the error sum of squares arising from the standard analysis of variance of their angular transforms

$[(820.7 \div 5) + \sigma^2] \times [\chi^2 \text{ on } 14 \text{ degrees of freedom}]$ where $0 \leq \sigma^2$.

Let us ask:

What are the σ^2 values for which the given data might reasonably be modelled?

The observed value of the error sum of squares for the present example is 3 014.4. So, for a co-ordination test of say $\sigma^2 = 0.04$ we find

$3.014.4 \div [(820.7 \div 5) + 0.04] = 18.36$, where $\Pr(\chi^2 \text{ on } 14 \text{ df} \geq 18.36) = 0.2028$,

showing that the mental correlate of the datum chi-square is then situated at

$(0.7972, \varepsilon, 0.2028)$ in the chi-square test distribution.

By testing different hypothesised values in this way we obtain the following array, where the co-ordinates are rounded to the nearest 0.00005 for convenient reading:

Hypothesised value	Statistical co-ordinate	
$\sigma^2 = 0.03$	(0.79725, ε, 0.20275)	
$\sigma^2 = 0.02$	(0.79730, ε, 0.20270)	
$\sigma^2 = 0.01$	(0.79735, ε, 0.20265)	
$\sigma^2 = 0$	(0.79740, ε, 0.20260)	(6.4.1)

We must now consider a corresponding array of significance tests, for which purpose we must carefully note that the form of our question is: 'How much might the value of σ^2 be?', not 'At least how much ...?' or 'At most how much ...?' When a substantive investigator asks us 'How much ...?', then *that* is the question we must try to answer. So if our hypothesised model is say $\sigma^2 = 0.04$, alternatives both of the form $\sigma^2 > 0.04$ and of the form $\sigma^2 < 0.04$ must be considered. Recalling that the value of the observed error sum of squares is 3 014.4, the relevant facts are now

> 3 014.4÷[(820.7÷5)+0.04] = 18.36, where $\Pr(\chi^2 \text{ on } 14 \text{ df} \geq 18.36) = 0.2028$,
> and where 0.2028 is also the value of $\Pr(\chi^2 \text{ on } 14 \text{ df} \leq 9.42)$.

Reasoning as at (6.3.11) then leads to a two-tailed significance test, as follows:

> Suppose that we were to accept chi-square$_{obs}$ = 18.36 as evidence against $\sigma^2 = 0.04$, and in favour of $\sigma^2 > 0.04$. Then we would be bound to accept chi-square$_{obs}$ = 9.42 as equally strong evidence against $\sigma^2 = 0.04$, and in favour of $\sigma^2 < 0.04$. And so we would be bound to accept any value of chi-square$_{obs}$ that is more than 18.36, or less than 9.42, as even stronger evidence against $\sigma^2 = 0.04$, and in favour of $\sigma^2 > 0.04$ or $\sigma^2 < 0.04$, respectively. Hence, the probability that we would mistakenly declare there to be evidence against $\sigma^2 = 0.04$, were we to regard the data under analysis as just decisive against $\sigma^2 = 0.04$, is given by

$$\begin{aligned} p_{obs} &= \Pr(\chi^2 \geq 18.36 \,|\, \sigma^2 = 0.04) + \Pr(\chi^2 \leq 9.42 \,|\, \sigma^2 = 0.04) \\ &= 0.2028+0.2028 \\ &= 0.4056. \end{aligned}$$

By reasoning in this way we can obtain the significance level of the evidence against any non-zero hypothesised value of σ^2. When the hypothesised value is zero however, we are forced to modify the reasoning in such a way that a one-sided test is obtained, otherwise the reasoning would not make sense. We thus obtain the following array:

Hypothesised value	Significance level	
$\sigma^2 = 0.03$	$p_{obs} = 0.40550$	
$\sigma^2 = 0.02$	$p_{obs} = 0.40540$	
$\sigma^2 = 0.01$	$p_{obs} = 0.40530$	
$\sigma^2 = 0.0000000001$	$p_{obs} = 0.40520$	
$\sigma^2 = 0$	$p_{obs} = 0.20260$	(6.4.2)

The discontinuity that separates the last term from its immediate predecessor at (6.4.2) has no counterpart at (6.4.1), as the discontinuity has derived from the introduction of an additional epistemological concept above and beyond the essential epistemological concept used at (6.4.1). The concept used at (6.4.1) is just that of pointing and saying:

> 'See for yourself how snugly within the hypothesised crowd (or awkwardly upon its outskirts) the mental correlate of the test datum is being situated.' (6.4.3)

The additional concept is then introduced at (6.4.2) by reasoning of the kind:

> 'Hence, the probability that we would mistakenly declare there to be evidence against the hypothesised model, were we to regard the data under analysis as just decisive against that model, is ... ' (6.4.4)

It is well worth noting that '*See for yourself ...*' is *predicative*, whereas '*Hence, the probability ...*' is *predictive*. So unquestionably, two distinctly different concepts are involved. Recall also, as was explained in the previous chapter, why the explanations at (6.3.4) and (6.3.5) carefully avoid making *a forecast*. Significance testing, just like co-ordination testing, concerns problems in *pointing*, and not problems in *forecasting* (*hypothesis* testing concerns problems in forecasting.)

The arrays at (6.4.1) and (6.4.2) strongly disagree about the solution to a given epistemological problem. This is so despite broad agreement on the nature of the problem itself. In order to come to grips with the disagreement, we must proceed from that broad agreement. There is broad agreement that we are pointing at a solitary data set in the real world, and trying to eliminate some of the members of a matching class of models in the human mind. It is agreed that those models are predictive in the sense of providing for the prediction of physical (bodily) experiences. It is agreed that the models are to be tested by testing predicted experiences against the corresponding experiential data. It is agreed that the outcomes of such tests must be capable of being forced upon the human body, if needs be by simulation. This brings us to the crux of the disagreement when we ask: 'What might one then need to simulate?' Clearly, all that might need simulation is the array at (6.4.1), as the array at (6.4.2) is then simply derived from the initial one at (6.4.1) by the introduction of a further concept. Equally clearly, that further concept is not needed, as the question to be answered has already been answered at (6.4.1) in terms of that which was agreed upon. And equally clearly, the further concept is incompatible with the initial concept, as its introduction results in *widely* different descriptions of the quality of fit of the two *infinitesimally* different models indexed by

$$\sigma^2 = 0.0000000001 \text{ and } \sigma^2 = 0, \tag{6.4.5}$$

respectively. It must be grasped firmly that for all practical purposes the two models indexed at (6.4.5) fit the given data *equally* well, *and that must be shown to be so by any reasonable procedure*. It remains only to note that there is no merit in replacing the array at (6.4.2) with a 'conservative' array based on two-tailed tests only. This is because there is no merit in replacing a given test with a less sensitive test – especially if that is being done in order to hide an unwelcome fact.

6.5 Another proof

The main thrust of the proof given in the previous section is an incongruity that arises in consequence of the inadvertent confounding. For a proof where the confounding and its immediate consequences are displayed directly at source, consider all the 2.×2

contingency tables whose 1st and 2nd row totals, and whose 1st and 2nd column totals, are fixed at 8 and 5, and at 6 and 7, respectively. Consider M(0): 'No association' as hypothesised model, and let O_T denote the ordering

$$\begin{bmatrix} T & 8\text{-}T \\ 6\text{-}T & T\text{-}1 \end{bmatrix} = \begin{bmatrix} 1 & 7 \\ 5 & 0 \end{bmatrix}, \begin{bmatrix} 2 & 6 \\ 4 & 1 \end{bmatrix}, \begin{bmatrix} 3 & 5 \\ 3 & 2 \end{bmatrix}, \begin{bmatrix} 4 & 4 \\ 2 & 3 \end{bmatrix}, \begin{bmatrix} 5 & 3 \\ 1 & 4 \end{bmatrix}, \begin{bmatrix} 6 & 2 \\ 0 & 5 \end{bmatrix}. \tag{6.5.1}$$

The recipe at (1.30.3) then shows that, under the hypothesised model, the probabilities of the ordered configurations at (6.5.1), conditional on the row and column totals, are

$$\frac{2}{429}, \frac{35}{429}, \frac{140}{429}, \frac{175}{429}, \frac{70}{429}, \frac{7}{429}, \tag{6.5.2}$$

respectively. Suppose now that each one of the two alternatives:

M(-): 'Negative association' and M(+): 'Positive association'

is, in its own right, substantively conceivable. Then a co-ordination test based on O_T is left-sensitive to M(-) and right-sensitive to M(+). Under this ordering the situation of, for instance, $T = 2$ in the test distribution, is given by

$$\left[\frac{2}{429}, \frac{35}{429}, \frac{140}{429}\right] = (0.01, 0.08, 0.91),$$

pointing weakly toward the left at M(–), (6.5.3)

and the situation of for instance $T = 6$ in the test distribution is given by

$$\left[\frac{422}{429}, \frac{17}{429}, \frac{0}{429}\right] = (0.98, 0.02, \varnothing),$$

pointing strongly toward the right at M(+). (6.5.4)

The problem Fisher wants to solve must now be grasped firmly. In Chapter 4 we saw that when investigation is mistaken for decision-making under risk, circular reasoning cannot be avoided. In Chapter 5 we saw how Fisher avoids such circular reasoning by using realised significance levels, not as forecasted frequencies, but as measurements of quality of fit. So the problem that significance testing proposes to solve, is that of appropriately measuring the quality of fit of alternative models brought to mind, so as to judge their tenability when considered as alternative explanations for how a given real-world data set might (or might not) have come about. In addition of course, it is required that those measurements be provided in terms that are physically meaningful, i.e. terms that can, if needs be, be explained by simulation. In short, we must devise an instrument that can physically measure the tenability of our three alternative models in respect of a given real-world data set. This is directly comparable to, for instance, devising an instrument that can physically measure the tenability of three alternative models called 'low', 'normal', or 'high', in respect of a real-world patient's blood pressure – that being the case because 'physical' here simply means 'can be grasped by the human body'. True, the patient is directly present, whereas the 2×2 table represents something not directly present. However, that need not concern us here, as we need not distinguish here between measurements directly made and measurements derived from the historical

record, that is to say, from previous measurements directly made. Also true, statistical co-ordinates are explicitly imprecise, whereas one measurement of blood pressure is only implicitly imprecise, as repeated measurement of such blood pressure would show. However, that is beside the point here, as our analogy does not concern the *particular* physical meanings involved but only the *physicality* of those meanings.

Co-ordination tests provide, in the required sense, a solution to the foregoing problem. In order to grasp this the reader should note that we have brought into mind an incipient class comprising just three models, {M(-), M(0), M(+)}. Though M(0) is the only mathematically explicit member of that class, the other two members are sufficiently well defined for us to grasp how, in respect of given data, the ordering at (6.5.1) might well separate the different causative explanations we associate with the different members; in cases of strong association, as follows:

> For *T* values arising from negative association, the human body can, by simulation, be physically forced to grasp that the mental correlates of such values tend mostly to be situated snugly within the M(-) crowd, but awkwardly so upon the left-hand outskirts of the other two.
>
> For *T* values arising from no association, the human body can, by simulation, be physically forced to grasp that the mental correlates of such values tend mostly to be situated snugly within the M(0) crowd, but awkwardly so upon the right-hand outskirts of the M(-) crowd, and upon the left-hand outskirts of the M(+) crowd.
>
> For *T* values arising from positive association, the human body can, by simulation, be physically forced to grasp that the mental correlates of such values tend mostly to be situated snugly within the M(+) crowd, but awkwardly so upon the right-hand outskirts of the other two. (6.5.5)

Our co-ordination test thus clearly provides a solution to the given problem. We note in passing that, for that problem, it does not seem possible that a better co-ordination test could be devised. This is, however, irrelevant at present, as the present development concerns our principle of testing, not our choice of test statistic. What *will* prove to be relevant is that the purpose of the ordering at (6.5.1) is *analytical* in the sense of trying to distinguish the three different models by separating, as far as possible, the sample patterns that point at this, that or the other respective model. This concludes the first of three developments that together comprise the present proof.

The foregoing development involved the concept of *measuring quality of fit*, but not that of *the probability of mistaken conclusion*. This must be grasped firmly. So, let us note that a co-ordination tester may grant, without involving any behavioural concepts, that it is convenient, though slightly imprecise, to summarise the statements made at (6.5.3) and (6.5.4) by saying:

> If $T = 2$, then $U+\varepsilon = 0.09$, pointing weakly at M(-). (6.5.6)
>
> If $T = 6$, then $\varepsilon+V = 0.02$, pointing strongly at M(+). (6.5.7)

And let us note further that a co-ordination tester may also grant, without involving any behavioural concepts, that it is correct, though irrelevant for the purposes of co-ordination tests, to say that it follows from the statements at (6.5.6) and (6.5.7) that:

> The probability under M(0) that $T = t$ points as strongly, or more strongly, at M(-) than $T = 2$ does, is given by U+ε = 0.09.
>
> The probability under M(0) that $T = t$ points as strongly, or more strongly, at M(+) than $T = 6$ does, is given by ε+V = 0.02.

The point here is that a co-ordination test involves no behaviourism whatsoever. This is important because, by way of contrast, significance testing relies, for its physical meaning, on the behavioural concept explained in (6.3.4), which entails, as explained in (6.3.11), the further notion of 'simultaneous statistical inference' as an inescapable logical consequence. Cox and Hinkley (1974, p. 77) have tried to motivate that notion by arguing that one must make allowance for behaviour amounting to

> 'selection of [the] test in the light of [the] data'. (6.5.9)

This leads them to remark (p. 79) that the recipe at (6.3.11) is available when

> 't and $-t$ represent essentially equivalent departures from H_0,' but that (6.5.10)
>
> 'commonly, however, large and small values of t indicate quite different kinds of departure and, further, there may be no very natural way of specifying what are equally important departures in the two directions'. (6.5.11)

The only immediate relevance of these remarks is that many orderings are incapable of adequately precise symmetric representation. So they argue that

> 'it is best to regard the tests in the two different directions as two different tests, both of which are being used'. (6.5.12)

(See also Cox 1984.) In the case of the present example the two different tests they refer to are obtained by considering the ordering at (6.5.1) as being right-sensitive to M(+), and the inverse of that ordering also as being right-sensitive, but to M(-), not M(+). The test statistic for the inverse ordering is say $7-T = S$. The two orderings are then as follows:

> O_T is right-sensitive to M(+), the hypothesised probabilities being given by
>
> $\frac{2}{429}, \frac{35}{429}, \frac{140}{429}, \frac{175}{429}, \frac{70}{429}, \frac{7}{429}$ for $T = 1, 2, 3, 4, 5, 6$, respectively.
>
> O_S is right-sensitive to M(–), the hypothesised probabilities being given by
>
> $\frac{7}{429}, \frac{70}{429}, \frac{175}{429}, \frac{140}{429}, \frac{35}{429}, \frac{2}{429}$, for $S = 1, 2, 3, 4, 5, 6$, respectively.

The behavioural concept with its *leitmotiv* of this, that or the other possibility of 'a mistaken conclusion', here in respect of M(0), together with its entailed concept of 'simultaneous

statistical inference', here in respect of M(-) and M(+) simultaneously, is then introduced by the following kind of reasoning where the phrase 'then we would be bound' means 'then we would be bound by way of a smaller probability of mistaken conclusion', and where the word 'hence' announces the entailment:

> Suppose we were to consider $S = 6$ as just decisive against M(0) in favour of M(-). Then we would be bound to consider that no value of T is decisive against M(0) in favour of M(+). Hence, the probability that we would mistakenly declare there to be evidence against M(0) in favour of either M(-) or M(+) is $p_{obs} = (2+0)\div 429$.
>
> Suppose we were to consider $T = 6$ as just decisive against M(0) in favour of M(+). Then we would be bound to consider $S = 6$ as even more decisive against M(0) in favour of M(-). Hence the probability that we would mistakenly declare there to be evidence against M(0) in favour of either M(-) or M(+) is $p_{obs} = (7+2+0)\div 429$.
>
> Suppose we were to consider $S = 5$ as just decisive against M(0) in favour of M(-). Then we would be bound to consider $T = 6$ as even more decisive against M(0) in favour of M(+), and bound to consider $S = 6$ as still more decisive against M(0) in favour of M(-). Hence the probability that we would mistakenly declare there to be evidence against M(0) in favour of M(-) or M(+) is $p_{obs} = (35+7+2+0)\div 429$.
>
> And so on. (6.5.13)

Cox and Hinkley (1974, p.79) note that such probabilities of 'mistaken' behaviour can as a general rule be obtained as the attainable significance levels when the test statistic is taken to be the sum of the original rounding and corresponding lesser co-ordinate. In our example, for instance, the values of the sum of the rounding and corresponding lesser co-ordinate for $T = 1, 2, 3, 4, 5, 6$, are seen from the hypothesised probabilities at (6.5.2) to be given by:

$$\frac{0+2}{429}, \frac{0+2+35}{429}, \frac{0+2+35+140}{429}, \frac{175+70+7+0}{429}, \frac{70+7+0}{429}, \frac{7+0}{429}, \text{ amounting to}$$

$$\frac{2}{429}, \frac{44}{429}, \frac{254}{429}, \frac{429}{429}, \frac{114}{429}, \frac{9}{429}, \text{ respectively.}$$

So, by re-ordering on the magnitude of these probabilities, the original ordering, is replaced by a re-ordering, , given by:

T'= 1, 2, 3, 4, 5, 6, for T = 4, 3, 5, 2, 6, 1, respectively. (6.5.14)

Table 6.5.1 shows how the behavioural probabilities arising at (6.5.13) then arise as the set of attainable significance levels for the re-ordering at (6.5.14). This completes

Table 6.5.1: The attainable significance levels of a two-tailed significance test of association in a 2 × 2 contingency table whose 1st and 2nd row totals, and 1st and 2nd column totals are fixed at 8 and 5, and fixed at 6 and 7, respectively

$$\Pr[T' \geq 6 \,|\mathrm{M}(0)] = \Pr(T = 1 \,|\mathrm{M}(0)] = \frac{2}{429}$$

$$\Pr[T' \geq 5 \,|\mathrm{M}(0)] = \Pr(T = 1 \text{ or } 6 \,|\mathrm{M}(0)] = \frac{2+7}{429}$$

$$\Pr[T' \geq 4 \,|\mathrm{M}(0)] = \Pr(T = 1 \text{ or } 6 \text{ or } 2 \,|\mathrm{M}(0)] = \frac{2+7+35}{429}$$

$$\Pr[T' \geq 3 \,|\mathrm{M}(0)] = \Pr(T = 1 \text{ or } 6 \text{ or } 2 \text{ or } 5 \,|\mathrm{M}(0)] = \frac{2+7+35+70}{429}$$

$$\Pr[T' \geq 2 \,|\mathrm{M}(0)] = \Pr(T = 1 \text{ or } 6 \text{ or } 2 \text{ or } 5 \text{ or } 3 \,|\mathrm{M}(0)] = \frac{2+7+35+70+140}{429}$$

$$\Pr[T' \geq 1 \,|\mathrm{M}(0)] = \Pr(T = 1 \text{ or } 6 \text{ or } 2 \text{ or } 5 \text{ or } 3 \text{ or } 4 \,|\mathrm{M}(0)] = 1$$

the second part of our proof. It shows how the significance test arises by replacing an initial ordering based on non-behavioural concepts of 'ordering according to fit', with an entirely different ordering based on behavioural concepts of 'ordering according to hypothesised probabilities of mistaken conclusion'.

Note that the first part of our proof now implies that the behavioural concept is unnecessary. This is the case because, as we explained at (6.5.5), the problem can be solved without its introduction. That brings us to the third part of our proof, which is simply to note that the introduction of the behavioural concept is not merely unnecessary, but has in fact destructive consequences when measuring quality of fit. The present case exemplifies such destruction inasmuch as the significance level decreases as T' increases, but by way of a scrambled ordering with

T' = 1 favouring M(+) as having the better quality of fit,
T' = 2 favouring M(-) as having the better quality of fit,
T' = 3 favouring M(+) as having the better quality of fit,
T' = 4 favouring M(-) as having the better quality of fit,
T' = 5 favouring M(+) as having the better quality of fit, and
T' = 6 favouring M(-) as having the better quality of fit. (6.5.15)

The ordering at (6.5.15) is clearly not compatible with the notion of a discriminative ordering that tries, as far as possible, to separate and place into three different parts of the ordering, those data patterns that point at M(-), or M(0), or M(+), respectively. A co-ordination tester might in fact object, with good reason, to our having described the result at (6.5.14) as a re-ordering; might in fact call, with good reason, for it to be described as a 'disordering'. And, of course, when that disordering is viewed as an ordering for a co-ordination test, its separating characteristics can be shown by simulation to be inferior to those of the original ordering. That completes the present proof, and we note that, although it differs in detail from the proof developed in the previous section, the same underlying reasoning

is used. In each proof we begin with a concrete example requiring assessment of quality of fit in physically understandable terms. In each proof, we show how that is achieved directly by the use of statistical co-ordinates. And in each proof we show that the use of significance levels for the same problem in measuring fit then requires the introduction of a further concept, which is both unnecessary and destructive.

6.6 A watershed

We can now finally come to grips with the definitive distinction between significance tests and co-ordination tests. It will be helpful to consider the following situation. Suppose that a family physician is consulted by a patient complaining of dizziness, and that the physician makes a measurement, $X = x$, of the patient's blood pressure. We suppose that such measurements on many individuals have provided a scale on which, for normal blood pressure, X is a $N(0, 1^2)$ variable, and for abnormal blood pressure X is a $N(\theta, 1^2)$ variable, with $\theta < 0$ (low blood pressure) or with $\theta > 0$ (high blood pressure). We now consider this situation in terms of four different scenarios.

Scenario 1

The physician will be concerned about the possibility of high blood pressure, as that is a serious condition. So, supposing that the physician practises hypothesis testing, a one-sided hypothesis test of H_0: $\theta = 0$ versus H_1: $\theta > 0$ is specified *without reference to the data*. The physician will be concerned by the possibility that slight tendencies might be overlooked. In order to rather be safe than sorry, a fairly high Type I error rate, say $\alpha = 0.10$, is specified *without reference to the data*. So the critical region for the test is

$$1.28 \leq x < \infty.$$

Let the observed blood pressure be $X = -1.55$. The physician will then be embarrassed by perhaps having, as Kendall and Stuart (1961, p. 182) express it, 'located the critical region in the tail of the statistic's distribution which turned out to be the wrong one for the true value of θ'. 'How silly of me,' the physician might think, 'for not having considered the possibility of low blood pressure. After all, dizziness is a symptom of low blood pressure. Yet, I will now have to pretend that I have not the foggiest notion of what causes my patient's complaint, whereas in fact the observed X value gives me a clear indication of what the cause may be.'

Scenario 2

The physician will be concerned about the possibility of high blood pressure, as that is a serious condition, but might also consider dizziness as being more indicative of low blood pressure. So, suppose that the physician practises hypothesis testing and is also mindful of a recommendation by Kendall and Stuart (1961, pp. 182, 201, 202) that, as 'a common sense way of insuring against' the possibility of certain embarrassing outcomes, an 'unbiased' hypothesis test is preferred. So a two-sided hypothesis test of H_0: $\theta = 0$ versus H_1: $\theta \neq 0$ with the 'unbiased' apportionment of half the Type I error rate in each tail is specified *without reference to the data*. The physician will be concerned by the possibility that slight tendencies might be overlooked. So as rather to be safe than sorry, a fairly high

Type I error rate, say $\alpha = 0.10$, is specified *without reference to the data*. Hence the critical region for the test is:

the union of $-\infty < x \leq -1.645$ and $+1.645 \leq x < +\infty$.

Recall that the observed blood pressure turned out to be $X = -1.55$. So, having sought protection against the embarrassment of an extreme value in the 'wrong' tail area, the physician will instead suffer the embarrassment of having been overly cautious. 'How silly of me,' the physician might think, 'for not having used a "biased" apportionment of say 0.75 of the Type I error rate in the lower tail area and then only 0.25 thereof in the upper tail area. After all, low blood pressure was in the first place indicated by the patient's complaint.' The critical region would then have been:

the union of $-\infty < x \leq -1.44$ and $+1.96 \leq x < +\infty$.

So, low blood pressure would have been diagnosed. Yet, as it is, I will now have to pretend that I have not the foggiest notion of what causes my patient's complaint, whereas in fact the observed X value gives me a clear indication of what that cause may be'.

Discussion of Scenarios 1 and 2

In both scenarios the physician has mistaken a problem in investigation for a problem in decision-making under risk. So, in both scenarios the physician begins by reasoning as if the patient is just one of a host of patients arriving in the real world. In both scenarios the physician then reasons (correctly in terms of that mistake) how, in such a case, there can be adjoined to each patient, a decision such that amongst those who are healthy the decision will be erroneous in a specified proportion of cases, and such that amongst those who are ill in a specified sense the decision will be correct in a larger proportion of cases. In each scenario the reasoning leads to embarrassment, not for being incorrect, but for being mistaken. As the problem is clearly an investigative one, the patient must be viewed as a *single, solitary* real-world individual, and not as one of a *host* of real-world individuals. So the population should, as an *explanatory* host, be brought into *the human mind*, and not, as a *fabricated* host, be brought into *the real world*. R. A. Fisher, as quoted at (6.3.1), clearly understood that, and we take account of that understanding in the next scenario.

Scenario 3

The physician will be concerned about the possibility of high blood pressure, as that is a serious condition, but should also consider dizziness as being more indicative of low blood pressure. So, supposing that the physician practises significance testing, a two-sided significance test of M_0: $\theta = 0$ versus M_1: $\theta \neq 0$ is indicated. The theory of significance testing would then have the physician reason that if $X = -1.55$ were to be regarded as just decisive against $\theta = 0$ and in favour of $\theta < 0$, then $X = +1.55$ would have to be regarded as just decisive against $\theta = 0$ in favour of $\theta > 0$. Therefore, the probability that the physician would mistakenly declare there to be evidence against M_0, were the physician to regard the given data as just decisive against M_0, is given by:

$$\begin{aligned} p_{obs} &= Pr(X \leq -1.55 \mid M_0) + Pr(X \geq +1.55 \mid M_0) \\ &= 0.06 + 0.06 \\ &= 0.12 \end{aligned}$$

As the physician will be concerned by the possibility that slight tendencies might be overlooked, the significance level of $p_{obs} = 0.12$ *coupled with the substantive fact that dizziness is a symptom of low blood pressure*, might well lead the physician to advise certain dietary precautions and to arrange a subsequent appointment, so as to keep the matter under observation. Note that this is very much *with* reference to *all* of the data, both the datum of pressure and the datum of dizziness. Note also that although the physician might act upon the given facts, there is no prescriptive rule that might impose an embarrassing decision.

Discussion of Scenarios 1, 2 and 3

Scenario 3 makes better sense than do Scenarios 1 and 2. However, here that is not the purpose for introducing the three scenarios. Here the purpose is to attract attention to something they have in common, namely if each of them is trying to explain *how these data might, or might not, have come about*, which is what they *should* be doing, why does each of them give an 'explanation' that, as it were, *partakes of a statistician*? In Scenarios 1 and 2, that partaking of the statistician is made glaringly obvious by the notion of specifications to be made *without reference to the data*. So we are offered an explanation involving a *specified* α, and a *specified* one-sided test, or a *specified* unbiased two-side test, or a *specified* biased two-sided test, etc., and these specifications are to be made *without reference to the data*. Each explanation thus involves *a statistician* in the guise of a physician who is unable to avoid embarrassing consequences that might arise from those specifications. The point here is that the purpose of data analysis is to provide an explanation of how given data might have come about. In the present case, those data are an experienced dizziness and a measurement of blood pressure, which data are not properly 'explained' by reasoning that partakes of a statistician who had absolutely nothing to do with the origin of those data. In Scenario 3 the specifications are avoided by replacing the notion of a specified Type I error rate with the notion of a significance level as measuring fit. However, as explained at (6.3.11), that measurement is by way of simultaneous statistical inference given *a meaning that partakes of the behaviour of a hypothetical statistician*. So the reasoning again leads to *a model that contains a statistician*! That is the case despite the fact that no statistician was involved in that which caused the datum of dizziness, or in that which caused the blood pressure measurement to be low. The idea of such a model must be firmly rebuffed. No matter how pretty its mathematics, a model that partakes of irrelevancy is bad science and bound to cause confusion and error. So the hypothetical statistician must, figuratively speaking, be seized by the scruff of the neck – and by the seat of the pants – and frog-marched from model, leaving us with Scenario 4.

Scenario 4

The physician will be concerned about the possibility of high blood pressure, as that is a serious condition, but should also consider dizziness as more indicative of low blood pressure. So, supposing that the physician practises co-ordination testing, such a test of $M_0: \theta = 0$ versus $M_1: \theta < 0$ or $M_2: \theta > 0$ is indicated. Given $X = -1.55$ as the datum of measurement, it will be found that the mental correlate of that datum is situated at $(0.06, \varepsilon, 0.94)$ in the test distribution. That finding, *coupled with the substantive fact that dizziness is a symptom of low blood pressure*, will lead the physician to make certain dietary prescriptions and to arrange a subsequent appointment to keep the matter under observation. Note that this is very much *with* reference to *all* the data, both the datum of pressure and the datum of dizziness. Note

also that even though the physician acts on the given facts, there is no prescriptive rule that might impose an embarrassing decision – no rule that runs contrary to the physician's rational opinion. And note, finally, that the physician's reasoning proceeds from a model that does not partake of a statistician's specifications or behaviour, as such specifications or behaviour, having been adjoined *post hoc*, cannot be part of any defendable model of how the patient's symptoms might have come about.

We have arrived at a watershed on which the reader is now compelled to take a stance. We have seen that significance tests are not intended to produce real-world frequency forecasts. They are intended to produce formal probabilities, so-called significance levels that are calculated from hypothesised models in respect of given data so as to measure the quality of fit of those models when viewed as possible explanations of how those particular data might, or might not, have come about. An inescapable question now arises:

'Why measure "quality of fit" in that particular way?' (6.6.1)

Cox and Hinkley (1974, p. 209) reply that one requires an evidential concept whose 'physical interpretation' gives 'an empirical meaning, which in principle can be checked by experiment'. Cox (1977, p. 50) states further that the interpretation in question must be

'relevant to the use to be made of the test'. (6.6.2)

These explanations posit two distinctly different requirements:

Requirement 1: The concept must be physically meaningful.
Requirement 2: That meaning must be 'relevant to the use to be made of the test'.

Requirement 1 is unexceptionable and clearly met, as the meaning of 'the significance level' can in any instance of its use be forced upon the human body by a simulation of that meaning. The same, however, is also true of the concept 'statistical co-ordinates', which brings us to Requirement 2, where Cox's phrase 'of the test' refers of course to a significance test, but could equally well be made to refer to a co-ordination test. So, going back to the question at (6.6.1), we must consider, not just one, but two different ways to measure that quality of fit. We must also note that the outcome of the significance test can be derived from that of the co-ordination test by adjoining the *further* concept of the knowing subject who, apart from this possible mistake, might also make that possible mistake and therefore compels the significance tester to adjoin *yet a further* concept of simultaneous statistical inference. However, we have developed proof beyond any reasonable contest that the two further concepts are not needed for measuring quality of fit. By adjoining those concepts, significance testing violates a fundamental principle of science:

Never, ever introduce a constituent that is not needed.

6.7 Fisher's 'knowing subject'

In Section 1.1 we cautioned that development of co-ordination tests would ultimately lead to a devastating outcome. So, it would not have done to precipitately dismiss such tests, as one colleague inanely did with the comment that co-ordination tests 'are just

significance tests with frills' – inanely, as it is the other way round; it is significance tests that adjoin the unneeded embellishment of a 'knowing subject'. And it would not have done to precipitately dismiss such tests, as another colleague short-sightedly did with the comment that 'this is just a shift of emphasis' – short-sightedly, as that fails to grasp that inasmuch as our removal of the knowing subject destroys simultaneous statistical inference at source (at the two-tailed level), further destruction cannot be escaped. This must be firmly grasped. So, let us develop two examples of that further destruction.

Example 6.7.1

Some 50 years ago Barbara McClintock discovered the existence in maize of what she called a travelling gene. Such a gene has, so to speak, two 'houses' – shall we say a town house and a country house. A travelling gene in the vinegar fly leads to the following problem (Boussy and Kidwell 1987): in order to establish, in any particular case, in which 'house' the gene is currently 'staying', one determines the proportions of sterile offspring in each of two different test crosses. If we examine a number n of offspring in each cross, we obtain a pair of proportions (p, q) that might be modelled satisfactorily as realisations of independent binomial random variables with estimated variances

$$s_p^2 = p(1\text{-}p)\div n \text{ and } s_q^2 = q(1\text{-}q)\div n,$$

respectively. The proportion pair (p, q) estimates the location of a point in one or the other of the two rectangular regions labelled P and Q in Figure 6.7.1. That point is not

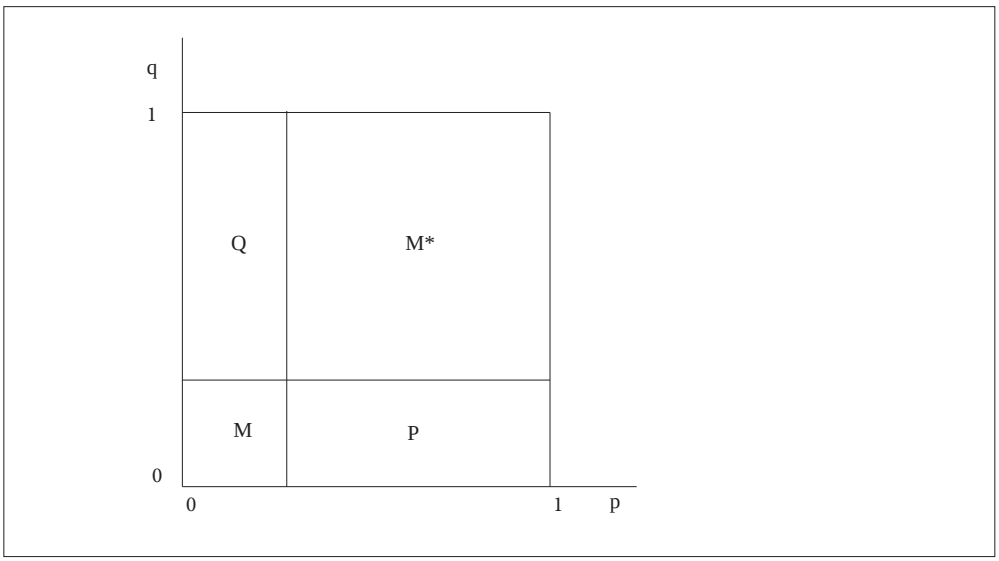


Figure 6.7.1: Possible locations of a point defined by a pair of binomial proportions

expected to be situated in either one of the two square regions labelled M and M*; but its estimated location (p, q) does sometimes stray into the region labelled M. So, we need a statistical method to judge such and other doubtful cases. For that purpose, we first develop a significance-test procedure, and then a co-ordination test procedure, as follows.

The significance test procedure for Example 6.7.1

Kempthorne and Folks (1971) develop what they call consonance (not confidence) regions, the general idea being to avoid a specified Type I error rate (or what amounts to the same thing, a specified confidence coefficient) allowing instead for significance levels that range over a number of different values of interest. Thus, for instance, we might use the standard normal approximation for the binomial to respectively obtain (1-0.10) = 0.90 and (1-0.05) = 0.95 upper bounding, one-sided consonance regions for E(p) in the travelling gene problem. Those regions are given by:

$$0 < E(p) < p+1.264\, s_p \text{ and } 0 < E(p) < p+1.645\, s_p, \text{ respectively.} \qquad (6.7.1)$$

Here, the knowing subject of significance testing knows that the probabilities of a mistaken conclusion would be 10% and 5%, respectively. These would then also be the probabilities for the lower bounding, one-sided consonance regions given by

$$p-1.264\, s_p < E(p) < 1 \text{ and } p-1.645\, s_p < E(p) < 1, \text{ respectively.} \qquad (6.7.2)$$

However, in respect of *just one solitary data set*, it is impossible for any investigator, whether real or imaginary, to be mistaken as at (6.7.1) and *also* at (6.7.2). In order for *both* kinds of mistakes to be *simultaneously* relevant, those mistakes, whether real or imaginary, must refer to different data sets arising from possibly different E(p). And so, by way of very subtle slippage, we fall into reasoning that no longer tries to explain 'how the given data might have come about', but instead is trying to explain 'how often an investigator might be "mistaken" in repetitive investigations of many different sets of such data'. And so, in order for the rates of those possible mistakes to be maintained at 10% and 5%, the forms at (6.7.1) and (6.7.2) are to be replaced by

$$p-1.645\, s_p < E(p) < p+1.645\, s_p, \text{ and}$$
$$p-1.960\, s_p < E(p) < p+1.960\, s_p, \text{ respectively.} \qquad (6.7.3)$$

There is more to come.

Our problem is how an observed proportion pair (p, q) might enable knowledge of an unknown proportion pair [E(p), E(q)]. In order to maintain the 10% and 5% norms in respect of rectangular consonance regions of the form

$$[p-Z\, s_p < E(p) < p+Z\, s_p]\times[q-Z\, s_p < E(q) < q+Z\, s_p], \qquad (6.7.4)$$

we must take Z to be the α percentage point of the N(0, 1) distribution for

$$\alpha = 1-\sqrt{1-(0.10\div 2)} \text{ and } \alpha = 1-\sqrt{1-(0.05\div 2)}, \text{ i.e. } \alpha = 0.0253 \text{ and } \alpha = 0.0126,$$

respectively. The corresponding values of Z, instead of those at (6.7.3), are

$$Z = 1.955 \text{ and } Z = 2.238, \text{ respectively.} \qquad (6.7.5)$$

The co-ordination test procedure for Example 6.7.1

The normal approximation used for the significance test procedure can also be used to obtain abbreviated traces for E(p) and E(q). For E(p) we find that

(0.10, ε, 0.90) and (0.90, ε, 0.10) are attained at
E(p) = p-1.264 s_p and p+1.264 s_p, respectively, and

(0.05, ε, 0.95) and (0.95, ε, 0.05) are attained at
E(p) = p-1.645 s_p and p+1.645 s_p, respectively.

For E(q) we obtain the same expressions but with q instead of p. The results can be expressed in terms of rectangular regions of the same *form* as that given at (6.7.4), but the meaning is of course a very different one, and instead of the Z values at (6.7.5) we now have, for *precisely the same physical norms,* 10% *and* 5%,

Z = 1.264 and Z = 1.645, respectively. (6.7.6)

Discussion of Example 6.7.1

There is a vast difference in the size of the regions arising from the formal expression at (6.7.4) when values, for precisely the same physical norm, are inserted from (6.7.5) or (6.7.6), respectively, the difference in area being of the order

$(2 \times 1.955)^2 = 15.3$ versus $(2 \times 1.264)^2 = 6.4$ when the norm is 10%, and
$(2 \times 2.238)^2 = 20.3$ versus $(2 \times 1.645)^2 = 10.8$ when the norm is 5%.

This of course reflects vastly different meanings. However, there is one meaning that is *absolutely identical* in the two cases, namely the meaning of a given physical norm. This must be firmly grasped. So, consider a significance tester who must simulate the physical meaning of the 10% norm for significance testing, and a co-ordination tester who must simulate the physical meaning of the 10% norm for co-ordination testing.

> The significance tester will challenge us to set up 1 000 *different population pairs* for as many different [E(p), E(q)] value pairs as we wish, will draw a sample pair from each population pair, calculate the 90% consonance region in each case, check whether or not the value of [E(p), E(p)] is covered, deposit a red chip in one bowl for each failure, a white chip in another bowl for each success, and then point *at the two bowls* for us to see for ourselves that about 10% of the deposits are red.
>
> The co-ordination tester will challenge us to set up *just a single population pair* for any [E(p), E(q)] value pair we wish, and to indicate any pointing co-ordinate we wish from amongst the four possibilities (two left and two right), will then draw a 1 000 sample pairs from the single population pair we set up, calculate the value of the indicated pointing co-ordinate in each case, check whether or not that co-ordinate exceeds 10%, deposit a red chip in one bowl for each one that does, a white chip in another bowl for each one that does not, and then point *at the two bowls* for us to see for ourselves that about 10% of the deposits are red. (6.7.7)

Each of these two simulations is entirely correct for what it explains. But they explain different reasoning. This is underscored by considering, how the two kinds of testers might express that difference, as follows:

Significance tester: When the coverage probability of your rectangle should be 90%, it is merely

$$1-[2(0.10)+2(0.10)-2(0.10)2(0.10)] = 0.64, \text{ i.e. } 64\%,$$

and when it is should be 95%, it is merely

$$1-[2(0.05)+2(0.05)-2(0.05)2(0.05)] = 0.81, \text{ i.e. } 81\%.$$

Co-ordination tester: By calculating the overall probability of possible mistakes on the part of your knowing subject, you blunt the instruments of investigation. For each of four instruments, your rectangle says the pointing co-ordinate equals 10%, when in fact it equals

$$\Pr(Z > 1.955) = 0.025, \text{ i.e. } 2.5\%,$$

and says it equals 5%, when in fact it equals

$$\Pr(Z > 2.238) = 0.013, \text{ i.e. } 1.3\%. \qquad (6.7.8)$$

The simulation procedures at (6.7.7) and the corresponding calculations at (6.7.8) are each in its own right correct for the reasoning it explains; but the reasoning cannot be appropriate in both cases. So we have to ask which of the two is appropriate. The question cannot be answered by mathematical statistics as such; it must be answered by substantive science. And, given substantive science that has escaped indoctrination by present-day literature on statistical methods for substantive investigation, there can surely be but one answer, as follows: any instrument of investigation that would have us confound data patterns pointing at large values of E(p) with data patterns pointing at small values of E(p), and have us confound data patterns pointing at large values of E(q) with data patterns pointing at small values of E(q), and further, for good measure, have us confound all four of the different kinds of patterns with each other, cannot be appropriate for our problem, because E(p) cannot at once be both large and small, and E(q) cannot at once be both large and small, not to mention that the pair [E(p), E(q)] cannot all and at once be [large, large], [small, large], [large, small], and [small, small].

Example 6.7.2

We revisit the problem of trying to establish the location of the earth's magnetic pole. We use the same class of models as in Section 4.28, first to develop a significance test procedure, and then to develop a co-ordination test procedure.

Significance test procedure for Example 6.7.2

Recall that if (μ_1, μ_2) denotes the position of the pole in two-dimensional Cartesian space, and our data are a number, n, of (X_1, X_2)-like measurements of that position, we suppose that our data can be modelled satisfactorily via a between-within analysis of variance as a realisation of three independent random variables, as follows.

a $N(\mu_1, \sigma^2 \div n)$ variable $\overline{X}_1$, which is the mean X_1-like measurement,
a $N(\mu_2, \sigma^2 \div n)$ variable $\overline{X}_2$, which is the mean X_2-like measurement, and
a $\sigma^2\chi^2 \div 2(n\text{-}1)$ variable S^2, which is the pooled error variance on 2(n-1) df.

Reasoning similar to that used in Sections 3.11 and 4.28 here leads to a nested array of (1-*P*) consonance regions for (μ_1, μ_2) centred at

$$(\mu_1, \mu_2) = (\overline{X}_1, \overline{X}_2) \text{ with radii } \sqrt{2(S^2 \div n)F(P)}, \text{ for } 0 < P < 1,$$

where $F(P)$ denotes the value that is exceeded with probability P by Snedecor's F on 2 and 2(n-1) degrees of freedom. Let n = 10. Then, choosing as norms $P = 0.10$ and $P = 0.05$, the radii of the corresponding circular consonance regions are given by

$$\sqrt{2(S^2 \div 10)2.62} \text{ and } \sqrt{2(S^2 \div 10)3.55}, \text{ that is to say,}$$
0.724S and 0.843S for the 10% and 5% norms, respectively. (6.7.9)

Co-ordination test procedure for Example 6.7.2

Consider an arbitrary degree of freedom of the form

$$T = \lambda_1\overline{X}_1 + \lambda_2\overline{X}_2, \text{ where } \lambda_1^2 + \lambda_2^2 = 1 \text{ and } E(T) = \lambda_1\mu_1 + \lambda_2\mu_2.$$

Then (T, S^2) is minimally sufficient for E(T). Consider, as elimination pivot

$$t = [T\text{-}E(T)] \div \sqrt{S^2 \div n}, \text{ distributed as Student's } t \text{ on } 2(n\text{-}1) \text{ df.} \quad (6.7.10)$$

With n = 10 as in the previous procedure, an abbreviated trace for E(T) is given by:

$T \pm 1.330\sqrt{S^2 \div 10}$, when the pointing co-ordinate equals 0.10, and
$T \pm 1.734\sqrt{S^2 \div 10}$, when the pointing co-ordinate equals 0.05,

thus

T-0.421S and T+0.421S at (U, ε, V) = (0.10, ε, •) and (• , ε, 0.10), respectively, and
T-0.548S and T+0.548S at (U, ε, V) = (0.05, ε, •) and (• , ε, 0.05), respectively.

Note that

$$T\text{-}E(T) = \lambda_1(\overline{X}_1 - \mu_1) + \lambda_2(\overline{X}_2 - \mu_2).$$

So, for every choice of (λ_1, λ_2) such that $\lambda_1^2 + \lambda_2^2 = 1$, the test statistic, t, takes the value zero when

$$(\mu_1, \mu_2) = (\overline{X}_1, \overline{X}_2), \text{ in which case (U, ε, V) = (0.50, ε, 0.50).}$$

This tells us that the variation of the abbreviated trace for all the possible variations of

$$(\lambda_1, \lambda_2) \text{ such that } \lambda_1^2 + \lambda_2^2 = 1,$$

takes the form of two of concentric circles centred at $(\mu_1, \mu_2) = (\overline{X}_1, \overline{X}_2)$ with radii

0.421 *S* and 0.548 *S* for the 10% and 5% norms, respectively. (6.7.11)

Discussion of Example 6.7.2

Here again, as in the case of the rectangular regions of the previous example, the two different kinds of test produce, in terms of the self-same norms of physical judgement, regions of the same form, but of vastly different size. For the regions arising at (6.7.9) and (6.7.11), respectively, the difference in area is of the order

$\pi(0.724)^2 = 1.65$ versus $\pi(0.421)^2 = 0.56$ when the norm is 10%, and
$\pi(0.843)^2 = 2.23$ versus $\pi(0.548)^2 = 0.94$ when the norm is 5%.

This again reflects the differences in meaning explicated in Example 6.7.1. We need not repeat that explication here. Instead, let us use graphical means to underscore the different meanings. The formal mathematics of the matter was introduced to deal with the target problem of Section 3.11, where we developed a recipe for providing, in each of a series of repetitions, a circular confidence region for the centre of the target's position in that particular repetition. Suppose the target was mounted against a wall in of course different positions from repetition to repetition. Following each repetition, we might then use our recipe to draw on the wall the circular confidence region for that particular repetition, maintaining say a 95 % confidence coefficient. Now bear in mind that the centre of the region would, in accordance with Section 3.11, vary from repetition to repetition, as would in fact also the radius. Figure 6.7.2 depicts the kind of display our drawings would provide.

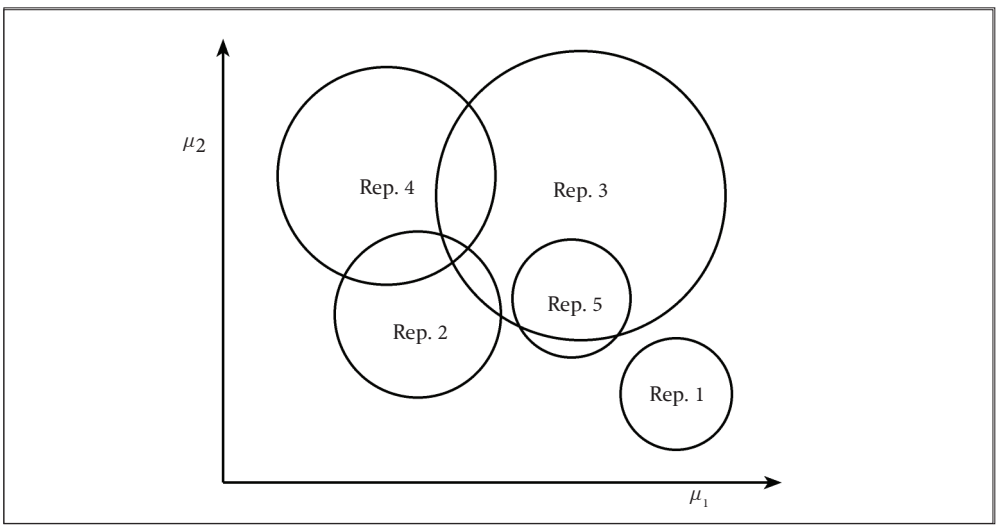


Figure 6.7.2: Depicting confidence regions from repetitive attempts at locating a target

Next, consider the present co-ordinate test procedure for using a solitary, real-world data set in order to develop an informed opinion on the current position of the earth's magnetic pole, and consider how we might then graphically display our evidence as in Figure 6.7.3;

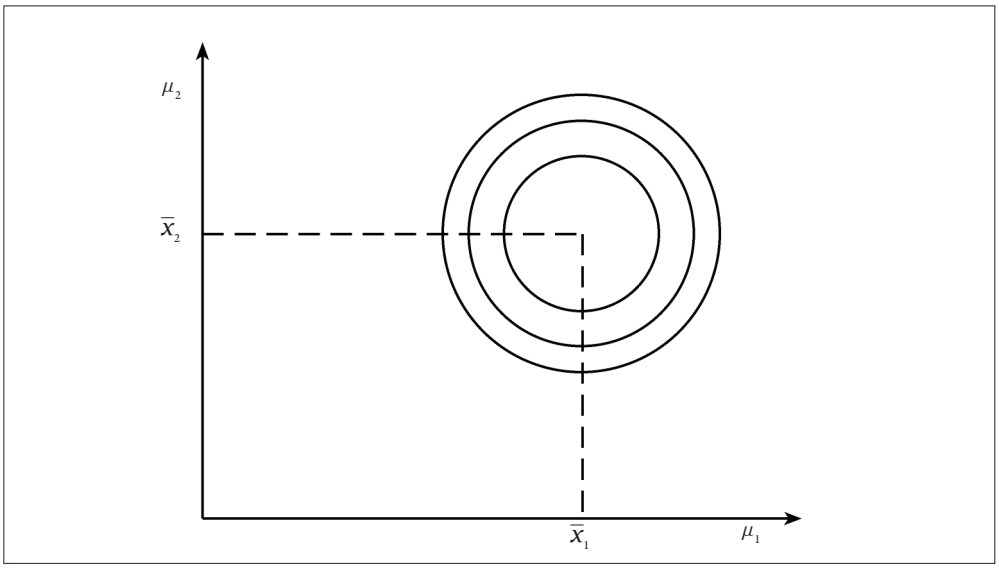


Figure 6.7.3: Depicting a suite if co-ordinating traces, given a data set for locating a target

any line through the centre of the concentric circles provides an abbreviated trace for the position of the pole on that line via its intersections with those circles. For instance, if n = 10, Student's *t* is based on 2(10-1) = 18 degrees of freedom. So if the circles must give the hypothesised positions of the pole (on that line) corresponding to the mental correlate of the test datum being situated in Student's test distribution, at say:

(0.01, •) and (•, 0.01) for the outer circle,
(0.05, •) and (•, 0.05) for the intermediate circle, and
(0.10, •) and (•, 0.10) for the inner circle,

then we must insert t = 2.55, 1.73 and 1.33, respectively, into the elimination pivot at (6.7.10). This holds for any line through the centre, i.e. for any line through $(\mu_1, \mu_2) = (\bar{x}_1, \bar{x}_2)$. If we rotate such a line through 180°, we obtain the concentric circles displayed in Figure 6.7.3. The following questions and answers now arise from contrasting Figures 6.7.2 and 6.7.3:

Question: Why do the circles in Figure 6.7.2 have different centres?
Answer: Because they arise from different data sets.

Question: Why do all the circles in Figure 6.7.3 have the same centre?
Answer: Because they arise from the self-same data set.

Question: Why do the circles in Figure 6.7.2 have different radii?
Answer: Because they arise from different error estimates.

Question: Why do the circles in Figure 6.7.3 have different radii?
Answer: Because they arise from different levels of co-ordination.

The reader should make sure that he or she clearly understands these questions and answers, as they concern the fundamental distinction between decision-making under risk, which is what hypothesis tests and confidence regions are intended for, and data analysis, which is what co-ordination tests and co-ordinate traces are intended for. This brings us to a further question: corresponding to Figures 6.7.2 and 6.7.3, where hypothesis tests and co-ordination tests are depicted, respectively, how can we depict significance testing? To begin with, let us copy Figure 6.7.3, but with the larger radii indicated at (6.7.9) instead of those indicated at (6.7.11). However, this is insufficient because the reasoning that would have us account not only for possible 'mistakes' as envisaged at (6.3.4), but also for *contrary* 'mistakes' as envisaged at (6.3.11), cannot involve just a single, solitary data set, just as one cannot, at one and the same time, make a 'mistake' and a *contrary* 'mistake'. So yes, we must for *any given data set* draw concentric circles with wider radii than those depicted in Figure 6.7.3, but in order to depict *the different data sets* involved by the reasoning, we must then also draw other, differently centred sets of concentric circles, in order to depict the behaviour of Fisher's knowing subject in respect of this, that and the other possible data set. The result would then be a sort of hybrid between Figures 6.7.2 and 6.7.3, thus depicting how significance testing tries to enjoy the best of both of two different worlds by occupying a sort of half-way station between hypothesis tests and co-ordination tests.

6.8 Further confounding

We have seen how two-tailed tests arise from confounding indicative data patterns with counter-indicative data patterns, and we have seen how that is motivated by the notion of the possible mistakes of a knowing subject. Inasmuch as a data pattern in actual evidence is confounded with a data pattern not in actual evidence, there is no clear principle of relevancy that limits the nature of the latter pattern. All that is needed for its introduction is that it must be capable of being envisaged as one that might cause a knowing subject, when viewed as a hypothetical investigator, to make a mistake in such an investigation. In this regard, the familiar *F* test of the analysis of variance is revealing, as the following examples will show.

Example 6.8.1

Table 6.8.1 displays the mean yields of k = 6 cultivars of lettuce averaged over n = 4 replicates (Mead and Curnow 1987, pp. 100-103); the error mean square arose from a wider source.

Table 6.8.1: Mean yields of six lettuce cultivars in a completely randomised design

Cultivar No.	1	2	3	4	5	6
Mean of n = 4 replicates	12.6	11.7	11.6	11.2	10.1	9.3
Error mean square = 5.050 on 45 degrees of freedom, drawn from a wider source						

We assume that the fixed model for analysis of variance is appropriate. First we consider how a data analyst would employ co-ordination tests to try to make sense of these data, and then we show that much confounding leads to the usual F test of the analysis of variance.

Co-ordinate tests applied to the lettuce data in Table 6.8.1

Let μ_i denote the population mean for the i^{th} cultivar, where interest is in μ_i-μ_j for all $i < j$ (i, j = 1, 2, 3, ..., k). The minimal sufficient statistic for μ_i-μ_j is the corresponding sample difference and the sample error variance. The expected error variance is a nuisance parameter, and is removed by using Student's t as elimination pivot. Consider trying to eliminate M_0: μ_i-μ_j = 0 for all $i < j$. For the lettuce data there are then 6-choose-2, i.e. 15 such elimination tests in total, these being, in obvious notation, as follows, in order of magnitude of the estimated differences in mean yield:

$\bar{x}_1$-$\bar{x}_6$ = +3.3	$\bar{x}_1$-$\bar{x}_5$ = +2.5	$\bar{x}_2$-$\bar{x}_6$ = +2.4	$\bar{x}_3$-$\bar{x}_6$ = +2.3	
t = +2.077	t = +1.573	t = +1.510	t = +1.447	
(•, ε, 0.022)	(•, ε, 0.061)	(•, ε, 0.069)	(•, ε, 0.077)	
$\bar{x}_4$-$\bar{x}_6$ = +1.9	$\bar{x}_2$-$\bar{x}_5$ = +1.6	...	$\bar{x}_2$-$\bar{x}_3$ = +0.1	
t = +1.196	t = +1.007	...	t = +0.063	
(•, ε, 0.118)	(•, ε, 0.160)	...	(•, ε, 0.475)	(6.8.1)

Such differences are of course not algebraically independent, and that often motivates poor epistemology because of a tendency to reason as if the given numerical data were the investigator's only source of knowledge. Such epistemology tends to reason that all the available information is accounted for by just 6-1 degrees of freedom. The reasoning often begins to go wrong by labelling the entries (cultivars in the present case) with numbers or letters, rather than with specific names, thus discounting any substantive knowledge. Concerning lettuce for instance, Wilson (1975, p. 57) lists four types, namely leaf or loosehead lettuce, butterhead lettuce, Romaine lettuce and head lettuce, with recommended varieties described as follows:

> **Leaf varieties:** Black Seeded Simpson (fast growing; outer leaves have a crumpled texture), Early Prizehead (brownish red leaves), Grand Rapids (slow to go to seed; will grow in a greenhouse in winter), Green Ice (glossy dark green leaves; slow to go to seed), Oakleaf (heat resistant), Ruby (reddish bronze leaves), Salad Bowl (slow to go to seed), Slo-bolt (compact dwarf plants; slow to go to seed)
>
> **Butterhead varieties:** Bibb (Limestone), Buttercrunch (heat-resistant Bibb), Butter King (disease resistant; slow to go to seed), Deer Tongue (slow to go to seed), Fordhook, Great Lakes (very productive, even under adverse conditions; resistant to tipburn and heat), Tom Thumb (miniature; good in containers)
>
> **Romaine (Cos) varieties:** Paris White, Valmaine (disease resistant)
>
> **Head varieties:** Iceberg (vigorous), Imperial No. 44 (forms good heads in warm weather), Premier Great Lakes (resistant to tipburn and heat)

As all these varieties are recommended, there must be reasons why none of them could be discarded without potential loss. Thus, supposing that Table 6.8.1 concerns six newly developed varieties of say leaf lettuce, there might be similar reasons for one to consider more than 6-1 comparisons amongst them. Amongst just three entries, say A (not resistant to tip burn), B (slow growing), and C (not slow to bolt), three contrasts, A vs. B, A vs.

C, and B vs. C, can be justified, and if in addition there exists conditions under which tip-burn, slow growth and early bolting are not of any concern, shortfall testing for those conditions bring the number of justifiable independent contrasts up to five, where algebra can find but two. So, instead of trying to derive an epistemology by counting degrees of freedom, we must instead establish just what the customer's questions are. For instance, we might find that the customer is interested in just seven different contrasts amongst the six population means, where the contrasts need not necessarily all be of the form indicated at (6.8.1). If each of those seven different contrasts were substantively well motivated, there would be no redundancy. All we then need to establish, one by one for each comparison in turn, would be just what the minimal sufficient statistic for that comparison is, and then make our statistical analysis rest on that statistic only.

The F test applied to the lettuce data in Table 6.8.1

Snedecor's *F* ratio for testing whether the cultivar means differ significantly is usually expressed as:

F = (cultivar mean square)÷(error mean square),

but it can also be expressed in an unusual form as

$$F = \binom{k}{2}^{-1} \sum_{i<j} \left(\frac{\bar{x}_i - \bar{x}_j}{\sqrt{2s^2 \div n}} \right)^2 , \text{ in obvious notation.} \qquad (6.8.2)$$

Here *F* is expressed as the mean square of all the values of Student's *t* listed at (6.8.1). The test is essentially due to R.A. Fisher, as *F* is a one to one transformation of a statistic that he originally developed for the same purpose. The form at (6.8.2) displays an extension of the confounding that we first met in Section 1.27, and that resurfaced at (6.3.11). At (6.8.2) each of the values of Student's *t* listed at (6.8.1) is squared, thus bringing about k-choose-2 instances of the confounding in Section 1.27. By computing the mean of those squared *t* values, all of those k-choose-2 instances of confounding are then also confounded with each other. Thus any positive value of say $\bar{x}_2 - \bar{x}_3$ is, by squaring, confounded with the corresponding negative value of $\bar{x}_2 - \bar{x}_3$, and any positive value of say $\bar{x}_4 - \bar{x}_6$ is, by squaring, confounded with the corresponding negative value of $\bar{x}_4 - \bar{x}_6$. Then, by averaging the squares, all those confounded pairs are confounded with each other. The confounding leads to possible contradictions that Fisher tried to reconcile. In the case of the lettuce data, for instance, the *F* test for cultivar differences is given by

> cultivar mean square ÷ error mean square = 5.686 ÷ 5.050, i.e. 1.126 on 5 and 45 df, whose mental correlate is situated at (0.640, ε, 0.360) in Snedecor's test distribution. (6.8.3)

Here the *F* test for cultivar differences does not produce a shred of evidence in favour of such differences, whereas the first four of the fifteen tests at (6.8.1) point quite strongly at the possibility of such differences. Note that in respect of the first of the 15 tests at (6.8.1) the contradiction is not removed by replacing the 15 co-ordinate tests with the corresponding two-tailed *t* tests. Fisher was aware of such possible contradiction. So he tried to reconcile the matter as follows (1966, p. 59):

> 'When ... the *F* test ... does not demonstrate significant differentiation, much caution should be used before claiming significance for special comparisons. Comparisons, which the experiment was designed to make, may, of course, be made without hesitation. It is comparisons suggested subsequently, by a scrutiny of the results themselves, that are open to suspicion, for if the variants are numerous, a comparison of the highest with the lowest observed value, picked out from the results, will often appear to be significant, even from undifferentiated material. Properly, such unforeseen effects should be regarded only as suggestions for future experimentation, in which they can deliberately be tested. To form a preliminary opinion as to the strength of the evidence, it is sometimes useful to consider how many similar comparisons would have been from the start equally plausible. Thus, in comparing the best with the worst of ten tested varieties, we have chosen the pair with the largest apparent difference out of 45 pairs, which might equally well have been chosen. We might, therefore, require the probability of the observed difference to be as small as 1 in 900, instead of 1 in 20, before attaching statistical significance to the contrast.' ('1 in 900' should be '1 in 90'.) (6.8.4)

We shall eventually come to disagree with almost all, if not all, of the ideas expressed in this proposal. Let us begin with the caution expressed in the opening sentence. And in doing so, let us avoid any misgivings that might arise from an involvement of more contrasts than there are degrees of freedom. So, consider say:

> $C_1 = \bar{x}_1 - \bar{x}_5$ and four further contrasts, C_2, C_3, C_4, and C_5, so chosen that the five contrasts form an orthogonal set. There might for instance possibly be interest in $C_1 = \bar{x}_2 - \bar{x}_6$ and $C_2 = \bar{x}_3 - \bar{x}_4$, with the remaining two contrasts simply chosen to complete the set, where (and this is a mathematical fact) that is always possible.

Let the j^{th} contrast be expressed as:

> $C_j = \sum \lambda_{ji} \bar{x}_i$, where $\sum \lambda_{ji} = 0$, and let $\sum \lambda^2_{ji} = D_j$, for $j = 2, 3, 4, 5$.

Instead of the form at (6.8.2), we then obtain:

$$F = \frac{1}{6\text{-}1}\left[\left(\frac{\bar{x}_1 - \bar{x}_5}{\sqrt{2s^2 \div n}}\right)^2 + \sum_{j=2}^{5}\left(\frac{C_j}{\sqrt{D_j s^2 \div n}}\right)^2\right]. \qquad (6.8.5)$$

Here the contrast of ~~i~~interest is $\mu_1 - \mu_5$, where the second term at (6.8.1) amounts to evidence that points quite strongly at $\mu_1 - \mu_5 > 0$ – evidence that was obtained by employing the minimal sufficient statistic for $\mu_1 - \mu_5$. At (6.8.5), however, the first term amounts to employing instead an insufficient statistic obtained by squaring the contrast in the minimal sufficient statistic, thereby confounding the sample pattern that *did* point at $\mu_1 - \mu_5 > 0$ with the contrary sample pattern that in fact was not in evidence. To that is then added the sum of four further terms involving irrelevant statistics only, statistics that convey nothing whatsoever by way of relevant evidence to the possible value of $\mu_1 - \mu_5$. Instead, the further terms have simply swamped the evidence of interest, resulting in the misleading *F* obtained at (6.8.3). Returning now to the opening sentence of the attempted reconciliation at (6.8.4), we might ask: 'Has Fisher not cautioned the wrong way round?' Surely the caution should

be that when the *F* test does not demonstrate significant differentiation, much caution should be used before claiming *insignificance* for special comparisons.

There is more to come.

Example 6.8.2

Table 6.8.2 displays the mean yields of k = 6 cultivars of lettuce averaged over n = 4 replicates, when cultivated under two different sets of microclimatic circumstances.

Table 6.8.2: Mean yields of six lettuce cultivars × two different microclimates

Cultivar	1	2	3	4	5	6
Climate 1	8.8	13.2	12.8	9.6	8.8	9.2
Climate 2	11.1	13.3	13.4	3.5	7.4	7.3
Error mean square = 5.050 on 45 degrees of freedom, drawn from a wider source.						

The data are adapted from the same source as the data of the previous example, and we assume again that the fixed model for analysis of variance is appropriate. Again we first consider how a data analyst would employ co-ordination tests to try and make sense of the data, and then show that much confounding leads to the usual *F* test of the analysis of variance.

Co-ordinate tests applied to the lettuce data in Table 6.8.2

We must begin by testing for interaction of cultivars × climates. Let μ_{il} denote the population mean for the i^{th} cultivar under the l^{th} climate, where interest is now in $(\mu_{i_1}-\mu_{j_1})-(\mu_{i_2}-\mu_{j_2})$ for all $i < j$ (i, j = 1, 2, 3, ..., k). The minimal sufficient statistic for $(\mu_{i_1}-\mu_{j_1})-(\mu_{i_2}-\mu_{j_2})$ is the corresponding sample difference and the sample error variance. The expected error variance is a nuisance parameter removed by using Student's *t* as elimination pivot. Consider trying to eliminate:

$$M_0: (\mu_{i_1}-\mu_{j_1})-(\mu_{i_2}-\mu_{j_2}) = 0 \text{ for all } i < j.$$

For the given data there are then (6-choose-2)(2-1) = 15 such elimination tests in total and, in an obvious notation, these are as follows in order of absolute magnitude of the estimated interaction:

$$\begin{array}{lll}
(\bar{x}_{11}-\bar{x}_{41})-(\bar{x}_{12}-\bar{x}_{42}) & (\bar{x}_{31}-\bar{x}_{41})-(\bar{x}_{32}-\bar{x}_{42}) & (\bar{x}_{21}-\bar{x}_{41})-(\bar{x}_{22}-\bar{x}_{42}) \\
t = -3.738 & t = -2.982 & t = -2.759 \\
(0.00026, \varepsilon, \bullet) & (0.00230, \varepsilon, \bullet) & (0.00418, \varepsilon, \bullet) \\
(\bar{x}_{41}-\bar{x}_{51})-(\bar{x}_{42}-\bar{x}_{52}) & (\bar{x}_{41}-\bar{x}_{61})-(\bar{x}_{42}-\bar{x}_{62}) & (\bar{x}_{11}-\bar{x}_{61})-(\bar{x}_{12}-\bar{x}_{62}) \\
t = +2.092 & t = +1.869 & t = -1.869 \\
(\bullet, \varepsilon, 0.02106) & (\bullet, \varepsilon, 0.03407) & (0.03407, \varepsilon, \bullet) \\
(\bar{x}_{11}-\bar{x}_{51})-(\bar{x}_{12}-\bar{x}_{52}) & (\bar{x}_{31}-\bar{x}_{61})-(\bar{x}_{32}-\bar{x}_{62}) & \ldots \\
t = -1.647 & t = -1.113 & \ldots \\
(0.05326, \varepsilon, \bullet) & (0.13581, \varepsilon, \bullet) & \ldots
\end{array} \qquad (6.8.6)$$

These tests reveal seven interactions arising from a simple pattern: the yield of Cultivar 4, contrary to the other 5 cultivars, drops sharply from Microclimate 1 to 2, whereas that of Cultivar 1, contrary to Cultivars 5 and 6, rises sharply from Microclimate 1 to 2.

The F test applied to the lettuce data in Table 6.8.2

Snedecor's *F* ratio to test for significant interaction is usually obtained as

> F = (cultivars × climates mean square)÷(error mean square),

but it can also be expressed in unusual form as:

$$F = \left\{ \binom{k}{2}^{-1} \sum_{i<j} \left(\frac{(\bar{x}_{i1}-\bar{x}_{j1})-(\bar{x}_{i2}-\bar{x}_{j2})}{\sqrt{4s^2 \div n}} \right)^2 \right\}. \qquad (6.8.7)$$

Here F is expressed as the mean square of all the values of Student's t listed at (6.8.6), thus displaying extensive confounding of the kind displayed in the previous example. Using the usual computational formula, we find the F value as

> cultivar × climates mean square ÷ error mean square = 16.64÷5.050, i.e. 3.295 on 5 and 45 df, whose mental correlate is situated at (0.987, ε, 0.013) in Snedecor's test distribution.

In this example, contrary to the previous one, the F test has revealed significance of some constituents of F. But how might we identify those constituents? There is just one way. We must scrutinise the treatment means. And have we not already done so at (6.8.6) and in the best possible way? You disagree? Are you saying that inasmuch as the co-ordinate for Cultivars 1 and 6 for instance might well be (0.03407, ε, •) due to error, that co-ordinate might equally well have been (•, ε, 0.03407) due to error? So are you saying that if a knowing subject considers that co-ordinate value to be 'just decisive' against the possibility of no interaction, the probability of that subject being mistaken in that regard would be 2×0.03407? And, are you saying that inasmuch as the knowing subject then considers (0.03407, ε, •) and (•, ε, 0.03407) the 'just decisive' co-ordinate values, the probability of a similar mistaken judgement in respect of for instance Cultivars 2 and 4 would also be 2×0.03407? And so, are you therefore saying that for instance for all four these possible mistakes to be accounted for, etc.?

6.9 A further development of the *F* test

If we express each difference between a pair of means in the expression at (6.8.2) as a deviation from its expected value, we obtain:

$$F = \left\{ \binom{k}{2}^{-1} \sum_{i<j} \left(\frac{(\bar{x}_i-\bar{x}_j)-(\mu_i-\mu_j)}{\sqrt{2s^2 \div n}} \right)^2 \right\}.$$

which is distributed as Snedecor's test statistic. So if $\Pr[F \geq F(\alpha)] = \alpha$, the solution of

$$\binom{k}{2}^{1} \sum_{i<j} \left[\frac{(\bar{x}_i - \bar{x}_j) - (\mu_i - \mu_j)}{\sqrt{2s^2 \div n}} \right]^2 \leq F(\alpha),$$

for the differences $\mu_i - \mu_j$ $(i < j)$, is a 1-α simultaneous confidence region for those differences. By reasoning similar to that used in Example 6.7.2, the region in fact provides simultaneous confidence intervals for all possible contrasts amongst μ. This idea was outlined by Fisher (1966, p. 206) and developed by Scheffé (1953), owing to which it has become known as Scheffé's multiple comparison procedure. We return to this matter in Chapter 13.

6.10 A different question

The developments in Section 6.8 might lead one to think (wrongly) that in any fixed-model analysis of variance situation the *F* test for differentiating means is worthless. Certainly the examples of that section show that the *F* test is grossly overused. However, there are certain problems in which the test is appropriate, as the following example shows.

Example 6.10.1

Hald (1952, pp. 441-444) reports the $CaCO_3$ contents of 40 specimens of raw meal, as determined by duplicate titrations. The specimens were collected at regular intervals while emptying a mixer, but Hald shows that the order of collection may be ignored. Interest is in the possible incompleteness of mix from specimen to specimen. Hald finds:

> The between-specimens mean square = 0.06827 on 39 df.
> The within-specimens mean square = 0.008813 on 40 df.

So, to test for incompleteness of mix his datum is

> $F = 7.75$ on 39 and 40 df.,

whose mental correlate is situated to the right of (0.9999, ε, 0.0001) in Snedecor's test distribution. The evidence for incompleteness of mix is overwhelming. Note that this conclusion does not presuppose the random model. (Hald makes a good case for such a model, but that is irrelevant for the present purposes.) Note also that we *do* presuppose that there is no interest whatever in drawing conclusions that distinguish between this, that or the other particular specimen. Interest is only in a hypothesis concerning their common source. Our test is valid for the fixed model. (That it would also be valid for the random model is irrelevant for the present purposes.)

6.11 The Studentised range test

Instead of using the *F* test in the fixed-model analysis of variance situation, we could for similar purposes use the Studentised range statistic, defined for k sample means as

$$W = \frac{\bar{X}_{(k)} - \bar{X}_{(1)}}{\sqrt{S^2 \div n}}, \text{ where } \bar{X}_{(k)} \text{ is the largest and } \bar{X}_{(1)} \text{ the smallest sample mean.}$$

The hypothesised model and the alternatives are the same as for the *F* test, but the two tests differ in sensitivity. We consider two examples, followed by a discussion.

Example 6.11.1

For the data in Example 6.8.1 we find the test datum is given by

$$w = \frac{12.6\text{-}9.3}{\sqrt{5.050 \div 4}} = 2.94 \text{ for } k = 6 \text{ means on } 45 \text{ df},$$

whose mental correlate is at approximately (0.89, ε, 0.11) in the test distribution. So again, just as in Example 6.8.1, there is not a shred of evidence to warn us of the substantial differences that are indicated by the analysis at (6.8.1). The Studentised range test is here just as inappropriate as the *F* test was in Example 6.8.1.

Example 6.11.2

For Hald's data referred to in Example 6.10.1, the W test datum turns out to be

$$w = \frac{0.85\text{-}0.10}{\sqrt{0.008813 \div 2}} = 11.3 \text{ for } k = 40 \text{ means on } 39 \text{ df},,$$

whose mental correlate is co-ordinated very far out to the right of (0.99, ε, 0.01) in the test distribution. So again, just as in Example 6.10.1, the evidence for incompleteness of mix is overwhelming. The Studentised range test is here just as appropriate as the *F* test was in Example 6.8.1.

Discussion of examples 6.11.1 and 6.11.2

Examples 6.11.1 and 6.11.2 differ from Examples 6.8.1 and 6.10.1, respectively, only by the use of *W* instead of *F*. In Example 6.11.1, as in Example 6.8.1, the *W* and *F* tests, respectively, are inappropriate for the same reasons. And in Example 6.11.2, as in Example 6.10.1, the *W* and *F* tests, respectively, are appropriate for the same reasons.

6.12 A further development of the Studentised range test

Similar reasoning to that in Section 6.9, but applied to the Studentised range statistic instead of the *F* statistic, provides simultaneous confidence intervals for all possible contrasts amongst μ. The method was developed by J.W. Tukey owing to which it has become known as Tukey's multiple comparison procedure (Scheffé 1959). We return to this matter in Chapter 13.

6.13 Simulation of a disagreement

Much confusion could be avoided if only all of us could come to understand that any findings by statistical data analysis *in respect of a solitary data set*, is fully as capable of laboratory demonstration as would be any findings by another physical science in respect of the solitary individual involved – even more so, as the tackle required for the

statistical demonstration is no more than a facility for simulating random numbers. So let us use this in order to explain why a data analyst must disagree with the reasoning at (6.8.4). The reasoning is supposed to concern the development of tenable values of parameters of interest, given a class of models as possible explanations of how given data might have come about. So, we must consider how to simulate different reasoning that in Section 6.8 leads to disagreement on the tenability of parameter values accountable for the differences in the mean performances of the six lettuce cultivars in Table 6.8.1. It is then worthwhile to identify the cultivars by name, as a reminder that the questions to be dealt with originate in substantive science. So, let the cultivars be:

(1) Fiesta (2) Dillie (3) Millie (4) Crispy (5) Billie (6) Beauty.

In Table 6.8.1 the cultivars are arranged in order of magnitude of the mean yields that happened to have been observed in the particular trial. And at (6.8.4) Fisher wants us to reason about the difference between the largest and the smallest of those means. It then appears here, at the very outset, that a disagreement is surfacing. At (6.8.4) Fisher's reasoning would have us model that difference as the realisation of an order-statistical quantity involving all six cultivars, whereas at (6.8.1) our co-ordination test models that difference as the realisation of a statistic that involves just two cultivars. Note that naming the varieties has underscored the disagreement, as it underscores the *substantive* nature of the question: 'How do the means of Fiesta and Beauty compare?' as compared to the *statistical* nature of the question: 'Is the largest difference amongst the six observed means too large to justify an order-statistical explanation?'

Simulating the co-ordinate tester's reasoning

First we simulate the hypothesised model: let the cultivars be modelled as six normal populations with common variance σ^2 to which we may assign an arbitrary value, as its contribution is eliminated by Student's t. Adjoin a 7th normal population also with variance σ^2 to provide a wider source for the error estimate. Assign an arbitrary, but common, value to the means of the populations named Fiesta and Beauty. Assign arbitrary values to the means of the other five populations. Draw samples of size n = 4 from each of the first six populations, and draw a sample of size n = 28 from the wider source. Calculate the sample means for Fiesta and Beauty, respectively, and calculate the error variance based on 6(4-1)+(28-1) = 45 df. In the previous notation, the parameter of interest is μ_1-μ_6, and the minimal sufficient statistic for that parameter is

$$\{\overline{X}_1\text{-}\overline{X}_6,\ S^2\},\ \text{with realised value}\ \{\overline{x}_1\text{-}\overline{x}_6,\ s^2\}. \qquad (6.13.1)$$

Calculate the realised value of the minimal sufficient statistic indicated at (6.13.1) and discard the rest of the data, as those data cannot tell us anything whatsoever about the value of μ_1-μ_6. Then calculate the simulated value of

$$\text{Student's } t = \frac{\overline{x}_1\text{-}\overline{x}_6}{\sqrt{s^2 \div n}}\ ,\ \text{and store that value.}$$

Repeat this simulation say 50 000 times. Display the stored values in a histogram. Pick out the bar that represents the value t = +2.077, which is that obtained from the actual

data as reported at (6.8.1). Shade that bar. Compute the proportions of the 50 000 simulated t values that are < +2.077, are = +2.077, and are > +2.077, respectively, and note that these are for all practical purposes the statistical co-ordinates given at (6.8.1) as (•, ε, 0.022). Now, in order to simulate different alternative models, let a full set of (6-1) orthogonal contrasts among the six population means comprise μ_1-μ_6 and four other contrasts. This may be done in many different ways, where the set shown in Table 6.13.1 will do as well as any other. The values of such a set determine (this is simply a mathematical fact) the value of any other contrast we might wish to consider.

Table 6.13.1: Contrasts among population means representing different cultivars

Cultivar	Fiesta	Dillie	Millie	Crispy	Billie	Beauty
Population mean	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6
Contrast θ_1	+1	-1	0	0	0	0
Contrast θ_2	0	0	+1	-1	0	0
Contrast θ_3	0	0	0	0	+1	-1
Contrast θ_4	+1	+1	-1	-1	0	0
Contrast θ_5	+1	+1	+1	+1	-2	-2

For instance,

μ_2-μ_3 = $(\theta_4$-θ_2-$\theta_1)\div 2$,
μ_4-μ_5 = $(\theta_5$-θ_4-$2\theta_3$-$2\theta_2)\div 2$,
$[(\mu_2$-$\mu_3)\div 2]$-μ_5 = $(\theta_5$-$2\theta_3$+θ_2-$\theta_1)\div 4$,
and so on.

The contrast of interest, Fiesta versus Beauty, is represented by:

μ_1-μ_6 = $(-\theta_6+\theta_4+2\theta_3+2\theta_1)\div 4$,

and if we now repeat the simulation with the θ values specified such that the value of μ_1-μ_6 remains zero as before, but the contrasts between the population means are otherwise given any arbitrary values, we will obtain precisely the self-same statistical co-ordinates as in the previous simulation. But if we then also specify non-zero values of μ_1-μ_6, the co-ordinates deviate from those previously obtained. Thus:

If we specify μ_1-$\mu_6 > 0$, the co-ordinates will be to the left of (•, ε, 0.022).

If we specify μ_1-$\mu_6 = 0$, the co-ordinates will be exactly (•, ε, 0.022).

If we specify μ_1-$\mu_6 < 0$, the co-ordinates will be to the right of (•, ε, 0.022).

This is as it should be and shows that our tests at (6.8.1) are untrammelled by irrelevant values.

Simulating the reasoning proposed in (6.8.4)

Proceed with the same instructions used before, until reaching the situation indicated at (6.13.1). The next instruction for the previous simulation then reads:

> 'Calculate the realised value of the minimal sufficient statistic indicated at (6.13.1) and discard the rest of the data, as those data cannot possibly be telling us anything whatsoever about the value of μ_1-μ_6.'

But if we now try to follow this instruction, we will not be able to simulate the *F* test, and so will not be able to simulate the reasoning at (6.8.4). Clearly then, the reasoning at (6.8.4) involves irrelevancy, and the source of that irrelevancy is the idea of trying to account for the possible mistaken inferences on the part of a knowing subject.

6.14 A source of confusion peculiar to statistics

There can be no question that the origins of the notion of simultaneous statistical inference are largely to be found in the work of R.A. Fisher, as evidenced for instance at (6.8.4) where he advances the idea that comparisons an experiment was designed to make, may be made without hesitation; but that comparisons suggested subsequently, by scrutiny of the results, are themselves open to suspicion. Nowhere in substantive science does one find this idea. Consider, for instance, the evidential role of a hominid fossil known as Turkana Boy in Richard Leakey's theories about human evolution (Leakey and Lewin 1993). Is it at all sensible to hold that the tenability of his theories depend in some way or another on whether Leakey formulated them before or after he found Turkana Boy? Did it detract from Johan Kepler's theories that he developed them in the light of Tycho Brahe's data? And are Jane Goodall's theories about social behaviour in chimpanzees open to suspicion because she did not formulate them prior to collecting the observations on which they are based? The point here is simply that any investigative science is directed at trying to explain *how given data came about*, and any successful outcome must invariably take the one and/or the other of just two possible forms, namely:

> 'This model fits the data' or 'This model does not fit the data'.

Whether we arrived at a model before or after obtaining the data is evidentially vacuous. This is not to say that an investigator cannot usefully proceed in certain ways that will become part of the data and influence the value of the data. It is only to emphasise that the data is a record of real-world events. It can be relevant to ask how the data were recorded, whether a properly randomised design was used, why certain plots are missing, which commencement tests were performed, and with what results, as these questions concern real-world acts and events that tell us how the data addressed by the minimal sufficient statistic came about, or might have come about. But in respect of the thoughts of a knowing subject, whether *pre hoc* or *post hoc*, we have to ask how those thoughts contribute to the model under test, and we must hasten to point out that those *thoughts* are not to be confused with any *acts* that might well be part and parcel of the data to be explained. In order to fend off the silly notion of simultaneous statistical inference, it is worth emphasising seven principles that help prevent us from being beguiled by those notions:

> It cannot possibly make scientific sense to reason in ways that confound indicative data patterns with counter-indicative data patterns.
>
> It cannot possibly make scientific sense to reason in ways that confound indicative data patterns with irrelevant data patterns.
>
> Scientific evidence can be found in the real world, and the record of the real world, only. Any thoughts about evidence that might have been, but in fact was not, are irrelevant. If thoughts prompted acts, only the acts are relevant.
>
> A scientific model might well be a *predictive* model; indeed that is more often than not the case. The method of gathering data against which such a model might be tested could well have been *planned in advance*; indeed that is often the case. But ultimately any relevant evidence can be found in the record of past events and in that record only.
>
> A forecast is not proper evidence, because evidence cannot at present be found in the future. Forecasted error rates of Type I and Type II can at best be a muddled way of reporting alternative statistical co-ordinates.
>
> Science as evidenced by the substantive sciences tries to avoid the investigation of complex models as such. Any model is preferably to be analysed into as many self-contained subsidiary models as possible, which are to be tested separately, rather than simultaneously.
>
> As explained in the case of the liberty ship in Example 2.14.1, a *purely numerical* model without substantive content is of little, if any, interest to substantive science.

The last of these principles concerns a popular notion that Fisher, quoted at (6.8.4), expresses when he states that: 'comparisons suggested subsequently, by a scrutiny of the results themselves ... are open to suspicion'. This requires careful thought. What are 'the results'? If the results are interpreted as 'the numerical data', nonsense arises; what statistician has ever come across a substantive investigator asking how to test the significance of *a purely numerical* pattern noticed upon scrutiny of the data? In fairness to Fisher, what he had in mind when referring to 'scrutiny of the results' was more likely 'retrospective conjecture about possible causes', and as our example of the liberty ship in Section 2.14.1 shows, he is right in warning us to be very leery of explanations not supported by outside evidence and facts. However, he is wrong in thinking that there can be purely statistical recipes that would 'protect' us from errors arising from such conjecture. No such recipes can exist. The belief that simultaneous statistical inference can protect us against such errors, rather than to simply propel us into alternative errors, is a snare and a delusion. It is true that scientists sometimes draw erroneous conclusions, but the only protection science has against that is the self-correcting nature of science, owing to repetitive investigation to check important findings, and to science practising a sharp look-out for anomalous 'findings'.

6.15 Concluding remark

We have seen in Chapter 5 that Fisher tries to evade the frequentist vicious circle by interpreting significance levels as measurements that are the realised values of certain statistics. Arguably, however, the knowing subject of significance testing, that is to say, the hypothetical investigator whose potential for mistaken conclusions leads to the introduction of simultaneous statistical inference, fails to achieve that.

Chapter 7

OPTIMAL ELIMINATION TESTS

Their derivation by drawing on existing literature

7.1 Introduction

In Chapter 1, *commencement tests* with good separating characteristics were contrived by the following method. All the sample patterns arising from a fully specified model were ordered in such a way that patterns broadly envisaged as occurring 'more frequently' under an alternative of interest, were placed in a specified extreme of the ordering. In this chapter we consider methods of finding *elimination tests* with good separating characteristics. As the alternatives are then an array of fully specified models, the notion 'more frequently' becomes capable of being expressed in terms of a measure of comparative frequency. In a so-called *likelihood ratio ordering*, that measure is a ratio of alternative frequencies, that is to say, the likelihood ratio ordering for separating models M_j and M_k ($j \neq k$) from one another is obtained by ordering all the sample patterns according to the magnitude of the ratio

$$\Pr(\text{the sample pattern}|M_j) \div \Pr(\text{the sample pattern}|M_k) \text{ for different patterns.}$$

Recall that a test statistic is an ordinal label. So, any order-preserving transformation of a test statistic yields an equivalent test statistic, and any order-preserving transformation of the ordinal label arising from a likelihood ratio ordering may be taken to be *the* likelihood ratio test statistic. The array of models involved should then be capable of being viewed as the members of a substantively sensible class of models, as many a motley collection of alternative models can be considered, purely as a mathematical possibility. This is shown by the following example.

Example 7.1.1

Let a data set comprising $\underline{n}$ positive fractions $x_1, x_2, x_3, \ldots, x_n$ bring to mind:

M_1: Sampling from the density $f(x) = \exp(-x)$ on $0 < x < \infty$.
M_2: Sampling from the density $f(x) = 1$ on $0 < x < 1$, and $f(x) = 0$ on $1 \leq x < \infty$.

Then the probability of a sample in

$$\prod_{j=1}^{n} dx_j \text{ is given by } \prod_{j=1}^{n} \exp(-x_j)dx_j \text{ for } M_1, \text{ and by } \prod_{j=1}^{n} (1)dx_j \text{ for } M_2.$$

Let S denote the set on which the probability of any sample under M_2 is non-zero. Then the likelihood ratio is given by:

$$[\prod_{j=1}^{n}(1)dx_j] \div [\prod_{j=1}^{n}\exp(-x_j)dx_j] = \exp\left(\sum_{j=1}^{n} x_j\right) \text{ on S, and zero otherwise.} \qquad (7.1.1)$$

Let X denote the sample total. As X is an order-preserving transform of exp(*X*), it may be taken to be *the* likelihood ratio test statistic. This choice is convenient, as 2X under M_1 is distributed as chi-square on 2n df. For instance, by reading the numbers in the first n = 5 rows of Table A1 of Snedecor and Cochran (1989) as fractions, we simulate M_2. Then, rounding at the second decimal, we obtain the following data set:

$$(x_1, x_2, x_3, x_4, x_5) = (0.54, 0.15, 0.86, 0.61, 0.05). \qquad (7.1.2)$$

For these data, X = 2.21. Interpreting 2×2.21 as a chi-square value on 2×5 degrees of freedom, we find that under M_1 the co-ordinates of

> *X* = 2.21 are (U_1, V_1) = (0.076, 0.924), a poor fit of M_1, as is to be expected from simulating M_2.

Under M_2 the sample total is approximately normal with expectation equal to n÷2 and variance equal to n÷12, in which case, with n = 5, the co-ordinates of:

> *X* = 2.21 are $(U_2, V_2) \approx (0.327, 0.673)$, a good fit of M_2, as is to be expected from simulating M_2.

For any data, the sample space will be discrete because of rounding. The investigator might thus report the foregoing co-ordinates as being approximately

$(U_1, \varepsilon_1, V_1) = (0.08, \varepsilon_1, 0.92)$ under M_1, and
$(U_2, \varepsilon_2, V_2) = (0.33, \varepsilon_2, 0.67)$ under M_2.

Here adequately precise measurement of the original data points would ensure that the values of ε_1 and ε_2 would round to zero. Consider, as a third model for this example,

> M_3: Sampling from the density f(x) = 2x on $0 < x < 1$, and f(x) = 0 on $1 \le x < \infty$.

The probability of a sample in

$$\prod_{j=1}^{n} dx_j \text{ is given by } \prod_{j=1}^{n} 2x_j dx_j \text{ for } M_3.$$

In order to separate models M_1, M_2 and M_3 from each other, we then have three quite different likelihood ratio tests, as follows: for separating M_1 and M_2 we found the requisite likelihood ratio at (7.1.1); for separating M_1 and M_3 the requisite likelihood ratio is given by

$$[\prod_{j=1}^{n} 2x_j dx_j] \div [\prod_{j=1}^{n}\exp(-x_j)dx_j] = 2^n \left(\prod_{j=1}^{n} x_j\right) \exp\left(\sum_{j=1}^{n} x_j\right) \text{ on S,}$$

and zero otherwise; (7.1.3)

for separating M_2 and M_3 the requisite likelihood ratio is given by:

$$[\prod_{j=1}^{n} (1)dx_j] \div [\prod_{j=1}^{n} 2x_j dx_j] = 2^n \prod_{j=1}^{n} x_j. \qquad (7.1.4)$$

The likelihood ratio orderings arising at (7.1.1), (7. 1.3) and (7.1.4) are very different, and that is the point of this example. The example goes to show that a likelihood ratio ordering *as such* is a method for the pair-wise separation of singletons. This of course presents no difficulty for the theory of co-ordination tests, as that theory is in the first place a theory for pair-wise separation of singletons. In order to separate models M_2 and M_3, for instance, the likelihood ratio statistic arising at (7.1.4) may be taken to be:

$$Y = -\ln[\prod_{j=1}^{n} x_j] = -\sum_{j=1}^{n} \ln x_j, \text{ as } -\ln[2^{-n}(\bullet)] \text{ is an order-preserving transformation.}$$

The expected value and the variance of Y are given by:

E(Y) = n(1) and variance(Y) = n(1) under M_2, and
E(Y) = n(0.5) and variance(Y) = n(0.25) under M_3.

Y = 6.47 for the data given at (7.1.2). Using normal approximations, we find that the co-ordinates of Y = 6.47 are approximately (0.6446, 0.3554) under M_2, and approximately (0.9998, 0.0002) under M_3.

Here again the separation is as expected, as the data simulates M_2.

7.2 The Neyman-Pearson lemma for data analysis

In Section 3.6 we remarked that several variants of the Neyman-Pearson lemma on optimal tests can be developed. The variant appropriate to the use of significance tests is developed by Kempthorne and Folks (1971, pp. 318-320) and we now wish to draw on it, so as to derive the variant appropriate to the use of co-ordination tests. As noted at the outset of Section 1.29, we can use Definitions 1.19.1 and 1.19.2 to obtain a pair of formal significance levels, as follows, where T denotes the test statistic, and M_0 denotes the hypothesised model:

$$SL_U(T|M_0) = U(T|M_0)+\varepsilon(T|M_0).$$
$$SL_V(T|M_0) = \varepsilon(T|M_0)+V(T|M_0).$$

When T = 3 say, we may denote the values taken by the various constituents in these two expressions, as follows:

$$SL_U(T = 3|M_0) = U(T = 3|M_0)+\varepsilon(T = 3|M_0).$$
$$SL_V(T = 3|M_0) = \varepsilon(T = 3|M_0)+V(T = 3|M_0).$$

Some such notation is needed, as we will wish to distinguish between for instance

$SL_U(T = 3|M_0)$ and $SL_U(S = 3|M_0)$, where T and S denote different statistics.

Corresponding forms in SL_U and SL_V can be obtained from one another by inverting the ordering, and the variant of the lemma appropriate to the use of significance tests can in terms of one or the other of these forms, be given as follows:

> Let T denote the likelihood ratio test statistic for a significance test of M_0 versus M_1, and let S denote any other statistic that could be used for the same purpose. Let T = t and S = s for the data in hand. Then, by choosing one or the other direction of ordering, or by re-labelling the two models, we can arrange that:
>
> $$SL_V(T = t|M_1) \geq SL_V(T = t|M_0), \tag{7.2.1}$$
>
> that is to say, we can arrange for the likelihood ratio test, if sensitive, to be a right-sensitive test, in which case the theorem states that
>
> $$\text{if } SL_V(T = t|M_0) = SL_V(S = s|M_0), \tag{7.2.2}$$
> $$\text{then } SL_V(T = t|M_1) \geq SL_V(S = s|M_1). \tag{7.2.3}$$
>
> Also, by choosing one or the other direction of ordering, or by re-labelling the two models, we can arrange that
>
> $$SL_U(T = t|M_1) \geq SL_U(T = t|M_0), \tag{7.2.4}$$
>
> that is to say, we can arrange for the likelihood ratio test, if sensitive, to be a left-sensitive test, in which case the theorem states that
>
> $$\text{if } SL_U(T = t|M_0) = SL_U(S = s|M_0), \tag{7.2.5}$$
> $$\text{then } SL_U(T = t \,|\, M_1) \geq SL_U(S = s \,|\, M_1). \tag{7.2.6}$$

The inequalities at (7.2.1) and (7.2.4) merely serve to remind us that for any ordering such that the significance level arises in the right-hand tail of the test distribution, the inverse of that ordering is such that the same significance level arises in the left-hand tail of the test distribution. So the hypothesis of the theorem can be stated as at (7.2.2) or as at (7.2.5), whereupon the conclusion of the theorem is stated as at (7.2.3) or (7.2.6), respectively. Measuring 'sensitivity' as in our Definition 1.29.1, Kempthorne and Folks (p. 318) formulate the theorem in words. We reproduce the same theorem in slightly different words, as follows:

> The likelihood ratio significance test is a most sensitive significance test for testing a hypothesised singleton M_0 against an alternative singleton M_1, at every level of significance that is attainable by the likelihood ratio test.

The theorem says 'a most' rather than 'the most', meaning that at any of the levels in question, another test could be equally sensitive, but not more so. The levels in question are described as only those that are 'attainable by the likelihood ratio test', meaning that other tests might well attain levels of significance not attainable by the likelihood ratio test. We will presently encounter such examples.

We derive (below) the following theorem:

Theorem 7.2.1 (The Neyman-Pearson lemma for data analysis):

Let given data retrospectively bring to mind an array of singletons M_1, M_2, M_3, ... , as possible models of how those data might have come about. Let T denote the likelihood ratio test statistic for separating any given pair of those models M_j and M_k ($j \neq k$). Let S denote any other test statistic selected for that same purpose. Let T = t and S = s for the data in hand. Let $V_j(T = t)$ and $V_k(T = t)$ denote the right co-ordinates of T = t under M_j and M_k, respectively, and let $V_j(S = s)$ and $V_k(S = s)$ denote the right co-ordinates of S = s under M_j and M_k, respectively. Let the direction of ordering be chosen so that:

$V_k(T = t) \geq V_j(T = t)$.

Under this arrangement,

if $V_j(T = t) = V_j(S = s)$,

then $V_k(T = t) \geq V_k(S = s)$.

Stated otherwise, let $U_j(T = t)$ and $U_k(T = t)$ denote the left co-ordinates of T = t under M_j and M_k, respectively, and let $U_j(S = s)$ and $U_k(S = s)$ denote the left co-ordinates of S = s under M_j and M_k, respectively. Let the direction of ordering be chosen so that:

$U_k(T = t) \geq U_j(T = t)$.

Under this arrangement,

if $U_j(T = t) = U_j(S = s)$,

then $U_k(T = t) \geq U_k(S = s)$.

To summarise: When trying to explain how given data might (or might not) have come about, the likelihood ratio co-ordination test is a most sensitive co-ordination test of any explanatory singleton M_j, against any other explanatory singleton M_k, at any level of co-ordination attainable by that test, and regardless of whether that test is sensitive toward the right or the left.

(In the formulation of this theorem the reader might well prefer the description '*a most separating co-ordination test*' instead of the description '*a most sensitive co-ordination test*', as the former better describes what one is trying to achieve. We will use these expressions synonymously.)

As explained in Section 1.29, the ordering for a significance test is usually arranged so as to be right-sensitive rather than left-sensitive. So, consider the right-sensitive orderings involved at (7.2.1), (7.2.2) and (7.2.3) as given by:

O_T for T = 1, 2, 3, ..., t, ..., k, and O_S for S = 1, 2, 3, ..., s, ...,

such that for the given data T = t and S = s, and such that the two tests then attain the same significance level. Let us express the foregoing in terms of roundings and right co-ordinates, as follows, where the subscripts 0 and 1 indicate 'under M_0' and 'under M_1', respectively. Then the hypothesis of the theorem as stated at (7.2.2) can be expressed as

$\varepsilon_0(T = t)+V_0(T = t) = \varepsilon_0(S = s)+V_0(S = s)$,
where possibly $V_0(T = t) = \emptyset$, and also possibly $V_0(S = s) = \emptyset$,

and the conclusion of the theorem as stated at (7.2.3) can be expressed as:

$\varepsilon_1(T = t)+V_1(T = t) \geq \varepsilon_1(S = s)+V_1(S = s)$,
where possibly $V_1(T = t) = \emptyset$, and also possibly $V_1(S = s) = \emptyset$.

The theorem is trivially true for t = 1. For t = 2, 3, 4, ..., k, the recurrence relations of Theorem 1.23.1 imply that the hypothesis and the conclusion of the theorem can be restated as follows:

$$\text{If } V_0(T = t) = V_0(S = s) \text{ then } V_1(T = t) \geq V_1(S = s), \text{ for } t = 1, 2, 3, ..., k - 1. \qquad (7.2.7)$$

The inequality at (7.2.7) expresses sensitivity as defined in our Definition 1.23.4, and not as defined in our Definition 1.23.3, which goes to show that the Neyman-Pearson lemma is most satisfactorily a theorem for continuous test statistics. We may anticipate that care is needed in extending its use to discrete test statistics. Henceforth we will use Definition 1.23.4 unless stated otherwise. So, noting that M_0 and M_1 may be relabelled M_1 and M_0, respectively, without further consequence if done throughout, we now draw on the foregoing to obtain, as Theorem 7.2.1 overleaf, another variant of the Neyman-Pearson lemma. Note that the summary sentence in Theorem 7.2.1 states the theorem completely. However, we have used a formulation involving much redundancy in order to invite comparison of the wording of the present version of the lemma with that of the version previously given as Theorem 3.6.1. Bear in mind that the different versions rest on the same underlying mathematical facts. Even the auxiliary random numbers of the previous version do not represent a difference in mathematical facts because such numbers are available to anyone. In the previous version their availability must be recognised, as that might serve the statistician's purpose. In the present version their availability must be ignored, as that is irrelevant to the statistician's purpose. The issue is this: beyond mathematical facts, there is a vast difference between the employment of certain predictive concepts for forecasting, and the employment of those selfsame predictive concepts for predicating. Section 3.6 concerned problems in forecasting. The present section concerns problems in predicating.

We note that Theorem 7.2.1 does not say that the likelihood ratio test statistic is *the same* statistic for different pairs of models. Example 7.1.1 shows that the likelihood ratio test statistic can differ from pair to pair.

7.3 Likelihood ratio tests in relation to commencement tests

Consider the class of models that arises if independent Bernoulli trials with a constant probability of success θ are performed until exactly m successes have occurred. The sample pattern will comprise success (S) and failure (F), and the probability of, for instance,

FFFSFS ... S, is given by $(1-\theta)\times(1-\theta)\times(1-\theta)\times\theta\times(1-\theta)\times\theta\times...\times\theta$.

Thus, for this class of models in general the probability of the sample is given by:

$\theta^m(1-\theta)^X$ for X failures, X = 0, 1, 2, ...

The class is known as the *negative binomial* class. Its different members are indexed by θ $(0 < \theta < 1)$ and there are $(X+m-1)$-choose-$(m-1)$ different ways in which the first $m - 1$ successes can precede the m^{th} success, at which point the sampling stops. So the probability of the sample can be expressed as:

$$\left[\binom{X+m-1}{m-1}\theta^m(1-\theta)^X\right] \times \left[\binom{X+m-1}{m-1}^{-1}\right] \text{ for } X = 0, 1, 2, \ldots < \infty. \qquad (7.3.1)$$

Here the first factor in square brackets gives the statistic that is minimally sufficient for the index, and the second factor in square brackets gives the statistic that is minimally sufficient for the characteristic. A commencement test would be based on the second factor. Consider, for instance, how this class of models might be used to form an opinion on the frequency of occurrence of a certain rare characteristic in plants of a given species. Proceeding in a straight line through a field of plants, an investigator might identify any plants of the species of interest and, starting from a plant with the rare characteristic, might thereafter, in each specific case, record whether the characteristic is present (S) or absent (F). In order to ensure that at least a modicum of suitable data will be obtained, recording might be continued until say just m = 10 successes after the start have been recorded. The numbers of failures straddled by successive successes might then turn out to be as follows, in order of occurrence:

2, 11, 0, 7, 21, 8, 37, 11, 14, 43.

Should there retrospectively arise concern that, contrary to the chosen class of models, these data might be reflecting a downward trend in the value of θ, a commencement test might address that, as follows: compute the product moment correlation between the inter-event numbers and their ordinal positions, and then refer the result to the test distribution arising from random assignment of the 10 inter-event numbers to the 10 ordinal positions. For instance, with m = 3 the inter-event numbers might be 5, 9, 0, in that order, where the characteristic factor at (7.3.1) then assigns equal probabilities to the following six ordered sets of inter-event numbers:

(0, 5, 9) (0, 9, 5) (5, 0, 9) (5, 9, 0) (9, 0, 5) (9, 5, 0).

Should we then be satisfied with the commencement, we will proceed to eliminate. So, consider testing M_1: $\theta = \theta_1$ versus M_1: $\theta = \theta_2$. The likelihood ratio is given by:

$$\left[\binom{X+m-1}{m-1}\theta_2^m(1-\theta_2)^X\right] \div \left[\binom{X+m-1}{m-1}\theta_1^m(1-\theta_1)^X\right], X = 0, 1, 2, \ldots . \qquad (7.3.2)$$

We note that the factor representing the class characteristic at (7.3.1) cancels when the likelihood ratio is formed and from the development in Section 1.12 it appears at once that such will always be the case, no matter what class of models might be considered. Consider for instance any likelihood ratios that arise from the forms at (1.8.1), (1.8.2) and (1.12.1). Hence we obtain Theorem 7.3.1, which follows so directly from the definitions given in Section 1.12 that formal proof would be repetitious. Nevertheless, the theorem is very important, as it conveys an inescapable consequence of the taxonomy of statistical models – a consequence that the following example underscores.

Theorem 7.3.1:

A likelihood ratio co-ordination test is an elimination test when used to find the range of tenable values of the index of a class of models.

Example 7.3.1

Consider, as a class of models of how a given positive integer might have come about: $\Pr_\theta(X = x)$ for $\theta = 1, 2$, where:

$$\Pr_1 (X = x) = \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}, \frac{1}{64}, \ldots, x = 1, 2, 3, \ldots, \text{respectively, and}$$

$$\Pr_2 (X = x) = \frac{1}{4}, \frac{1}{2}, \frac{1}{16}, \frac{1}{8}, \frac{1}{64}, \frac{1}{32}, \ldots, x = 1, 2, 3, \ldots, \text{respectively.} \quad (7.3.3)$$

The likelihood ratio ordering for separating the two models is obtained from:

$$\Pr_2 (X = x) \div \Pr_1 (X = x) = \frac{1}{2}, 2, \frac{1}{2}, 2, \frac{1}{2}\ 2, \ldots, x = 1, 2, 3, \ldots, \text{respectively.}$$

The likelihood ratio ordering classifies the set of all possible sample patterns into just two mutually exclusive and exhaustive subsets, the one when X is even, and the other when X is odd. The likelihood ratio test statistic may therefore be taken to be:

T = T(X) where T = 1 when X is odd, and T = 2 when X is even.

Let the range of X be partitioned into the pairs (1, 2), (3, 4), (5, 6), ... For $\theta = 1$, the odd member of any given pair accounts for two thirds of the samples contributed to the population by that pair. For $\theta = 2$, the even member of any given pair accounts for two thirds of the samples contributed to the population by that pair. So:

$$\Pr_1(T = t) = \frac{2}{3} \text{ and } \frac{1}{3} \text{ for t = 1 and t = 2, respectively, and}$$

$$\Pr_2 (T = t) = \frac{1}{3} \text{ and } \frac{2}{3} \text{ for t = 1 and t = 2, respectively.} \quad (7.3.4)$$

Let the pairs (1, 2), (3, 4), (5, 6), ..., be enumerated by:

C = C(X) where C = 1, 2, 3, ..., for (x, x+1) = (1, 2), (3, 4), (5, 6), ..., respectively.

Then $X \leftrightarrow (T, C)$ is a one to one transformation, as follows:

$$1 \leftrightarrow (1, 1), 2 \leftrightarrow (2, 1), 3 \leftrightarrow (1, 2), 4 \leftrightarrow (2, 2), 5 \leftrightarrow (1, 3), 6 \leftrightarrow (2, 3), \ldots .$$

So

$$\Pr_\theta(X = x) = \Pr_\theta[(T, C) = (t, c)].$$

As (2c–1, 2c) is the pair of values enumerated by C = c, it follows from (7.3.3) that:

$$\Pr_1[(T, C) = (t, c)] = \left(\frac{1}{2}\right)^{2c-1} \text{ and } \left(\frac{1}{2}\right)^{2c} \text{ for t = 1 and t = 2, respectively, and}$$

$$\Pr_2[(T, C) = (t, c)] = \left(\frac{1}{2}\right)^{2c} \text{ and } \left(\frac{1}{2}\right)^{2c-1} \text{ for t = 1 and t = 2, respectively.} \qquad (7.3.5)$$

Referring back to (7.3.4) we now see that these probabilities factor as follows:

$$\Pr_1[(T, C) = (t, c)] = \left(\frac{2}{3}\right) \times \left[3\left(\frac{1}{2}\right)^{2c}\right] \text{ and } \left(\frac{1}{3}\right) \times \left[3\left(\frac{1}{2}\right)^{2c}\right] \text{ for t = 1 and t = 2,}$$

respectively, and:

$$\Pr_2[(T, C) = (t, c)] = \left(\frac{1}{3}\right) \times \left[3\left(\frac{1}{2}\right)^{2c}\right] \text{ and } \left(\frac{2}{3}\right) \times \left[3\left(\frac{1}{2}\right)^{2c}\right] \text{ for t = 1 and t = 2,}$$

respectively. So we have found that the probability of the sample can be expressed as

$$\Pr_\theta[(T, C) = (t, c)] = \Pr_\theta(T = t) \times \Pr(C = c \mid T = t),$$

where the first factor on the right depends on θ, and the second one is independent of θ. The factorisation also shows that C is independent of T, with distribution given by:

$$\Pr(C = c) = \frac{3}{4^1}, \frac{3}{4^2}, \frac{3}{4^3} \ldots, \text{ for c = 1, 2, 3, ..., respectively.}$$

T is minimally sufficient for the class index. So C is minimally sufficient for the class characteristic. The distribution of C shows that large values of C are uncharacteristic of the class of models as a whole, at least to the extent that a very large value of C would prompt us to wonder whether re-investigation would produce another similarly large value of C. So, consider the co-ordinates of C as providing a commencement test for the present problem. For C = c these co-ordinates are given by:

$$\left[1 - \left(\frac{1}{4}\right)^{c-1}, 3\left(\frac{1}{4}\right)^{c}, \left(\frac{1}{4}\right)^{c}\right] \qquad (7.3.6)$$

The co-ordinates of T = t provide a likelihood ratio elimination test, for which test the co-ordinates are obtained from the distributions given at (7.3.4) as

$$(\text{Ø}, \frac{2}{3}, \frac{1}{3}) \text{ under } M_1, \text{ and } (\text{Ø}, \frac{1}{3}, \frac{2}{3}) \text{ under } M_2, \text{ for T = 1, and}$$

$$(\frac{2}{3}, \frac{1}{3}, \text{Ø}) \text{ under } M_1, \text{ and } (\frac{1}{3}, \frac{2}{3}, \text{Ø}) \text{ under } M_2, \text{ for T = 2.} \qquad (7.3.7)$$

Using the formulae given at (7.3.6) and (7.3.7), we obtain the co-ordinates displayed in Table 7.3.1.

Table 7.3.1: Co-ordinates of some values of an elimination test statistic T, and of a commencement test statistic C, for a given data set under different models

X	(T, C)	Co-ordinates of C	Co-ordinates of T under M_1	Co-ordinates of T under M_2
1	(1, 1)	(∅, 0.75, 0.25)	(∅, 0.67, 0.33)	(∅, 0.33, 0.67)
2	(2, 1)	(∅, 0.75, 0.25)	(0.67, 0.33, ∅)	(0.33, 0.67, ∅)
3	(1, 2)	(0.750, 0.19, 0.06)	(∅, 0.67, 0.33)	(∅, 0.33, 0.67)
4	(2, 2)	(0.750, 0.19, 0.06)	(0.67, 0.33, ∅)	(0.33, 0.67, ∅)
5	(1, 3)	(0.938, 0.05, 0.02)	(∅, 0.67, 0.33)	(∅, 0.33, 0.67)
6	(2, 3)	(0.938, 0.05, 0.02)	(0.67, 0.33, ∅)	(0.33, 0.67, ∅)
7	(1, 4)	(0.984, 0.01, 0.00)	(∅, 0.67, 0.33)	(∅, 0.33, 0.67)
8	(2, 4)	(0.984, 0.01, 0.00)	(0.67, 0.33, ∅)	(0.33, 0.67, ∅)

We find that:

If X = 1 or 2:
C test: The class characteristic, as tested, fits the data very well.
T test: Both members, as tested, fit the data very well.

If X = 3 or 4:
C test: The class characteristic, as tested, fits the data well.
T test: Both members, as tested, fit the data very well.

If X = 5 or 6:
C test: The class characteristic, as tested, fits the data poorly.
T test: Both members, as tested, fit the data very well.

If X = 7 or 8:
C test: The class characteristic, as tested, fits the data very poorly.
T test: Both members, as tested, fit the data very well.

Etc.

The issue here is this: if substantive science and commencement testing have failed to bring a suitable class of models into the human mind, elimination tests are useless, if not misleading. This concerns a universal principle of taxonomy. Thus, for instance, according to *Roberts' birds of southern Africa* (Maclean, 1985), pipits (the species of the genus *Anthus*) 'are among the hardest of all birds to identify with certainty in the field; some are hard to identify even in the hand.' Cisticolas (the species of the genus *Cisticola*) too, 'are among the hardest of all small southern African passerines to identify by sight alone'. So, when pipits are mistaken for cisticolas, or *vice versa*, any criteria for separating species of the same genus are useless, if not misleading.

Remark on the wording of Theorem 7.3.1

Consider testing for homogeneity of variance, as a commencement test preceding the use of Student's *t* for the comparison of the means of two data sets being modelled as independent normal samples. Typically we then use Snedecor's *F* to test whether the two variances can tenably be modelled as equal. Such a test is a commencement test, not an elimination test, because we are then interested in the tenability of just a single value for the ratio of those two variances, not in a range of such values.

7.4 Likelihood ratio rounding

In this section we display certain circumstances under which likelihood ratio ordering is defective. We present this by way of three examples followed by a discussion.

Example 7.4.1

Consider a life-testing problem where the lifetimes of just n items under test can be modelled as a random sample from a population with density function

$e^{-(x-\theta)}$ for $0 < \theta \leq x$, and zero otherwise.

As the probability of failure before θ is zero, θ represents a threshold. So the smallest sample value, say S, cannot be less than θ. In fact S is minimally sufficient for θ, and it turns out that for a sample of size n, its density function is:

$ne^{-n(s-\theta)}$ for $0 < \theta \leq s$, and zero otherwise.

Consider:

$M_1: \theta = \theta_1$ versus $M_2: \theta = \theta_2$, labelled such that $\theta_2 < \theta_1$,

where such models cannot be sensibly put forward until the data have been examined because, letting s_{obs} denote the observed value to be assigned to the range of S for the data in hand, we must have:

$$0 < \theta_2 < \theta_1 \leq s_{obs}.$$

Otherwise we would be 'modelling' our datum in self-contradictory terms as one that could not have occurred. The thrust of this is that we are pointing at a solitary datum in the real world, and striving to bring to mind the matching members of a given class of models. The situation is depicted in Figure 7.4.1. In order to test the quality of fit of the models as depicted, it is appropriate to order on the values of S, as the values of S tend toward θ. The right co-ordinate of S = s is then given by:

$$\Pr(S > s|\theta) = e^{-n(s-\theta)} \text{ for } 0 < \theta \leq s_{obs} < \infty.$$

For the two models being considered, the co-ordinates of the mental correlate of s_{obs} are thus given, in terms of a token ε, by:

$$\left\langle 1-e^{-n(s_{obs}-\theta j)}, \varepsilon, e^{-n(s_{obs}-\theta j)} \right\rangle \text{ for } j = 1, 2. \qquad (7.4.1)$$

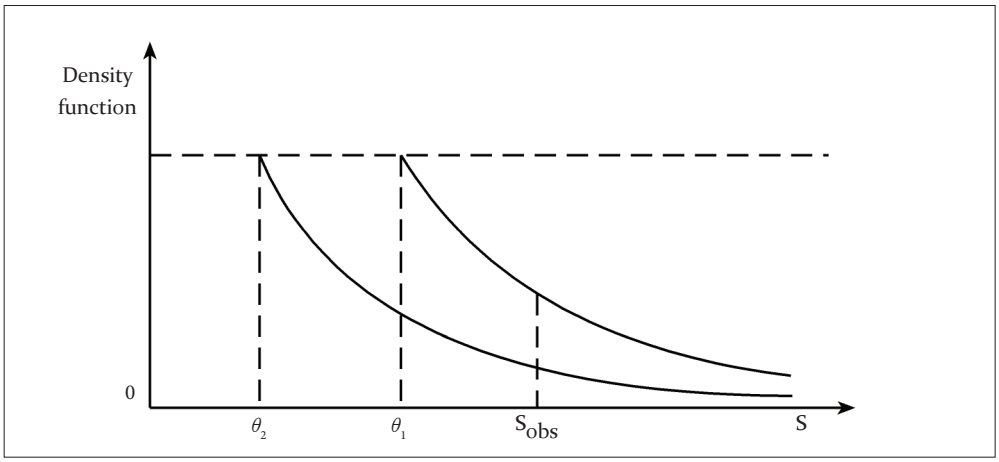


Figure 7.4.1: Two alternative models indexed by θ_2 and θ_1, respectively, brought to mind as possible explanations of how a solitary datum s_{obs} came into the real world. The statistical co-ordinates of the situation within each model of the mental correlate of datum s_{obs} are the areas under the density function, and to the left and right of s_{obs}

Inspection of Figure 7.4.1 shows that the possible co-ordinates of the mental correlate

range all the way from (0, ε, 1) to the right under M_1, (7.4.2)

but only from $\left\langle 1\text{-}e^{-n(\theta_1-\theta_2)},\ \varepsilon,\ e^{-n(\theta_1-\theta_2)} \right\rangle$ to the right under M_2, (7.4.3)

as s_{obs} cannot be to the left of θ_1.

Now consider the likelihood ratio test for this problem. The likelihood ratio equals

zero for $\theta_2 < s < \theta_1$, and $ne^{-n(s-\theta_2)} \div ne^{-n(s-\theta_1)}$ for $s \geq \theta_1$,

that is to say,

zero for $\theta_2 < s < \theta_1$, and $e^{-n(\theta_1-\theta_2)}$ for $s \geq \theta_1$. (7.4.4)

Recall that the actual values taken by the likelihood ratio are irrelevant here. It is only the ordering that matters. So all that matters here is that the likelihood ratio is

zero for $\theta_2 < s < \theta_1$, and a positive constant for $s \geq \theta_1$,

which means that the ordering for the likelihood ratio test statistic can be taken to be the subscript of O_T when

$$O_1 = \{S = s | \theta_2 < s < \theta_1\}, \text{ and } O_2 = \{S = s | s \geq \theta_1\}. \quad (7.4.5)$$

The mental correlate of s_{obs} is necessarily situated in the ordinal class on the right, as s_{obs} cannot be modelled to the left of θ_1. Thus, no matter what the value of s_{obs} might be, the corresponding value of T is necessarily $t_{obs} = 2$. So, the co-ordinates of the mental correlate of t_{obs} in the T test distribution are given by

$$(0, 1, \emptyset) \text{ under } M_1, \text{ and } (1\text{-}e^{-n(\theta_1-\theta_2)}, e^{-n(\theta_1-\theta_2)}, \emptyset) \text{ under } M_2. \quad (7.4.6)$$

Consider

$$n = 5, \theta_1 = 0.4, \text{ and } \theta_1 = 0.1, \text{ when } 0.4 \leq s_{obs}.$$

Then, no matter what the value of s_{obs} might have been (apart from being ≥ 0.4), the T test tells us that $t_{obs} = 2$, whose mental correlate in the T distribution is, according to (7.4.6), found at

$$(0, 1, \emptyset) \text{ under } M_1, \text{ and } (0.78, 0.22, \emptyset) \text{ under } M_2. \quad (7.4.7)$$

Thus, no matter what the value of s_{obs} might have been (apart from being ≥ 0.4), the T test would have us conclude that

both models, as tested, fit the data very well.

The form at (7.4.1) shows, as follows below, that for values of s_{obs} close to θ_1 the S test essentially agrees with this conclusion, but the form at (7.4.1) also shows that for certain larger values of s_{obs} the S test strongly disagrees:

If $s_{obs} = 0.40$, its mental correlate in the S test distribution is to be found at
$(0, \varepsilon, 1)$ under M_1, and $(0.78, \varepsilon, 0.22)$ under M_2. (7.4.8)

If $s_{obs} = 0.45$, its mental correlate in the S test distribution is to be found at
$(0.22, \varepsilon, 0.78)$ under M_1, and $(0.83, \varepsilon, 0.17)$ under M_2. (7.4.9)

If $s_{obs} = 0.65$, its mental correlate in the S test distribution is to be found at
$(0.71, \varepsilon, 0.29)$ under M_1, and $(0.94, \varepsilon, 0.06)$ under M_2. (7.4.10)

Compared to the pair of directions given at (7.4.10), the pair of directions given at (7.4.7) is extremely misleading. The immediate reason for this is that a continuum of possible S values have at (7.4.5) been rounded into just two discrete classes to which two arbitrary class marks (1 and 2) have been assigned. The effect of this rounding is displayed when we express the co-ordinates given at (7.4.6) as

$$(0, \varepsilon+1, \emptyset) \text{ under } M_1, \text{ and } (1\text{-}e^{-n(\theta_1-\theta_2)}, \varepsilon+1\text{-}e^{-n(\theta_1-\theta_2)}, \emptyset) \text{ under } M_2,$$

showing how they arose from those at (7.4.2) and (7.4.3), respectively, by a rounding which, in effect, rounds s_{obs} down to $s_{obs} = \theta_1$, regardless of the actual value of s_{obs}. Consequently, instead of being given the actual co-ordinates of the mental correlate of s_{obs}, we are in effect given the leftmost of its possible right co-ordinates. The fundamental reason for this is that when different sample patterns have the same likelihood ratio, they are put into the same ordinal class by a likelihood ratio ordering, though such patterns might nevertheless differ significantly with regard to the quality of fit of the different models under test. Thus, in the case of our example, any sample patterns such that $s \geq \theta_1$ are put into the same ordinal class at (7.4.3). Yet, the results given at (7.4.8), (7.4.9) and (7.4.10) show that this groups together patterns that have a very different bearing on the quality of fit of the different models under test.

Example 7.4.2

In the problem on the number of buses in Example 2.3.2, the minimal sufficient statistic for θ (the population number of buses) is $S = X_{(n)}$ (the largest sample bus number) and the corresponding real-world datum is $s_{obs} = x_{(n)}$ (the largest observed bus number) where θ cannot be $< s_{obs}$. The expression at (2.3.2) then shows that for

M_1: $\theta = \theta_1$ versus M_2: $\theta = \theta_2$, labelled such that $\theta_1 < \theta_2$,

the likelihood ratio is

a positive constant for $s \leq \theta_1$, and zero for $\theta_1 < s < \theta_2$. (7.4.11)

So, the likelihood ratio test statistic may be taken to be T in O_T when

$O_1 = \{S = s | s \leq \theta_1\}$, and $O_2 = \{S = s | \theta_1 < s < \theta_2\}$.

The mental correlate of s_{obs} is necessarily found in the ordinal class on the left, as θ cannot be $< s_{obs}$; so the value of the likelihood ratio test statistic is $t_{obs} = 1$. Consider for instance any value of $s_{obs} \leq 10$. Then $\theta_1 = 10$ and $\theta_2 = 12$ are possible values of θ, and $t_{obs} = 1$. Using the expression at (2.3.2) we find that the likelihood ratio test then tells us that the mental correlate of t_{obs} is situated in the T test distribution at

(Ø, 1, 0) under M_1, and at (Ø, 0.227, 0.773) under M_2. (7.4.12)

Thus if $s_{obs} \leq 10$, the likelihood ratio test tells us that no matter what the value of s_{obs} might otherwise be,

M_1 and M_2, as tested, both fit the data very well.

However, again using the expression at (2.3.2), we find that the S test tells us:

If $s_{obs} = 10$, its mental correlate is situated within the S test distribution at (0.400, 0.600, 0.00) under M_1, and at (0.091, 0.136, 0.773) under M_2. (7.4.13)

If $s_{obs} = 9$, its mental correlate is situated within the S test distribution at (0.133, 0.267, 0.600) under M_1, and at (0.030, 0.061, 0.909) under M_2. (7.4.14)

If $s_{obs} = 8$, its mental correlate is situated within the S test distribution at (0.033, 0.100, 0.867) under M_1, and at (0.007, 0.023, 0.970) under M_2. (7.4.15)

By expressing the pair of T co-ordinates given at (7.4.12) as

(Ø, 0.4+0.6, 0) under M_1, and (Ø, 0.091+0.136, 0.773) under M_2,

we display their relationship to the S co-ordinates given at (7.4.13), thus showing that all the values of s_{obs} that are possible under both models are in effect rounded up to $s_{obs} = 10$ (the value of θ_1) by the likelihood ratio ordering. Thus, for that particular s_{obs} value, the S and T tests essentially agree. However, whenever s_{obs} deviates from that value, for instance as at (7.4.14) and (7.4.15), the two kinds of tests disagree, owing to the likelihood ratio ordering having misleadingly rounded away informative variation in the minimal sufficient statistic. Certainly the pair of directions given at (7.4.12) is a misleading version of the pair given at (7.4.15).

Example 7.4.3

In Example 2.3.3 the vector $S = (X_{(1)}, X_{(n)})$ is minimally sufficient for the class index θ (an unknown integer). As explained in Example 2.3.3, the class of models involved can only be introduced retrospectively. Thus if $n = 3$ and $s_{obs} = (12, 15)$ for given data, any value of θ other than 8, 9, 10, 11, is impossible, as we must necessarily have

$$\theta+1 \le x_{(1)} \text{ and } x_{(n)} \le 2\theta-1, \text{ implying } \theta+1 \le 12 \text{ and } 15 \le 2\theta-1 \text{ for those given data.}$$

Consider, say

$$M_1: \theta = 9 \text{ versus } M_2: \theta = 11.$$

Using the formulae at (2.3.3) we previously found the distribution of S under M_2, as given in Table 2.3.3. Using the same formulae we find the distribution of S under M_1 as given in Table 7.4.1. Hence, the values of the likelihood ratio

$$\Pr(\text{the sample pattern}|M_2) \div \Pr(\text{the sample pattern}|M_1),$$

are those given in Table 7.4.2.

Table 7.4.1: Joint and marginal distributions of $X_{(3)}$ and $X_{(1)}$ for $\theta = 9$ and $n = 3$. $X_{(1)}$ ranges from $\theta+n-2$ to $2\theta-n$ inclusive, i.e., from 10 to 15 for $\theta = 9$ and $n = 3$. $X_{(3)}$ ranges from $\theta+n$ to $2\theta-1$ inclusive, i.e., from 12 to 17 for $\theta = 9$ and $n = 3$.

	$X_{(1)}$	10	11	12	13	14	15	Total
$X_{(3)}$								
12		1/56						1/56
13		2/56	1/56					3/56
14		3/56	2/56	1/56				6/56
15		4/56	3/56	2/56	1/56			10/56
16		5/56	4/56	3/56	2/56	1/56		15/56
17		6/56	5/56	4/56	3/56	2/56	1/56	21/56
Total		21/56	15/56	10/56	6/56	3/56	1/56	1

Table 7.4.2: Values of Pr(the sample when $\theta = 11$)÷Pr(the sample when $\theta = 9$)

	$X_{(1)}$	10	11	12	13	14	15	16	17	18	19
$X_{(3)}$											
12		0									
13		0	0								
14		0	0	7/15							
15		0	0	7/15	7/15						
16		0	0	7/15	7/15	7/15					
17		0	0	7/15	7/15	7/15	7/15				
18				∞	∞	∞	∞	∞			
19				∞	∞	∞	∞	∞	∞		
20				∞	∞	∞	∞	∞	∞	∞	
21				∞	∞	∞	∞	∞	∞	∞	∞

The likelihood ratio test statistic may be taken to be T in O_T, as follows:

T = 1 when the likelihood ratio is 0, as indicated in Table 7.4.2.
T = 2 when the likelihood ratio is 56/120 = 7/15, as indicated in Table 7.4.2.
T = 3 when the likelihood ratio is ∞, as indicated in Table 7.4.2.

For the given data, the sample pattern corresponding to s_{obs} is necessarily $\in O_2$, and so, no matter what that pattern might otherwise be, $t_{obs} = 2$, whose mental correlate in the T test distribution will be found to be situated

at (36/56, 20/56, Ø) under M_1, and at (Ø, 20/120, 100/120) under M_2,

that is to say,

at (0.64, 0.36, Ø) under M_1, and at (Ø, 0.17, 0.83) under M_2.

Thus if $s_{obs} \in O_2$, the likelihood ratio test will tell us that no matter what the value s_{obs} might otherwise be,

M_1 and M_2, as tested, both fit the data very well. (7.4.16)

Now consider two tests separately based on the two components of S, respectively, as developed in Example 2.3.3. Here the test datum for an $X_{(1)}$ test is $x_{(1)} = 12$, of which the mental correlate in the $X_{(1)}$ test distribution will be found to be situated

at (36/56, 10/56, 10/56) under M_1, and at (Ø, 36/120, 84/120) under M_2, (7.4.17)

and the test datum for an $X_{(3)}$ test is $x_{(3)} = 15$, of which the mental correlate in the $X_{(3)}$ test distribution will be found to be situated

at (10/56, 10/56, 36/56) under M_1, and at (1/120, 3/120, 116/120) under M_2, i.e.:
at (0.18, 0.18, 0.64) under M_1, and at (0.01, 0.02, 0.97) under M_2. (7.4.18)

The results at (7.4.18) cannot be described as at (7.4.16), and the disagreement cannot be brushed aside by the phrase 'as tested'. Instead, we must note that likelihood ratio ordering in this case destroys evidence in two different ways, as follows: firstly, as established from the outset in Section 1.12, and further explicated in Section 2.3, each of the two statistics $X_{(1)}$ and $X_{(n)}$ tells us something about the value of θ that the other one cannot tell us. More specifically, when we choose a θ value that is too large only $X_{(1)}$ can tell us that, and when we choose a θ value that is too small only $X_{(n)}$ can tell us that. Thus the tests at (7.4.17) are uninformative in the present case; it is the tests at (7.4.18) that inform us in this case. Secondly, as shown in Table 7.4.2, much evidence is destroyed by rounding $X_{(1)}$ and $X_{(n)}$ into dichotomous variables, thereby replacing the original sample space with the Cartesian product:

$$\{X_{(1)} < 12, X_{(1)} > 11\} \times \{X_{(n)} < 18, X_{(n)} > 17\}.$$

Discussion

Let us remind ourselves that ultimately any scientific model must be tested against its predictions. In the case of co-ordination testing, those predictions take the form of the predicted frequencies we refer to as left and right statistical co-ordinates. A likelihood ratio ordering will recognise, as different under different models, only those predicted frequencies whose variation under the different models is concomitant to variation in the likelihood ratio. However, our examples show that predicted co-ordinates can vary in ways that have a significantly different bearing on the tenability of certain alternative models, without there being any concomitant variation in the likelihood ratio. In much of the literature on statistical methods such examples are not found; for instance, there do not seem to be any such amongst the hundreds of examples given by Snedecor and Cochran in *Statistical methods* (1989). Nevertheless, Example 7.4.1 indicates the utility of such models in a specialised field. That, however, is not the reason for having dealt with them here. The reason for the present section is that certain literature tries to use such examples to persuade us to abandon the principle of testing statistical models against predicted frequencies (Jeffreys 1961, Edwards 1972, Basu 1975). In subsequent chapters we will have to consider the alternative epistemologies proposed in that literature. For the present it must suffice to note that when a particular principle fails to satisfactorily separate two models in terms of the predictions derived by means of that principle, we must try to find another principle to separate those models in terms of predictions. What we must *not* do, is fall into the fallacy of thinking that there can be non-empirical tests that can stand in for empirical tests. This is not to deny that a theory must be tested for inherent ambiguity, self-contradiction, or incoherence in some or other sense that would prevent us from establishing what in fact is being predicted by that theory. Such tests, however, are not empirical tests, whereas data analysis concerns the analysis of empirical data, and tests of the empirical kind.

Concluding remark

For a clear understanding of this section, the following is worth noting. Consider co-ordination tests of alternative singletons M_1 versus M_2 as models of how given data might

have come about. Let the co-ordinates of the mental correlate of the test datum in different test distributions for the likelihood ratio ordering then be:

$$C_1 = (U_1, \varepsilon_1, V_1) \text{ under } M_1, \text{ and } C_2 = (U_2, \varepsilon_2, V_2) \text{ under } M_2,$$

and for some other ordering be:

$$c_1 = (u_1, e_1, v_1) \text{ under } M_1, \text{ and } c_2 = (u_2, e_2, v_2) \text{ under } M_2.$$

Let it, for each ordering, be arranged that should the situation of the mental correlate deviate under M_2 from its situation under M_1, that deviation will be toward the right. The Neyman-Pearson lemma then tells us that if C_1 and c_1 coincide, C_2 and c_2 either coincide or else C_2 is further to the right than c_2. So, consider the possibility that c_1 is to the left of C_1. Could it then happen that c_2 is further to the right than C_2? The reader should verify that Examples 7.4.1, 7.4.2 and 7.4.3 produce no such cases. In general, such cases do not exist (Kempthorne and Folks 1971, Theorem 12.2, p. 321).

7.5 Exhaustively continuous test statistics

At (7.4.2) the co-ordinates of S = s range continuously from (0, ε, 1) to (1, ε, 0), as the value S = s ranges from θ_1 to $+\infty$, but at (7.4.3) no co-ordinate can be to the left of:

$$\left(1\text{-}e^{-n(\theta_1-\theta_2)},\ \varepsilon,\ e^{-n(\theta_1-\theta_2)}\right).$$

This motivates an asymmetry between the hypothesised model and the alternative in Definition 7.5.1.

Definition 7.5.1:

A test statistic is said to be *exhaustively continuous* if the co-ordinates of its values under the *hypothesised* model range continuously from (0, ε, 1) to (1, ε, 0).

Theorem 7.5.1 follows directly from this definition and the Neyman-Pearson lemma for data analysis.

Theorem 7.5.1:

If a likelihood ratio statistic for testing a hypothesised singleton M_j, against an alternative singleton M_i, is exhaustively continuous, then the test is a most sensitive test, invariably so over all levels of hypothesised co-ordination, and regardless of whether the test is sensitive toward the right or the left.

Example 7.5.1

Let circumstantial facts, a data set $S_x = \{x_1, x_2, x_3, ..., x_n\}$, and commencement tests bring into the human mind a random sample $S_X = \{X_1, X_2, X_3, ..., X_n\}$, from an $N(\mu,1)$ population, where μ is of interest. The raw data may then be replaced with the data mean $\bar{x}$, as the sample mean $\overline{X}$ is minimally sufficient for μ. The distribution of $\overline{X}$ is $N(\mu, n^{-1})$. So,

the likelihood ratio for testing any two distinct index values μ_i and μ_j against each other is given by:

$$\frac{1}{\sqrt{2\pi \div n}}\ \exp[-(\overline{X}-\mu_j)^2 \div \frac{2}{n}] \div \frac{1}{\sqrt{2\pi \div n}}\ \exp[-(\overline{X}-\mu_i)^2 \div \frac{2}{n}],$$

that is to say, by $\exp[\overline{X}n(\mu_j-\mu_i)] \times \exp[-(\mu_j^2-\mu_i^2) \div \frac{2}{n}]$.

Here the likelihood ratio test statistic may be taken to be $\overline{X}$, as the transformation

$$[\ln(\bullet)+(\mu^2-\mu^2) \div \frac{2}{n}] \div n(\mu_j-\mu_i)$$

is order preserving. Since $\overline{X}$ is an $N(\mu, n^{-1})$ random variable, it follows that no matter what value might be assigned to μ, the co-ordinates of the values that $\bar{x}$ might take range continuously from (0, ε, 1) to (1, ε, 0) as $\overline{X}$ ranges continuously from -∞ to +∞. Hence, an $\overline{X}$ test of μ_i against μ_j is a most separating test, invariably so over all levels of co-ordination, and regardless of whether the test is sensitive toward the right or the left.

7.6 Monotone likelihood ratios

An elimination tester would like a co-ordination test to be uniformly most separating over all parameter value pairs of interest and all levels of hypothesised co-ordination, regardless of whether the test is sensitive to the left or to the right. We are now ready to develop a large class of such tests for the case of a single parameter whose range is any set of values that are ordered in some or other substantively meaningful way, and where the probability of the sample differs for those different parameter values. This includes numerous problems of practical interest. And, as will appear in a subsequent section, it extends to further development in the case of nuisance parameters. The following example will help us come to grips with the matter.

Example 7.6.1

Let a solitary real-world data set, S_x, bring into the human mind, as explanatory model of how those data might (or might not) have come about, a random sample, S_X, from one or other member of a class of Poisson populations indexed by a parameter, θ, which ranges over a set of values ordered in some or other a substantively meaningful way. Let t = t(x) and T = T(X) denote the sum of the data values and the sum of the sample values, respectively. Then the likelihood ratio ordering for a test of $\theta = \theta_i$ versus $\theta = \theta_j$, for any $\theta_j \neq \theta_i$, is obtained by ordering on the magnitude of

$$[\exp(-\theta_j)\theta_j^t \div t!] \div [\exp(-\theta_i)\theta_i^t \div t!] = \exp[-(\theta_j-\theta_i)] \times (\theta_j \div \theta_i)^t.$$

Given any $\theta_j > \theta_i$, this likelihood ratio is a monotone increasing function of *t*, and the larger θ_j is in relation to θ_i, the larger the increase. So, the co-ordinates of *t* might (by way of a small right co-ordinate) point to the right of θ_j in favour θ_i. Conversely, given any $\theta_j < \theta_i$,

the likelihood ratio is a monotone decreasing function of *t*, and the smaller θ_j is in relation to θ_i, the larger the decrease. So, the co-ordinates of *t* might (by way of a small left co-ordinate) point to the left of θ_j in favour θ_i. Here (θ_i, θ_j) is *any* pair such that $\theta_j > \theta_i$, in the first instance, or *any* pair such that $\theta_j < \theta_i$ in the second instance, and the tests in question are likelihood ratio tests, as *t* is an order-preserving one to one transform of the likelihood ratio. Recall that according to the Neyman-Pearson lemma for data analysis, when trying to explain how a given data set might (or might not) have come about, a likelihood ratio co-ordination test is a most separating test of any explanatory singleton, here indexed by θ_i, versus any other explanatory singleton, here indexed by θ_j, at any level of co-ordination attainable by that test, and regardless of whether the test is sensitive toward the right or the left. It then becomes apparent that the foregoing has exemplified Theorem 7.6.1.

Theorem 7.6.1:

Let a class of statistical models indexed by a scalar parameter θ posses a monotone likelihood ratio test statistic T = T(X). Then an ordering on the magnitude of T will provide a most separating co-ordination test of $\theta = \theta_i$ vs $\theta = \theta_j$, invariably so over every level of hypothesised co-ordination attainable by the test, uniformly so over all (θ_i, θ_j) pairs, and regardless of whether the test is sensitive toward the right or the left. (If T is exhaustively continuous, the phrase 'attainable by the test' falls away.)

This theorem applies to a great many examples. It is in fact rather difficult to find an example of any substantive interest where a suitable class of statistical models is to be indexed by a scalar, but where the class does not posses a monotone likelihood ratio. And indeed, Theorem 7.6.2, being a consequence of Theorem 7.6.1, displays a fundamental form of many of those examples

Theorem 7.6.2:

Let a solitary real-world data set $S_x = \{x_1, x_2, x_3, ..., x_n\}$, bring into the human mind a class of explanatory models of how those data might have come about by way of a random sample $S_X = \{X_1, X_2, X_3, ..., X_n\}$, drawn from a population whose density $f(x; \theta)$ is indexed by a scalar, θ, and is of the 'Koopman-Darmois' form, i.e. one of the many forms covered by:

$$f(x; \theta) = a(\theta)b(x)\exp[c(\theta)d(x)],$$

when

$a(\theta)$ is any function of θ whose value is independent of x,

$b(x)$ is any function of x whose value is independent of θ,

$c(\theta)$ is any monotone function of θ whose value is independent of x, and

$d(x)$ is any function of x.

If so, that class of explanatory models will posses a monotone likelihood ratio test statistic T = T(X), which in fact will be given by:

$$T(X) = d(X_1) + d(X_2) + d(X_3) \ldots + d(X_n).$$

We note in passing that 'Koopman-Darmois' is a misnomer; the form originated in the work of R.A. Fisher. Theorems 7.6.2 and 7.6.1 hold for the following examples, in each of which we give a conventional expression for $f(x; \theta)$, and then use square brackets to exhibit the constituents $a(\theta)$, $b(x)$, $c(\theta)$, $d(x)$, in that order.

Example 7.6.2

For sampling a binomial population, f(x; θ) is given by:

$$\binom{n}{x}\theta^{x}(1-\theta)^{n-x} = [(1-\theta)^{n}]\left[\binom{n}{x}\right]\exp\left\{\left[1n\left(\frac{\theta}{1-\theta}\right)\right][x]\right\}.$$

Example 7.6.3

For sampling a negative binomial population, f(x; θ) is, as at (7.3.1), given by:

$$\binom{x+m-1}{m-1}\theta^{m}(1-\theta)^{x} = [\theta^{m}]\left[\binom{x+m-1}{m-1}\right]\exp\{[1n(1-\theta)][x]\}.$$

Example 7.6.4

For sampling a Poisson population, f(x; θ) is given by:

$$\frac{\exp\{-\theta\}\theta^{x}}{x!} = [\exp\{-\theta\}]\left[\left(\frac{1}{x!}\right)\right]\exp\{[1n\theta][x]\}.$$

Example 7.6.5

For sampling an exponential population, f(x; θ) is given by:

$$\theta\exp\{-\theta x\} = [\theta][1]\exp\{[-\theta][x]\}.$$

Example 7.6.6

For sampling a Pareto population when modelling a distribution of incomes exceeding a specified income x_0, f(x; θ) is given by:

$$\left(\frac{\theta}{x_0}\right)\left(\frac{x_0}{x}\right)^{\theta+1} = \left[\left(\frac{\theta}{x_0}\right)\right][1]\exp\left\{[(\theta+1)]\left[1n\left(\frac{x_0}{x}\right)\right]\right\}.$$

Example 7.6.7

For sampling an N(θ, k^2) population (where k is any known constant), f(x; θ) is given by:

$$\frac{1}{\sqrt{2\pi k^2}}\exp\{-\frac{(x-\theta)^2}{2k^2}\} = [\exp\{-\frac{\theta^2}{2k^2}\}][\frac{1}{\sqrt{2\pi k^2}}\exp\{-\frac{x^2}{2k^2}\}]\exp\{[\frac{\theta}{k}][\frac{x}{k}]\}.$$

Example 7.6.8

For sampling an N(k, θ^2) population (where k is any known constant), f(x; θ) is given by:

$$\frac{1}{\sqrt{2\pi\theta^2}}\exp\{-\frac{(x-k)^2}{2\theta^2}\} = [\frac{1}{\theta}][\frac{1}{\sqrt{2\pi}}]\exp\{[-\frac{1}{2\theta^2}][(x-k)^2]\}.$$

In all these examples, and in numerous others covered by Theorem 7.6.1, we thus have a completely satisfactory solution to the elimination problem – a solution that simply cannot be improved upon.

7.7 Unbiased hypothesis testing

By 'unbiased hypothesis testing' we mean that the frequency in repetitive testing with which a given null hypothesis is rejected whenever it is false, is expected to be at least as large as the frequency with which it is rejected whenever it is true. For instance, if k = 1 in Example 7.6.7, the likelihood ratio statistic may be taken to be:

$[d(X_1)+d(X_2)+d(X_3)...+d(X_n)]\div n = \overline{X}$, which is an $N(\theta, 1\div n)$ random variable.

So, following Kendall and Stuart (1961, p. 182), consider three possible recipes for an array of repeated hypothesis tests of H_0: $\theta = \theta_0$ based on this statistic, as follows:

Test a: The critical region is in the lower tail of the test distribution. For instance, in order to test $\theta = \theta_0$ repetitively we specify the Type I error rate as $\alpha = 0.050$. So the critical region is $\overline{X} \leq \theta_0 - 1.645(\sqrt{1\div n})$.

Test b: The critical region is in the upper tail of the test distribution. For instance, in order to test $\theta = \theta_0$ repetitively we specify the Type I error rate as $\alpha = 0.050$. So the critical region is $\overline{X} \geq \theta_0 + 1.645(\sqrt{1\div n})$.

Test c: The critical region is in both tails of the test distribution and equally so. For instance, in order to test $\theta = \theta_0$ repetitively we specify the Type I error rate in each of the two tails equally as 0.025, the overall the rate thus being $\alpha = 0.050$. So the critical region is $\overline{X} \leq \theta_0 - 1.960(\sqrt{1\div n})$ and $\overline{X} \geq \theta_0 + 1.960(\sqrt{1\div n})$ jointly.

Note that here we speak of tests labelled a, b and c, as if singular, though in reality each 'test' is an array of such tests, otherwise it would be meaningless to refer to frequencies or rates associated with that test. Such ambiguity is ubiquitous in the language of hypothesis testing, and must at all times be grasped firmly because, as we have seen in Chapter 4, it is the ambiguous language of a profoundly mistaken discourse. We must note for instance that Test c involves simultaneous statistical inference, because in the case of just a single, solitary data set, the realised value $\bar{x}$ of $\overline{X}$, cannot belong to one tail area while at the same time belonging to the opposite tail area. Bearing this in mind, we now note that the frequency of rejecting $\theta = \theta_0$ is, for each of these tests, a function of θ, the so-called *power function* of that test. The three power functions are compared diagrammatically in Figure 7.7.1, where a single fixed value of n and a single fixed value of α are illustrated. The power at $\theta = \theta_0$, which is the Type I error rate, equals α for each test.

Test a is biased in that the power is $< \alpha$ for any $\theta > \theta_0$.
Test b is biased in that the power is $< \alpha$ for any $\theta < \theta_0$.
Test c is unbiased in that the power is $> \alpha$ for any $\theta \neq \theta_0$.

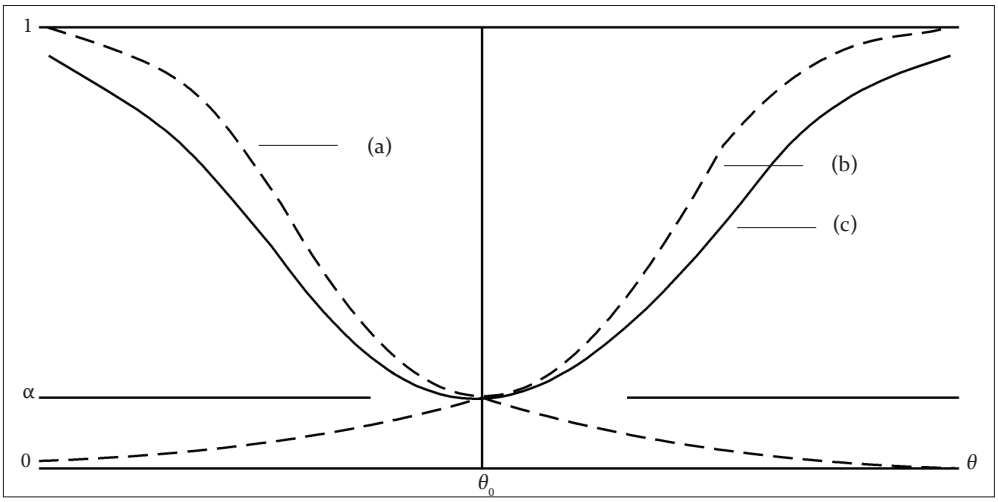


Figure 7.7.1: Power functions of three infinite arrays of hypothesis tests, where the critical region is (a) in the lower tail, (b) in the upper tail, and (c) in both tails equally

Now if, within the restricted class of unbiased tests for any given problem, there is a test that is uniformly most powerful, then we say it is *a uniformly most powerful unbiased test* (a UMPU test). In fact, Test c is such a test. However, the problem of finding a UMPU test, if extant in any given case, is often (even usually) not at all as simple as our example might lead one to suppose. In fact, there is an elaborate theory for finding such UMPU tests (see Lehmann 1986). However, we must note that the concept of 'an unbiased hypothesis test' can have no counterpart in co-ordination testing, as any co-ordination test involves only a single, solitary real-world data set. For instance Tests a, b and c above concern a certain paradigm in repetitive decision-making under risk. But they might be considered (in fact, they often *are* considered) by a mistaken view of a superficially similar, but profoundly different, paradigm in data analysis. For instance, commencement testing, substantive reasoning and circumstantial facts might make a data analyst investigate the quality of fit of the members of an $N(\theta, 1\div n)$ class of models $(-\infty < \theta < +\infty)$ as explanations of how just a solitary datum, $\overline{X} = \bar{x}$, might (or might not) have come about. In that case a suite of co-ordination tests that, for that solitary datum, are the most separating over all pairs of possible θ values, over all levels of hypothesised co-ordination, and are so regardless of whether sensitivity is to the left or to the right, is given by tracing the mental correlate of $\overline{X} = \bar{x}$ when that correlate is transported from model to model in the human mind. That trace is computed as:

$$[U(\theta), \varepsilon, V(\theta)] = [\Pr(\overline{X} < \bar{x} \mid \theta), \varepsilon, \Pr(\overline{X} > \bar{x} \mid \theta) \text{ for various } \theta\ (-\infty < \theta < +\infty),$$

and its general form is shown diagrammatically in Figure 7.7.2, where on the left it tends toward (1, ε, 0), thus pointing toward better fitting values of θ on the right, and on the right it tends toward (0, ε, 1), thus pointing toward better fitting values of θ on the left. In between it tends of course toward (0.5, ε, 0.5) thus pointing at $\theta = \bar{x}$ as the index of the best fitting model. Comparisons between the diagram in Figure 7.7.2 and those in Figure 7.7.1 are well worthwhile. Thus, for instance, the diagram in Figure 7.7.2 depends on $\bar{x}$, *which is a datum*, whereas the diagrams in Figure 7.7.1 depend on α, *which is a specification*.

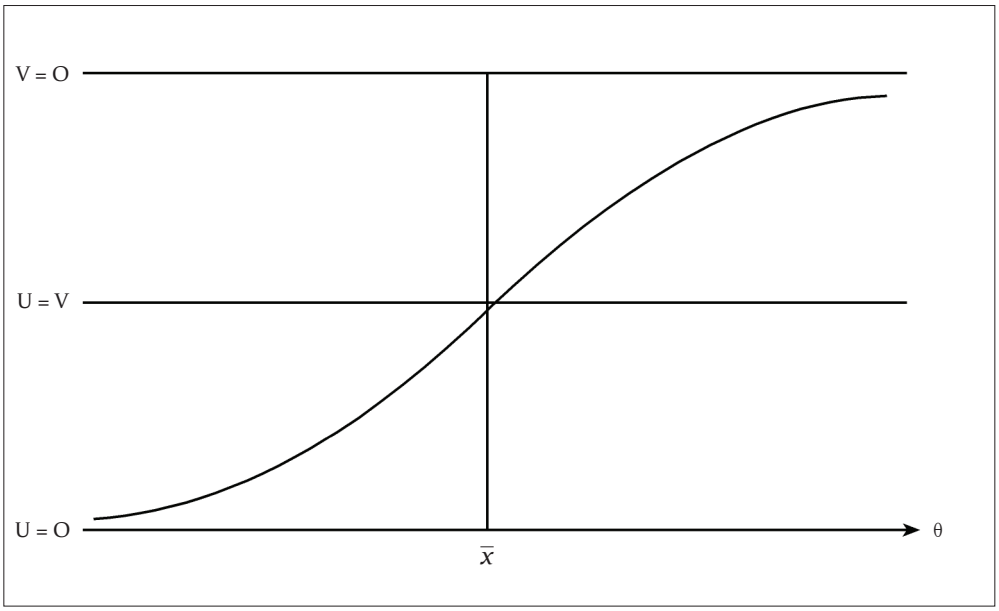


Figure 7.7.2: The trace of the mental correlate of just a solitary real-world datum, $\bar{x}$, whose mental correlate is being transported from model to model in the human mind

The datum concerns *a single real-world item*, whereas the specification concerns *a host of many real-world items*. Thus in each of the three diagrams given in Figure 7.7.1, $\theta = \theta_0$ is a fixture that applies to a host of many cases, whereas in case of the diagram in Figure 7.7.2, different values called θ_0 can be inserted on the θ axis in order to read off the corresponding co-ordination applicable to the solitary case under investigation. Such comparisons help us grasp how very different the matters are that these two different kinds of diagram depict, and thus to grasp that an array of hypothesis tests and a suite of co-ordination tests involve incompatibly different ideas, owing to which the idea of 'unbiased testing', just like the wider idea of 'simultaneous statistical inference' under which it resorts, simply has no place at all in a sound discourse on data analysis. The reader might also find it instructive to compare an 'ideal' array of hypothesis tests as depicted in Figure 7.7.3, to an 'ideal' suite of co-ordination tests as depicted in Figure 7.7.4.

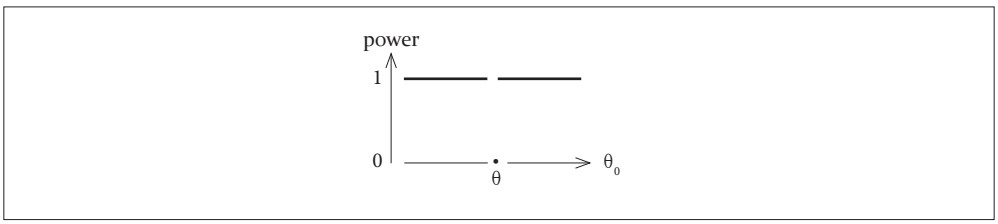


Figure 7.7.3: A depiction of the operating characteristics of an infinite array of ideal hypothesis tests of size α nearly = 0. The depiction imagines α = 0 as a possible specification.

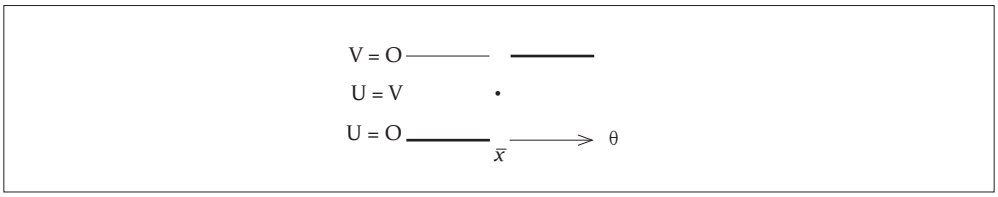


Figure 7.7.4: A depiction of the separating characteristics of a single suite of ideal co-ordination tests. All outcomes depend on just one, solitary real-world datum, $\bar{x}$.

7.8 Similar tests

Example 7.6.7 might interest an investigator of animal dispersion. A number of snails might for instance be released at a point represented as k in N(k, θ^2), and the model might then be used to represent the variance after a period of time, of the distances of individual snails from point k. Example 7.6.8 might interest an investigator when k^2 is known owing to a variance-stabilising transformation of given data, who might then wish to represent the data as a sample from an N(θ, k^2) population. A more commonly occurring problem arises when we wish to represent a given data set as a sample from an N(μ, σ^2) population with μ and σ^2 both unknown, and when σ^2 is a nuisance in case we are only interested in eliminating the untenable values of μ. A familiar solution to this problem using Student's *t* test, exemplifies the use of a *similar test*, the essence of the matter being that the problem is reduced to one involving only the parameter of interest when we condition on the observed value of a minimal sufficient statistic for the nuisance parameter. We have already met a very simple example of such a test in Example 2.5.2, as follows: given data modelled satisfactorily by independent Poisson samples indexed by μ_1 and μ_2, respectively, we wish to test the quality of fit of the model for different values of $\theta_1 = \mu_1 \div \mu_2$, in which case $\theta_2 = \mu_1 + \mu_2$ is a nuisance. If X_1 and X_2 denote the total counts for the two samples, respectively, we find (as shown in Example 2.5.2) the problem reduced to one where a binomial sample is indexed by a 1-1 transform of the parameter of interest, i.e. the problem is thus reduced to the one considered in Example 7.6.2. We will not pursue this matter further, as the developments are rather technical. Suffice it to say that similar tests greatly extend the scope of the development presented in Section 7.6.

7.9 Concluding remarks

We gave a brief explication of how optimal elimination tests of co-ordination can be derived from the existing literature. In Section 7.6 we saw that if the density sampled is one for which the likelihood ratio is monotone in some or other statistic, a completely satisfactory solution is obtained. In Section 7.8 we saw that the ideas in Section 7.6 can be made applicable to more complicated problems by sensibly restricting the class of tests to be considered. It is of course not to be expected that every problem in testing quality of fit will have a uniquely best solution. So, we might find we have to select two or more co-ordination tests that offer different separating characteristics, as we saw in for instance Example 2.3.3. The main trick in exploiting the existing literature for the

purpose of finding co-ordination tests is to be very leery of two-sided tests, unbiased tests, and other such ideas arising from notions of simultaneous statistical inference. It will in fact be found that we have in this and previous chapters provided an understanding of co-ordination tests that should be sufficient to enable for instance the customers of Snedecor and Cochran (1989) to page through that book and readily see how almost all, if not all, of the hundreds of tests it recommends for the analysis of this, that and the other given data, can be replaced by suitable co-ordination tests.

7.10 Economy of presentation

In developing the theory of co-ordination tests we have insisted that the statistical co-ordinates of the mental correlate of a given test datum cannot be conveyed precisely in terms of fewer than two numbers. This insistence is motivated by the need to make utterly clear that a set of statistical co-ordinates is intended to give directions, and is not at all intended to convey a probability. However, such insistence would burden statistical reportage unattractively; so to avoid that, we introduce Definition 7.10.1.

Definition 7.10.1:

Let (U, ε, V) denote a given set of statistical co-ordinates. The *augmented left co-ordinate* and the *augmented right co-ordinate* are given by:

$u = U + \varepsilon$ and $v = \varepsilon + V$, respectively,

where the notations 'u' and 'v', instead of 'U' and 'V', respectively, would allow tacit understanding to omit the term 'augmented' when referring to u or v.

Example 7.10.1

In an experiment on the effects of crowding on phenotypic variation in wheat, 1 000 plants were grown in pots (15.24 cm in diameter) with two plants per pot. Circumstantial details and commencement tests brought the standard random model of analysis of variance into the human mind. Results of analyses of the variance of number of seeds per ear and of number of ears per plant are given in Table 7.10.1.

Table 7.10.1: Analyses of variance of the two components of numbers of seeds/plant, obtained from an experiment on the effects of competition on phenotypes in wheat

Source	df	MS (seeds/ear)	MS (ears/plant)
Between pots	499	86.82	0.1213
Within pots	500	81.65	0.9248
		F = 1.063	F = 0.131
		v = 0.248	u = 0.000

For each one of the two analyses in this table, the expected mean squares in terms of the intraclass correlation ρ and the total variance σ^2 are as follows:

E('between-pots' mean square) = $[(1-\rho)+n\rho]\sigma^2$, and
E('within-pots' mean square) = $(1-\rho)\sigma^2$,

where $0 < \sigma^2 < \infty$, and, with n plants per pot, $-1\div(n-1) < \rho < 1$. From the *F* tests shown in the table, the competition between plants occupying the same pot seems to result in negative intraclass correlation for the number of ears per plant, but not for the number of seeds per ear. In order to develop an opinion about the magnitude of the intraclass correlation for the number of ears per plant, by way of an abbreviated trace, note that the pair of mean squares is minimally sufficient for the pair of parameters ρ and σ^2. The nuisance parameter, σ^2, is removed by forming the *F* ratio, which is distributed as

Snedecor's $F \times [1+n(\rho\div(1-\rho))]$.

So, by solving for ρ from equations of the form

observed $F \div [1+n(\rho\div(1-\rho))]$ = an appropriately specified percentage point of *F*,

the mental correlate of the observed pivotal value is found to be situated in Snedecor's test distribution

at v, u = 0.01 for ρ = -0.80, -0.73, respectively, and
at v, u = 0.05 for ρ = -0.79, -0.74, respectively.

Augmented co-ordinates are *formal* significance levels. However, when using such co-ordinates our notation and language must avoid the terms 'significance test' and 'significance level' for two reasons – firstly, because those terms are already being used ambiguously in the current statistical literature, and it can serve no positive purpose for us to add to that confusion, and secondly, because the current usage posits 'probabilities of "mistaken inference" by some or other knowing subject', whereas such a 'knowing subject' and corresponding probabilities of 'mistaken conclusion' have no place at all in the theory of co-ordination tests. Hence, our response to any attempt at interpreting augmented co-ordinates in probabilistic terms must be to revert at once to reasoning in terms of the (U, ε, V) presentation, it being imperative to grasp that statistical data analysis is not at all about 'how to take a chance'.

Chapter 8

STATISTICAL INTERVALS

Contriving to accommodate the *idée fixe*

8.1 Introduction

It is not at all uncommon for various members of a class of models to fit a given data set satisfactorily. In fact, in certain problems that will necessarily be the case. For instance, it is inconceivable that a well-planned data set on the difference in productive potential of two rooibos tea cultivars will not have the investigator conclude that the difference is for instance 'anything from say 10 kg/ha to say 15 kg/ha'. Hence the *idée fixe* of mathematical statistics has prompted 'statistical inference' to try to invent some or other recipe according to which such an investigator might add: '... and there is a 95 % probability that this conclusion is correct'. In Chapter 4 we saw how a mathematically ingenious but scientifically ham-fisted and ultimately circular attempt at incorporating this idea, leads to the dual theories of 'hypothesis tests' and 'confidence intervals'. In Chapter 5 we met R.A. Fisher's method for trying to evade that circularity. However, in Chapter 6 we saw that his method fails to avoid the introduction of another version of 'the knowing subject', and thereby relapses into transforms of the very defects it tries to avoid. In this chapter we display a further defect of that method, and we use that as a platform from which to show that a *probabilistic* idea of interval estimation, as an *investigative* tool, is inherently defective, regardless of the approach used. That is so because it contrives to foist an entirely unnecessary and ill-conceived statistical embellishment onto the discourse of investigative science.

8.2 A proposal of Kempthorne and Folks

Let X_j for j = 1, 2, 3, ..., denote an infinite array of independent $N(\mu_j, 1^2)$ random variables ($-\infty < \mu_j < +\infty$). An array of hypothesis tests of size α for any hypothesised value μ of μ_j versus larger values of μ_j rejects the hypothesised μ whenever

$$X_j - \mu > z_\alpha \text{ for } z_\alpha \text{ such that } \Pr(X_j - \mu_j > z_\alpha) = \alpha \text{ for } j = 1, 2, 3, \ldots . \tag{8.2.1}$$

A corresponding array of 1-α confidence regions is obtained by re-expressing this as

$$\Pr(X_j - \mu_j \leq z_\alpha) = 1-\alpha \text{ no matter what the value of } \mu_j, \text{ for } j = 1, 2, 3, \ldots . \tag{8.2.2}$$

The confidence interval when $X_j = x_j$ is thereby bounded from below by:

$$x_j - z_\alpha, \text{ for } j = 1, 2, 3, \ldots . \tag{8.2.3}$$

We thus obtain the array of intervals

$$x_j\text{-}z_\alpha \leq \mu_j < \infty \text{ for } j = 1, 2, 3, \ldots . \tag{8.2.4}$$

A decision-maker might thus bring into the real world an array of such intervals made to specification. Kempthorne and Folks (1971, p. 364 and onward) propose that a data analyst employ the same formal mathematics to envisage, for one solitary real-world datum x, a system of *consonance* intervals (not *confidence* intervals) in the human mind (not in the real world). This is accomplished by replacing the *specified constants* (α, z_α) with the *observed values* (p, z_p) for any hypothesised value μ, as follows:

$$\Pr(X\text{-}\mu > z_p) = p \text{ for variable } \mu \ (-\infty < \mu < +\infty). \tag{8.2.5}$$

The system of consonance intervals is obtained by re-expressing this as

$$\Pr(X\text{-}\mu \leq z_p) = 1\text{-}p \text{ for variable } \mu \ (-\infty < \mu < +\infty). \tag{8.2.6}$$

The consonance intervals envisaged for X = x are thereby bounded from below by

$$x\text{-}z_p \text{ for } -\infty < \mu < +\infty. \tag{8.2.7}$$

We thus obtain the system of intervals

$$x\text{-}z_p \leq \mu < \infty \text{ for } -\infty < \mu < +\infty. \tag{8.2.8}$$

Thus, if p = 0.100, 0.050, 0.025, then $x\text{-}z_p$ = x-1.282, x-1.645, x-1.960, respectively, and *vice versa*. Kempthorne and Folks (1971, p. 366) graphically display a system of such intervals. The reader should note that the development at (8.2.1), (8.2.2), (8.2.3) and (8.2.4) belongs to the discourse of decision-making under risk, which is the discourse of forecasting, whereas the development at (8.2.5), (8.2.6), (8.2.7) and (8.2.8) belongs to the discourse of data analysis, which is the discourse of pointing.

8.3 An epistemological consideration

Each interval at (8.2.8) includes its own boundary because at (8.2.5) we have used the exclusive Definition 6.2.2 of a significance level. If the inclusive Definition 6.2.1 were to be used at (8.2.5), each interval at (8.2.8) would exclude its own boundary. In the case of a continuous parameter space that might seem to be of little consequence. We will find, however, that in the case of a discrete parameter space, it does not make sense for an interval of interest to exclude its own boundary. The issue arises because habitual thought will try to achieve a set that *includes* the unknown parameter value, thereby neglecting the implication of a complementary set that *excludes* the unknown parameter value. In either case it will be found that sound epistemology requires a set that *includes* its own boundary. So we will have to reason in terms of a partition of the parameter space into two complementary intervals that may be called 'the *consonance* interval' and 'the *dissonance* interval', respectively, where these intervals will have to be constituted as follows:

> The values that comprise the consonance interval are its boundary and all those values that, by the tests performed, are as consonant or more consonant with the data, than the boundary is.

> The values that comprise the dissonance interval are its boundary and all those values that, by the tests performed, are as dissonant or more dissonant with the data, than the boundary is.

We will thus strictly maintain that such intervals always contain their own boundaries. We subsequently give examples showing that to reason otherwise, sows confusion. It will also be found that the ensuing development rests more easily on the term 'region' rather than on the term 'interval'.

8.4 Consonance regions and dissonance regions

We revisit the pregnant rabbit of Section 6.2. To simplify matters, suppose that by subjecting extensive historical records to suitable commencement and elimination testing, it has been found that any given amnion measurement x can be modelled satisfactorily as a realisation of an $N(\mu, 1^2)$ random variable X, where μ ($\mu = 1, 2, 3, \ldots$) denotes an unknown number of foetuses present. For the present purposes we avoid the notion of simultaneous statistical inference. So consider, for a given datum $X = x$, how to obtain tenable lower bounds for the possible values of μ. This is a problem in elimination testing, and it must be firmly grasped that such testing is subject to the following two constraints.

> The reasoning behind a commencement test whose purpose it is to provide a basis for elimination tests, cannot rely upon the notion that one can (somehow) arrive at 'the "probability" of having drawn a correct conclusion'. The very essence of any such real-world commencement problem is that it is utterly impossible to arrive at such a probability. Such a commencement test can provide nothing more than factual knowledge about quality of fit in a particular case. (8.4.1)

> It then follows inexorably that should our commencement tests (and there might be several of them) encourage us to adopt a particular class of models, any subsequent elimination tests cannot in turn arrive at anything in the nature of 'the "probability" of having drawn a correct conclusion'. So, owing to the commencement tests that they rely upon, elimination tests in turn can also provide nothing more than further factual knowledge about quality of fit in the particular case. (8.4.2)

It is precisely these constraints that motivate the methods of Chapters 1, 2 and 5, respectively. A further matter to be firmly grasped is that any exemplifying paradigm is always at best deliberately chosen so as to be of such extreme simplicity that the conclusions to be drawn from it are from the very outset obvious beyond reasonable contest. That is accomplished here by taking a given amnion measurement to be $X = 4$, its standard error to be $\sigma = 1$, and by adopting the following familiar norms, whose familiarity testifies to it that they can be forced upon the human body:

> Firstly, the model indexed by $\mu = 3$ fits the given data well, $\mu = 3$ being a mere one standard error unit less than $X = 4$ (the observed X value). (8.4.3)

> Secondly, the model indexed by $\mu = 2$ fits the given data poorly, $\mu = 2$ being all of two standard error units less than $X = 4$ (the observed X value). (8.4.4)

We must maintain these norms throughout the following development otherwise the issues are confused with different norms. Subject to the constraints recognised at (8.4.1) and (8.4.2) we must establish how different instruments of investigation try to correctly partition the parameter space with respect to the real-world pregnant rabbit at which we are pointing. And so, for each instrument in turn, we must establish whether or not it delivers an equivalent to the correct answers, that is to say, the answers already given at (8.4.3) and (8.4.4). So, given the class of models

$Z = X\text{-}\mu$ is a $N(0, 1^2)$ random variable, $\mu \in \{1, 2, 3, ...\}$,

and given the datum:

$X = 4$ in the particular case,

we must, using different instruments in turn, try to choose the appropriate partition of the parameter space from amongst the following partitions, where the models indexed to the left of the partition must, as tested, fit the given data poorly, and those indexed to the right of the partition must, as tested, fit the given data well:

$... \mu \in \{1, |2, 3, 4, ...\}, \mu \in \{1, 2, |3, 4, 5, ...\}, \mu \in \{1, 2, 3, |4, 5, 6, ...\}, ...$.

The parameter space is of course correctly partitioned by our choice of norms as

$$\mu \in \{1, 2, |3, 4, 5, ...\}. \tag{8.4.5}$$

Now, if for instance our instrument is co-ordination testing, we would consider the co-ordinates of $Z = X\text{-}\mu$ for various possible values of μ, and we would judge the better fitting of any two of the μ values to be the one giving Z co-ordinates that are nearest to $(U, V) = (0.5, 0.5)$. In the present case one then cannot disagree with the following:

> Given that $X = 4$, the co-ordinates of the Z values arising from $\mu = 3$ and $\mu = 2$ are $(0.84, 0.16)$ and $(0.98, 0.02)$, respectively. So there would seem to be three or more foetuses, and not two or less. (8.4.6)

In the present context, one cannot disagree with the statement at (8.4.6), as it is just a re-statement of the partition and motivating norms agreed upon at (8.4.3) and (8.4.4).

We are now ready to tackle the rabbit problem by means of the instruments indicated in Sections 8.2 and 8.3. So, consider $Z = (X\text{-}\mu)\div\sigma$ and let $P = p$ be the P value arising from $Z = z$ when Z is an $N(0, 1^2)$ random variable. Two distinctly different recipes for a 1-p lower bound for μ then arise from Definitions 6.2.1 and 6.2.2, as in the following, where those recipes appear on the extreme right at (8.4.7) and (8.4.8), respectively:

$$\Pr\left(\frac{X\text{-}\mu}{\sigma} > z\right) = p. \text{ So } \Pr(X\text{-}\mu \le z\sigma) = 1\text{-p. Recipe: } \mu \ge X\text{-}z\sigma. \tag{8.4.7}$$

$$\Pr\left(\frac{X\text{-}\mu}{\sigma} \ge z\right) = p. \text{ So } \Pr(X\text{-}\mu < z\sigma) = 1\text{-p. Recipe: } \mu > X\text{-}z\sigma. \tag{8.4.8}$$

The recipe arising at (8.4.7) places the limit inside the region. The recipe arising at (8.4.8) places the limit outside the region. A useful mnemonic arises:

> The exclusive definition of the P value leads to an inclusive region. The inclusive definition of the P value leads to an exclusive region.

Now, if p = V is attained at (8.4.7), then p = ε+V is attained using the same class mark z at (8.4.8). So let us take the co-ordinates given at (8.4.6) to have arisen as follows:

> (0.840, 0.159) ≈ (0.84, 0.16) by rounding (ε = 0.001).
> (0.980, 0.019) ≈ (0.98, 0.02) by rounding (ε = 0.001).

Bearing in mind that $\sigma = 1$, we now derive certain terms from each of the two systems of regions arising at (8.4.7) and (8.4.8), respectively. First we consider the system of regions arising from the exclusive definition and leading to the recipe at (8.4.7), thus obtaining the following two terms of that system:

> $\mu \in \{3, 4, 5, ...\}$ is a 0.841 inclusive region where the limit ($\mu = 3$) is *inside* the region, and p = 0.159 at the limit.
>
> $\mu \in \{2, 3, 4, ...\}$ is a 0.981 inclusive region where the limit ($\mu = 2$) is *inside* the region, and p = 0.019 at the limit. (8.4.10)

Next we consider the system of regions arising from the inclusive definition and leading to the recipe at (8.4.8), thus obtaining the following two terms of that system:

> $\mu \in \{4, 5, 6, ...\}$ is a 0.840 exclusive region where the limit ($\mu = 3$) is *outside* the region, and p = 0.160 at the limit.
>
> $\mu \in \{3, 4, 5, ...\}$ is a 0.980 exclusive region where the limit ($\mu = 2$) is *outside* the region, and p = 0.020 at the limit. (8.4.11)

The two systems are completely distinct; no term from the one system can be obtained from the other system. It follows from Section 8.3 that two consonance regions arise at (8.4.10) and two dissonance regions arise at (8.4.11) albeit that habitual thought has made the latter two regions arise awkwardly, as if expressing consonance. In order to clarify the matter we reformulate as follows: at (8.4.10) the two terms of the system are better expressed as:

> $\mu \in \{3, 4, 5, ...\}$ is a 1-p = 1 − 0.159 consonance region where p = 0.159 at the limit and p is larger for larger μ.
>
> $\mu \in \{2, 3, 4, ...\}$ is a 1-p = 1 − 0.019 consonance region where p = 0.019 at the limit and p is larger for larger μ. (8.4.12)

At (8.4.11) the two terms of the system are better expressed as follows:

> $\mu \in \{1, 2, 3\}$ is a 1-p = 1-0.160 dissonance region where p = 0.160 at the limit and p is smaller for smaller μ.
>
> $\mu \in \{1, 2\}$ is a 1-p = 1-0.020 dissonance region where p = 0.020 at the limit and p is smaller for smaller μ. (8.4.13)

It then appears that, in terms of the norms agreed upon, the P values at (8.4.12) point at the first of the two regions as the appropriate *consonance region*, and the P values at (8.4.13) point at the second of the two regions as the appropriate *dissonance region*. Thus the correct partition, as expressed at (8.4.5), is attained. However, it is attained by a means that contrives concomitantly to introduce unneeded concepts. The following two examples and a discussion of their import will help make that clear.

Example 8.4.1

An appropriate partition of the entries in Table 4.5.1 is {D, F, |C, A, B, E}, as can be seen directly by inspection of the co-ordinate trace given in the table. We must use the exclusive definition of the P value, so as to include F as the boundary of the 1-0.522 consonance region for the identity of the best entry. That leaves the status of C open to question. So we also have to use the inclusive definition of the P value, to include C as the boundary of the 1-0.006 dissonance region for the identity of the best entry.

Example 8.4.2

Two distinctly different, but both appropriate, partitions of the entries in Table 2.9.1 are {A, B, C, D, |E, F} and {A, B, C, D, E, | F}, as can be seen directly by inspection of the co-ordinate trace given in the table. To obtain the first partition we must use the exclusive definition of the P value, so as to include D as the boundary of the 1-0.126 consonance region for the identity of the best entry. That leaves the status of E open to question. So, we also have to use the inclusive definition of the P value for the first partition in order to include E as the boundary of the 1-0.055 dissonance region for the identity of the best entry. That still leaves the status of F open to question. So we again have to use the inclusive definition of the P value in order to include F as the boundary of a further 1-0.012 dissonance region for the identity of the best entry. The complement of this further dissonance region is an unwanted consonance region.

Discussion of the import of Examples 8.4.1 and 8.4.2

Both examples are of practical interest, but that is not the reason for introducing them here. They are introduced here because the partitions of interest are seen directly by inspection of the co-ordinate traces given in Tables 4.5.1 and 2.9.1, are indeed made glaringly obvious by those traces. So we are compelled to ask:

> Considering that in each case the given problem is from the very outset solved by inspection of the trace, what does a consonance-dissonance development then add to or subtract from that solution *as such*?

We are compelled to answer that it neither adds to nor subtracts from that solution *as such*. It merely contrives to express the solution in terms of coverage probabilities, which are formal probabilities that, as such, cannot be taken seriously in any case, lest we fall into a vicious circle of the kind developed in Section 4.8. So here again, just as in Chapter 6, 'probability inference' violates one of the most fundamental principles of science:

Never, ever introduce a constituent that is not needed. (8.4.14)

8.5 An unneeded embellishment

At (8.4.14) we paraphrased William of Occam's *principle of paucity*: 'Never try to do with more, what can be done with fewer.' Our customers in substantive science will know it as an indispensable principle. It is for instance the reason why the notion of intelligent design can find no place in science. It should also prevent various kinds of statistical regions (confidence regions, consonance regions, fiduciary regions, likelihood regions, credibility regions, etc.) from invading the discourse of investigative science. The following two examples and the ensuing discussion will help clarify this.

Example 8.5.1

Bliss (1967, p. 154) considers the survival times of 19 mice infected with tubercle bacilli. He notes that: 'From the evidence of this and other similar series, a suitable metameter is y = log days.' Of course he refers here to circumstantial evidence, data inspection and experience. The numerical data, in [ln(days)-1]×1 000 code, are given in Table 8.5.1 and are plotted against the expected values of corresponding normal-order statistics in Figure 8.5.1; the requisite order-statistical values are given by Harter (1961). Such a plot

Table 8.5.1: Survival times of 19 tubercle infected mice [ln(days)-1] × 1 000 code

Code	161	290	312	332	352	371	389	406	439	455
Frequency	1	1	1	3	1	3	5	2	1	1

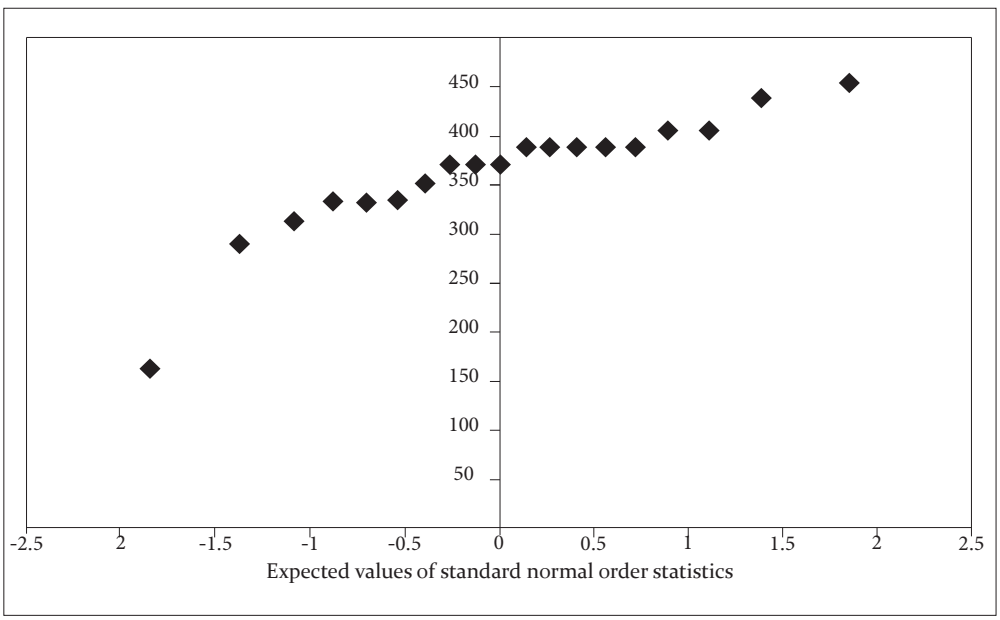


Figure 8.5.1: Order-statistical plot of the linearly coded survival times of 19 mice infected with tubercle bacilli

would tend to scatter round a straight line for sampling from a normal population. So the observed plot encourages one to model the data as

> '... a sample from a normal population, except for the earliest death as premature, that is to say, as an outlier'. (8.5.1)

Bliss uses significance tests to test more incisively whether or not the earliest death is indeed modelled satisfactorily as an outlier. But, inasmuch as those tests complement normal sampling, let us first test, more incisively, whether or not the model expressed in words at (8.5.1) can account satisfactorily for the 18 longer survival times. Using the facilities of D'Agostino (1970) and D'Agostino and Tietjen (1971), we then find:

> The mental correlate of Pearson's observed measure of skewness, $\sqrt{b_1} = -0.06612$, is situated at (0.420, ε, 0.580) in the test distribution. (8.5.2)

> The mental correlate of Pearson's observed measure of kurtosis, $b_2 = 2.623$, is situated at (0.496, ε, 0.504) in the test distribution. (8.5.3)

So the model expressed at (8.5.1), as tested, fits the 18 longer survival times very well. Turning now to whether the earliest death is to be modelled as an outlier, let us follow Bliss using the tests of Dixon (1950, 1951). If the observed values, arranged in order of magnitude and starting with the suspected outlier, are denoted by $y_1, y_2, y_3, \ldots, y_n$ the test datum is

$$\frac{y_2-y_1}{y_n-y_1} \text{ for } 3 \le n \le 8, \quad \frac{y_3-y_1}{y_{n-1}-y_1} \text{ for } 9 \le n \le 13, \quad \frac{y_3-y_1}{y_{n-2}-y_1} \text{ for } 14 \le n \le 30, \ldots$$

where the numerator, referred to as the *gap* to be tested, is taken to be the larger of two possible candidates in absolute value. For instance, in Table 8.5.1, the candidates are the earliest death with corresponding gap 312-161 = 151, and the latest death with corresponding gap 406-455 = -49. So the earliest death is identified as the suspected outlier. Using Dixon's gap test and prepared tables for the test distribution, we find that:

> The test datum is (312-161)÷(406-161) = 0.616, whose mental correlate is situated far to the right of (0.995, ε , 0.005*) in Dixon's test distribution.

Whenever an outlier is indicated, as in this case, we must test whether or not there might be another outlier. So we consider the n =18 remaining observations for which the two gaps are now 332-290 = 42 and 406-455 = -49, and where the latter is now the gap to be tested. We find that:

> The test datum is (406-455)÷(332-455) = 0.398, whose mental correlate is situated at approximately (0.906, ε, 0.094*) in Dixon's test distribution.

The fit is slightly awkward but tolerable. We note in passing that the current statistical literature would have us double the pointing co-ordinates obtained in any Dixon test if an outlier could occur at either extreme; however, a co-ordination tester must of course ignore that. Turning now to the various members of the class of models, each member comprising a sample of size n = 18 from an $N(\mu, \sigma^2)$ population and one outlier (the latter is part of each member model), we test for tenable values of μ treating σ^2 as

a nuisance parameter. The appropriate test statistic is Student's t as applied to the 18 longer survival times, and the mental correlate of the test datum is found in Student's test distribution at $(0.500, \bar{\varepsilon}, 0.500)$ for $\mu = \bar{x}$, the observed mean survival time, and

$$\begin{aligned}
&\text{at } (0.080, \varepsilon, 0.920) \text{ and } (0.920, \varepsilon, 0.080) \text{ for } \mu = \bar{x}\text{-}15 \text{ and } \mu = \bar{x}\text{+}15, \text{ respectively,}\\
&\text{at } (0.040, \varepsilon, 0.960) \text{ and } (0.960, \varepsilon, 0.040) \text{ for } \mu = \bar{x}\text{-}19 \text{ and } \mu = \bar{x}\text{+}19, \text{ respectively,}\\
&\text{at } (0.020, \varepsilon, 0.980) \text{ and } (0.980, \varepsilon, 0.020) \text{ for } \mu = \bar{x}\text{-}22 \text{ and } \mu = \bar{x}\text{+}22, \text{ respectively,}\\
&\text{at } (0.010, \varepsilon, 0.990) \text{ and } (0.990, \varepsilon, 0.010) \text{ for } \mu = \bar{x}\text{-}26 \text{ and } \mu = \bar{x}\text{+}26, \text{ respectively,}\\
&\text{at } (0.005, \varepsilon, 0.995) \text{ and } (0.995, \varepsilon, 0.005) \text{ for } \mu = \bar{x}\text{-}29 \text{ and } \mu = \bar{x}\text{+}29, \text{ respectively,}\\
&\ldots . \qquad (8.5.4)
\end{aligned}$$

Cox and Hinkley (1974) would have us report these same facts as follows in terms of consonance intervals, except that they would have us refer to the consonance intervals as confidence intervals:

$$\begin{aligned}
&\bar{x}\text{-}15 < \mu < \bar{x}\text{+}15 \text{ is a } 1\text{-}2(0.080) \text{ consonance interval for } \mu,\\
&\bar{x}\text{-}19 < \mu < \bar{x}\text{+}19 \text{ is a } 1\text{-}2(0.040) \text{ consonance interval for } \mu,\\
&\bar{x}\text{-}22 < \mu < \bar{x}\text{+}22 \text{ is a } 1\text{-}2(0.020) \text{ consonance interval for } \mu,\\
&\bar{x}\text{-}26 < \mu < \bar{x}\text{+}26 \text{ is a } 1\text{-}2(0.010) \text{ consonance interval for } \mu,\\
&\bar{x}\text{-}29 < \mu < \bar{x}\text{+}29 \text{ is a } 1\text{-}2(0.005) \text{ consonance interval for } \mu, \ldots . \qquad (8.5.5)
\end{aligned}$$

Note that the facts at (8.5.4) and (8.5.5) are one to one transforms of each other. An advocate of confidence intervals in the original sense intended by Neymann, as explained in Chapter 4, would have us specify *without reference to the data* a Type I error rate, say $\alpha = 0.020$ for argument's sake, and then dichotomise the parameter space as follows:

> We must, *as an act of will,* conclude that the true value of μ can be, and can only be, one of those in the interval $\bar{x}\text{-}22 < \mu < \bar{x}\text{+}22$, where such conclusions are true in $1\text{-}2(0.020)$ of cases, i.e. 96% of cases. (8.5.6)

Despite continuing to use the term 'confidence', as not intended by Neymann, rather than following the usages proposed by Kempthorne and Folks (1971), Cox and Hinkley (1974) would nevertheless object (and rightly so) to the idea exemplified in (8.5.6). They say for instance (pp. 207-208): '... in general, interval estimates cannot be taken as probability statements about parameters, and foremost is the interpretation "such and such parameter values are consistent with the data"'. This is precisely why Kempthorne and Folks characterise an acceptable model as 'consonant with the data'. The advocates of consonance intervals are entirely correct in rejecting Neymann's notions of specifying error rates *without reference to the data*, and of performing *acts of will*, where instead an understanding of factual evidence is required. So, we have to concede that the presentation of factual evidence at (8.5.5) serves investigative needs better than does the reasoning at (8.5.6) because the latter has mistaken a problem in investigation for a problem in forecasting. But, having agreed to that, and inasmuch as the presentations at (8.5.4) and (8.5.5) are one to one transforms of each other, we must ask why advocates of consonance intervals employ a presentation that, as at (8.5.5), courts the very misunderstanding they wish to avoid. The answer is simply that we are as much the victims of our education, as we are its beneficiaries. When Cox and Hinkley explain (p. 209) that the coverage probability is a 'key requirement' that 'gives a physical interpretation to the confidence limits' because, though it is only 'a hypothetical statement', it 'gives an empirical meaning, which in principle can be checked by experiment', they have

fallen victim to their education; it does not occur to them that for a statistical statement to have a physical meaning, does not imply that it must necessarily take the form of a forecasted frequency 'which in principle can be checked by experiment'. The presentation at (8.5.4) comprises statements whose meanings can by simulation be forced onto the human body – not just in principle, but by simulation that provides the *actual* evidence. It should be noted, however, that significance testers are justified in being leery of Neymann's reasoning, that is to say, of reasoning that interprets the coverage probabilities as real-world probabilities. In order to grasp that we merely have to note that whereas at (8.5.4) it is *explicitly* the case that a very wide variety of models fit the given data well, at (8.5.2) and (8.5.3) it is *implicitly* the case that there too a very wide variety of models would fit the given data well. So, to think that any Type II-like errors at (8.5.2) and (8.5.3) can be made to have zero probability by means of reasonable assumptions or acts of will, would be extremely unrealistic. Here the reader should note that the assumptions seem to be reasonable because there is nothing in evidence to make us think otherwise. So the crux of the argument is simply this: Why court a fallacy by the introduction of an assumption that is not needed?

Example 8.5.2

The differences in yield between the n = 14 sprayed strips of corn and their unsprayed controls, given in Table 2.4.1, might at first seem to present a very similar problem to the one we dealt with in the previous example, but it turns out to be fundamentally different. If we proceed, without forethought, to test for outliers as in the previous example, we find that Dixon's test points strongly at the largest difference, 24.0, and at the smallest difference, -5.7, as outliers, whilst for the second largest difference, 8.8, we find the mental correlate of the test datum placed at approximately (0.94, ε, 0.06*) in Dixon's test distribution. However, given the purposes of the experiment, we have to deal with the mean of all 14 differences, because there is no reason to suspect that the outliers do not measure actual differences in yield satisfactorily. Also note that inasmuch as the differences are from 14 different farms, the outliers represent an understandable heterogeneity. In this respect the experiment resembles *a survey*, suggesting that we might simply judge the observed mean difference in yield, 4.70 bu./acre, in relation to its estimated standard error, 1.73 bu./acre. However, the data are from *an experiment* that was conducted by means of a properly randomised design. And we also know that Student's *t* test for a mean of paired differences is remarkably robust. So it was reasonable for Snedecor to consider the use of Student's *t* in this case. What must concern us here is whether an interval estimate, when it is adjoined to an informative co-ordinate trace, contributes positively, vacuously, or negatively to whatever the trace has already supplied for investigative needs. The trace of present interest is given by the following, where δ denotes the effect of the spray, the D-like symbol denotes conceptual difference in yield, the d-like symbol denotes real difference in yield, and the notation is otherwise obvious:

$$\left[\Pr\left(t < \frac{\bar{d}\text{-}\delta}{s_{\bar{D}}}\right), \varepsilon, \Pr\left(t > \frac{\bar{d}\text{-}\delta}{s_{\bar{D}}}\right)\right] \text{ for } \bar{d} = 4.70,\ s_{\bar{D}} = 1.73, \text{ and for } 0 \le \delta < \infty.$$

Hence the mental correlate of the test datum is situated in Student's test distribution

at (•, ε, 0.009) for δ = 0.0 bu./acre,
at (•, ε, 0.026) for δ = 1.0 bu./acre,
at (•, ε, 0.071) for δ = 2.0 bu./acre,
at (•, ε, 0.172) for δ = 3.0 bu./acre, (8.5.9)

and the co-ordinates for any other δ values that might be of interest may of course be adjoined to this abbreviated trace. We have seen that the reasoning leading up to this trace is by rough and ready approximation. Nevertheless, the trace can hardly be said to be uninformative, as the error estimate is appropriately modelled as inflated. So the trace clearly shows that the models indexed by values of the order $\delta = 2.0$ bu./acre or less, fit the data poorly, requiring larger values for satisfactory fit. Now let us consider the corresponding consonance regions. They are of course one sided. But inasmuch as the notion of a two-sided region is *entailed* by the more basic notion of a coverage probability, it is appropriate to evaluate here just the more basic notion. So, by a one to one transformation of the abbreviated trace at (8.5.9) we obtain the following four terms of a system of one-sided consonance intervals:

0.0 bu./acre $\leq \delta$ is a 1-0.009 consonance region for δ.
1.0 bu./acre $\leq \delta$ is a 1-0.026 consonance region for δ.
2.0 bu./acre $\leq \delta$ is a 1-0.071 consonance region for δ.
3.0 bu./acre $\leq \delta$ is a 1-0.172 consonance region for δ. (8.5.10)

If we select, *without reference to the data*, just one such consonance region it would be interpretable as a confidence region, but as explained in Chapter 4, we then fall into circular reasoning. Hence, as explained in Chapter 5, an advocate of consonance would have us avoid interpreting a consonance coefficient as an *actual* (real-world) probability and would instead have us interpret the coefficient as a measure of fit that, for want of a better *form*, is expressed as a *formal* (conceptual) probability. However, if we *do* agree to that interpretation, then whatsoever is conveyed at (8.5.10) *has already been conveyed* at (8.5.9), which means that we are trying to 'do with more, what can be done with fewer'.

8.6 The notion of 'interval estimation' in general

In the foregoing, confidence intervals and consonance intervals provided sufficient exemplification of the general idea of interval estimation. Further exemplification is not needed, as it suffices to note that the general idea is that a coverage probability must characterise the tenability of many parameter values comprising an interval (or region) of such values. This is bound to have silly consequences, as those values are not equally tenable. Typically, values on or near the boundary of the interval are less tenable than those in or near the centre of the interval. Seemingly, this defect might be removed whilst retaining the notion of coverage probability, by a series of nested intervals with a range of different coverage probabilities. However, each one of the nested intervals by itself residually retains the defect. So, the only way to remove the defect entirely is to utterly abandon the notion of coverage probability in favour of descriptions of the tenability of the individual parameter values. In retrospect we see that the idea of interval estimation contrives to somehow accommodate the *idée fixe*, and it thereby introduces a constituent that is not needed – and that is bad science. To prove that that is indeed bad science is very easy. All one has to do, is to show how, under certain circumstances, the idea inexorably leads to the replacement of one-sided intervals by two-sided intervals, and to show how that amounts to replacement of each test in an array of co-ordination tests, with a corresponding co-ordination test whose separating characteristics are uniformly inferior to those of the test it replaces.

Chapter 9

ANCILLARY STATISTICS

Selecting the correct frame of reference for data analysis

9.1 Introduction

In Section 1.5 we met the mixed sampling problem. One version of the problem is as follows: the outcome of a Bernoulli trial, A, determines one or the other of two further Bernoulli trials, B and C, with equal probability. The conditional probability of success for the further trial is 0.2 for B, and 0.8 for C. We ask:

> For B arising from A, must the probability of success be taken to be conditionally 0.2, or taken to be unconditionally 0.5?
>
> For C arising from A, must the probability of success be taken to be conditionally 0.8, or taken to be unconditionally 0.5? (9.1.1)

These questions are meaningful *in the logic* of mathematical statistics without having to resort to anything beyond that logic; yet they *cannot be answered in that logic* and thus exemplify *the incompleteness* of that logic. Each question posits, as possible, one or the other of two different frames of reference, and asks for 'the correct one of the two' to be supplied, where that can only be supplied by substantive science. These facts can be forced upon us in that, depending on substantive science, any one of three different frames of reference might in its own right be the correct one, as shown in Example 9.1.1.

Example 9.1.1

A decision-maker must classify each one of a host of many items as 'good' (send it off to market) or as 'bad' (send it off to waste). A regulatory authority allows at most 5% of marketed items to be bad. Machine A misclassifies only 1% of items, but is more expensive to use than Machine B, which misclassifies 9% of items. So, for each item in turn, the decision-maker flips an unbiased coin to determine which machine makes the classification, and then erases any record of which machine was used (because otherwise the regulatory authority would rule out any items classified by machine B). So the decision-maker's intention is that the items sent off to market will constitute a host of items of which 5% are bad. However, in a particular case, a customer purchases an item and discovers that by oversight a mark, showing that that item was classified by machine B, was not erased. So, supposing the customer understands the significance of the mark, it would require a rather stupid customer not to realise the advantage of demanding a replacement from the pool of unmarked items. Thus the correct frame of reference is that *that particular item is known to belong* to a population where 9% of items are bad. Similarly, if oversight left a mark showing that a particular item was classified by machine A it would require a rather stupid customer not to realise the disadvantage, in such a case,

of demanding a replacement from the pool of unmarked items. Thus the correct frame of reference is then that *that particular item is known to belong* to a population of which 1% of items are bad. The reader should note also that for any particular item, whenever the tell-tale mark has been erased, we would *know* that *that particular item* belongs to a population where 5% of the items are bad.

Several remarks are in order:

(1) The example is not intended to exemplify problems of practical importance. It is a paradigmatic example; it is intended to display, in the simplest possible terms, certain principles of reasoning. The paradigm is a good one inasmuch as it compels us to recognise from the outset that in each of the three possible cases there can, in respect of the particular item being pointed at, be only one satisfactory answer as to what is *known* about that *particular* item.

(2) It is crucially important to grasp (and this is difficult to the extent that we have to keep returning to it) that although the questions at (9.1.1) display, in mathematical logic, the possibility of three distinctly different frames of reference, they do not – and they cannot – supply in that logic, any principle or rule by which they can be answered. We have to note (and this is difficult) that whether *a particular item* is marked A, is marked B, or is unmarked, refers to *the substance* (the bodily experiences) of the matter; it is not part of *the logic* (the mathematical statistics) of the matter.

(3) The decision-maker's intention *to create* to specification *a host of many items* is of trivial importance. Since we understand it completely, there is nothing further to explain or to disagree about. The thrust of the example is what we *come to know* about how, in any particular instance, *a single solitary item* has come about. Hence, here and in the rest of this chapter, we are concerned with a problem in *investigative* statistics. This point is crucial, because the problem of ancillary partitions is one of a number of problems in investigative statistics that are aggravated by the ubiquitous use of a certain device in statistical teaching. Typical examples of that device are given below.

> **Question:** To what does the standard error of the mean refer?
> **Answer:** Suppose you were to do this over again, many times, and to calculate the mean in each case. Then the standard deviation of the population comprising those many means will be ... .
>
> **Question:** To what does the Type I error rate refer?
> **Answer:** Suppose you were to do this over again, many times, and to employ this accept-reject rule in each case. Then the long-run frequency of Type I errors will be ...
>
> **Question:** To what does confidence coefficient refer?
> **Answer:** Suppose you were to do this over again, many times, and to use this recipe for an interval estimate in each case. Then the proportion of intervals that include the true value will be ...

There is an inadvertent miss-education imbedded in the phrase 'suppose you were to do this over again'. The phrase conditions one to the 'frequentist' notion of 'statistical evidence' as a *forecast* of what '*will be*', which notion is wrong; whatsoever evidence

we can point at cannot belong to what 'will be' *in the future*. So it must be very firmly grasped that an investigator of given data cannot 'do this over again'. Investigation of any particular data set must always ask how *this single, solitary, real-world individual might have come about*, and one cannot properly answer such a question by turning a blind eye to anything that is known about *that particular individual*. In short, we have to proceed from Definition 1.2.1, and not from Definition 1.2.2, because investigative problems in science belong to the discourse of *predicting and pointing*; such problems do not belong to the discourse of *predicting and forecasting*.

9.2 A PROPOSAL THAT FAILS

An *ancillary statistic* is a function of the minimally sufficient statistic for the index of a class of models, but is independent of that index. By thus restricting the definition to that minimal sufficient statistic, the concept of 'an ancillary statistic' is prevented from being confused with that of 'a class characteristic'. The concept 'an ancillary statistic' is due to R.A. Fisher. He noticed that amongst the simplest examples of such statistics are indicators of sample size. For instance, imagine flipping two unbiased coins successively and letting the outcomes in terms of heads (H) or tails (T) determine the size of samples to be drawn from some or other population of interest, as follows:

> For outcomes (H, H), (H, T), (T, H), and (T, T),
> draw samples of size $n = 1$, $n = 10$, $n = 100$, and $n = 1000$, respectively.

Define a statistic,

> Y = 1, 2, 3, 4, for outcomes (H, H), (H, T), (T, H), (T, T), respectively.

Then Y is an indicator of sample size, and is ancillary, since its distribution, given by

> $\Pr(Y = y) = \frac{1}{4}$ for y = 1, 2, 3, 4,

is independent of any parameter, θ, that might index the members of a class of models for the sample in question. If the variance of the population is given by σ^2, what is the variance of the sample mean?

> Is it $\sigma^2 \div n$ for *the* particular n?
>
> Or is it $\frac{1}{4}(\sigma^2 \div 1)+\frac{1}{4}(\sigma^2 \div 10)+\frac{1}{4}(\sigma^2 \div 100)+\frac{1}{4}(\sigma^2 \div 1\,000)$ for *any* particular n?

In the discourse of *investigative* statistics, the only sensible answer to this question is:

> $\sigma^2 \div n$ for $n = 1$, or 10, or 100, or 1 000, depending on the *particular* data set.

So here the correct frame of reference is conditional on the ancillary. Returning to Example 9.1.1 we find that there too the correct frame of reference for any *particular* marked individual is conditional on an ancillary, as follows: if X = 0 or 1 denotes the state of a random item as good or bad, respectively, X is a Bernoulli variable with $\Pr(X = 1) = \theta$ say $(0 \le \theta \le 1)$. So if Y = A or B then denotes the use of machine A or B, respectively, and

Z = 0 or 1 denotes correct classification or misclassification, respectively, the sample space is given by the following possible values of (X, Y, Z)

(0, A, 0), (0, A, 1), (1, A, 0), (1, A, 1), (0, B, 0), (0, B, 1), (1, B, 0), (1, B, 1),

whose probabilities are:

$(1-\theta)½(0.99)$, $(1-\theta)½(0.01)$, $(\theta)½(0.99)$, $(\theta)½(0.01)$, ..., $(\theta)½(0.91)$, $(\theta)½(0.09)$

respectively, where Pr(Y = A) and Pr(Y = B) are of course given by:

$(1-\theta)½(0.99) + (1-\theta)½(0.01) + (\theta)½(0.99) + (\theta)½(0.01) = ½$,

$(1-\theta)½(0.91) + (1-\theta)½(0.09) + (\theta)½(0.91) + (\theta)½(0.09) = ½$, respectively.

Thus, Y is ancillary, and by conditioning on Y = A, or on Y = B, we obtain the correct frame of reference for *a particular item* marked A, or *a particular item* marked B, respectively. For *a particular item* that is unmarked the relevant ancillary is the entire sample space and the empty set, with probabilities 1 and 0, respectively (independent of θ). Another example where conditioning on ancillary statistics corresponds to the correct frame of reference for investigative statistics is developed by Basu (1964), as follows: the variance of the mean of a random sample drawn *with replacement* from a finite population of values with variance σ^2, is given by $\sigma^2 \div n$, where n denotes sample size. However, owing to the population being finite, some items might then be drawn more than once, where, of course, investigative statistics must ignore any such replicates as additionally uninformative. Since the number of times any particular item is drawn does not depend on its value, or on that of any of the other N-1 items, replicate items are accounted for by ancillary statistics (see for example Table 9.2.1). Basu shows that by ignoring any such

Table 9.2.1: An example of sample size as an ancillary statistic. $\{X_1, X_2\}$ sampled at random with replacement from $\{x_1, x_2, x_3\}$. The sample space comprises 3×3 = 9 equally likely cases.

X_1	x_1	x_1	x_1	x_2	x_2	x_2	x_3	x_3	x_3
X_2	x_1	x_2	x_3	x_1	x_2	x_3	x_1	x_2	x_3
n	1	2	2	2	1	2	2	2	1
$\Pr(n = 1) = 1/3$, $\Pr(n = 2) = 2/3$, regardless of the values x_1, x_2, x_3. If one conditions on n, sampling is without replacement; then the sample space comprises 3-choose-n equally likely cases.									

replicates we are in effect conditioning on those ancillaries. So, if m of the N different items are drawn ($m \le n \le N$) their values comprise, in terms of the conditioning, a random sample of size m drawn *without replacement* from the population, where the variance of the mean of such a sample is given by the well-known formula:

$$\frac{\sigma^2}{m}\left[1-\frac{m}{N}\right]. \qquad (9.2.1)$$

Yet another case of the correct frame of reference for investigation being obtained by conditioning on an ancillary statistic was developed in Sections 4.14, 4.15, 4.16 and 4.17. Various such examples caused Fisher (e.g. 1935, 1936, 1973) to propose that the correct frame of reference for investigative statistics in the presence of any ancillary statistics will, as a general principle, *always* be found by conditioning on the ancillary statistics.

He was mistaken.

He was trying to make a logical principle supply the answer to a substantive problem, and thus mistook a problem in substantive science for a problem in logical reasoning. This runs counter to Gödel's incompleteness principle, and therefore cannot be viable. That the proposal is not viable was uncovered and demonstrated by Basu (1964) using remarkably simple paradigms involving dice.

Example 9.2.1

Let X model the outcome of just one roll of a biased tetrahedral die, as in Table 9.2.2.

Table 9.2.2: A class of models for the outcome of rolling a biased tetrahedral die.

Outcome X (value on bottom)	X = 1	X = 2	X = 3	X = 4
Probability of outcome X	$\frac{1-2\theta}{4}$	$\frac{1-\theta}{4}$	$\frac{1+\theta}{4}$	$\frac{1+2\theta}{4}$
Ancillary label Y	Y = 1	Y = 2	Y = 2	Y = 1
Maximum likelihood estimate $\hat{\theta}$	-½	-½	+½	+½
$-\frac{1}{2} < \theta < +\frac{1}{2}$				

In order to convey the import of this example, we have to ensure that the *substantive* context of the problem is firmly grasped. So we suppose that the biased die was made by covering the four different sides of an unbiased tetrahedral die with appropriately weighted triangular sheets, without distorting its tetrahedral shape. And we suppose also that the biased die is rolled in the usual way. X is minimally sufficient for θ, the index of the class of models displayed in Table 9.2.2. The table displays an ancillary statistic Y. We note that the maximum likelihood estimator of θ is insufficient for θ. However, a one to one transform of the minimal sufficient statistic for θ is given by

$$(Y, \hat{\theta}), \text{ where } \hat{\theta} \text{ denotes the maximum likelihood estimator of } \theta. \tag{9.2.2}$$

To see that this is a one to one transform, note that for X = 1, 2, 3, 4, Table 9.2.2 shows that:

$$(Y, \hat{\theta}) = (1, -\tfrac{1}{2}), (2, -\tfrac{1}{2}), (2, +\tfrac{1}{2}), (1, +\tfrac{1}{2}), \text{ respectively,} \tag{9.2.3}$$

and *vice versa*. So, by expressing the minimal sufficient statistic in the form at (9.2.2), the expressions at (9.2.3) and the probabilities in Table 9.2.2 show that the minimal sufficient statistic for θ, conditional on Y = 1, is given by $(1, \hat{\theta})$ for

$$\hat{\theta} = -\tfrac{1}{2} \text{ or } +\tfrac{1}{2} \text{ with probabilities } \frac{1-2\theta}{2} \text{ or } \frac{1+2\theta}{2}, \text{ respectively,} \tag{9.2.4}$$

and the minimal sufficient statistic for θ, conditional on Y = 2, is given by $(2, \hat{\theta})$ for

$$\hat{\theta} = -½ \text{ or } +½ \text{ with probabilities } \frac{1 - \theta}{2} \text{ or } \frac{1 + \theta}{2}, \text{ respectively,} \qquad (9.2.5)$$

However, Basu would object (rightly so) to using the model at (9.2.4) or the model at (9.2.5) to represent how, by rolling the biased tetrahedral die, any particular outcome was brought about, because (so he would object) each conditional model requires the rolling of the die to have been such that the ancillary would take its observed value, but that could not have been the case because the die was rolled in the usual way, that is to say, we placed the die in a cup, we shook the cup, and we tossed the die from the cup. Basu calls that 'a performable experiment', whereas rolling the die such that Y = 1, as required by the model given at (9.2.4), or such that Y = 2, as required by the model given at (9.2.5), are 'non-performable'. In order to make this entirely clear, Basu also considers as follows, how models such as those at (9.2.4) and (9.2.5) might represent 'performable experiments': Suppose the substantive investigator has two bent coins, each of which is

> marked $\theta = -½$ on one side, and marked $\theta = +½$ on the other side,

and suppose that when Coin 1 is flipped, the probabilities are as given at (9.2.4), and when Coin 2 is flipped, the probabilities are as given at (9.2.5). If the investigator then flips an unbiased coin to determine which one of the two bent coins is to be flipped, a performable experiment has been performed, and its outcome would validly be represented by a performable subsidiary experiment corresponding to one or the other of the two conditional models at (9.2.4) and (9.2.5). But that is not at all the case for the tetrahedral die because, as Basu (rightly) observes, the investigator had 'a die to experiment with, but where are the coins?' So he concludes that the 'trouble lies in our failure to recognise the difference between a real (performable) and a conceptual (non-performable) statistical experiment'. In short: Basu holds that if conditioning on an ancillary statistic fails to pick out any performable subsidiary experiment, we must ignore the ancillary as substantively vacuous for investigative purposes. In the present example he would have us consider Y to be substantively vacuous, and so would have us use the unconditional model. We subsequently find the requirement 'performable' somewhat too restrictive; but for the time being that need not concern us.

9.3 Conditional models that are 'performable', even 'performed', yet vacuous

Any consulting statistician learns to distrust a substantive investigator's understanding of the protocol that produces a randomised design, and so learns that instead of describing the protocol, it is wiser to supply the field plan. So we can well imagine the following example.

Example 9.3.1

A substantive investigator knows that statisticians 'are inclined to make a fuss' about failure to randomise, and tries to make 'doubly sure' that 'the completely randomised design', as instructed, is used. So, through muddled understanding, the substantive investigator divides 15 units at random into three 'blocks' of five units each, and then randomly

assigns, separately for each block, each of just five treatments to a different unit of that block. The statistician might at first be dismayed. But on second thoughts would be relieved to realise that after all the completely randomised design *was* used, as the muddled procedure replicates each of the field plans of the completely randomised design exactly $(3!)^5$ times, as follows:

> The procedure used generates one of
>
> $$\binom{15}{5}\binom{10}{5}\binom{5}{5}$$
>
> sets of blocks, and assigns one of 5!5!5! treatment patterns to that set, which amounts to altogether 15! possible field plans involving each appropriate plan just $(3!)^5$ times as follows: the number of possible field plans for a completely randomised design is given by
>
> $$\binom{15}{3}\binom{12}{3}\binom{9}{3}\binom{6}{3}\binom{3}{3} = \frac{15!}{3!12!} \times \frac{12!}{3!9!} \times \frac{9!}{3!6!} \times \frac{6!}{3!3!} \times \frac{3!}{3!0!} \text{ amounting to } \frac{15!}{(3!)^5} \text{ field plans}$$
>
> where $(3!)^5$ is the number of ways in which the units of any given one of those field plans can be grouped into three complete blocks, i.e., $(3)^5$ ways to form the first block, times $(2)^5$ ways to form the second block, times $(1)^5$ ways to form the third block.

This example shows that 'a performable subsidiary experiment', and even 'an actually performed subsidiary experiment', might, with good reason, be ignored as one that is substantively vacuous. So we are compelled to interpret Basu's position as demanding 'a performable subsidiary experiment' if and only if that subsidiary is substantively non-vacuous. This example also shows that every-day statistical practice often involves models that embrace several, even a great many, different ancillary statistics, which suggests that statistical practice already has an intuitive understanding of how to deal with such ancillary statistics. So, when considering purely *mathematical* theories of how we are to deal with an ancillary statistic, it is advisable to keep our feet firmly on *substantive* ground, which is precisely what Basu (1964) would have us do.

9.4 Another proposal that fails

Cox (1971) has proposed, at least in certain cases, to disagree with Basu (1964). In order to explain this, he considers a simplified version of Basu's main example. We now present the simplified example, first as Basu would have us view it, next as Cox would have us view it, and we then consider how we ought to view it.

Example 9.4.1

Let X model the outcome of just one roll of a biased tetrahedral die, as in Table 9.4.1.

Table 9.4.1: A class of models for the outcome of rolling a biased tetrahedral die

Outcome X (value on bottom)	X = 1	X = 2	X = 3	X = 4
Probability of outcome X	$\frac{2-\theta}{6}$	$\frac{1-\theta}{6}$	$\frac{1+\theta}{6}$	$\frac{2+\theta}{6}$
Ancillary label Y_1	$Y_1 = 1$	$Y_1 = 2$	$Y_1 = 2$	$Y_1 = 1$
Ancillary label Y_2	$Y_2 = 1$	$Y_2 = 2$	$Y_2 = 1$	$Y_2 = 2$
Maximum likelihood estimate $\hat{\theta}$	-1	-1	+1	+1
$-1 < \theta < +1$				

In order to convey the import of the present example we must, as we did in Example 9.2.1, make sure that the *substantive* context of the problem is clearly understood. So we suppose that the biased die is carefully made by covering the sides of an unbiased die with appropriately weighted triangular sheets, without distortion of the tetrahedral shape of the die. We also suppose that the resulting biased die is rolled in the usual way, i.e. we put the die in a cup, shake the cup, and toss the die from he cup. Table 9.4.1 displays two different ancillary statistics denoted by Y_1 and Y_2, respectively, and also shows that the maximum likelihood statistic is insufficient for θ. In this example, the minimal sufficient statistic, X, may alternatively be represented by $(Y_1, \hat{\theta})$, which is a one to one transform of X, or by $(Y_2, \hat{\theta})$, which is also a one to one transform of X. If we condition on the value of Y_1, a pair of formal models arise, as follows:

$\hat{\theta}$ = -1 or +1 with probabilities $\frac{2-\theta}{4}$ or $\frac{2+\theta}{4}$, respectively, when $Y_1 = 1$, and

$\hat{\theta}$ = -1 or +1 with probabilities $\frac{1-\theta}{2}$ or $\frac{1+\theta}{2}$, respectively, when $Y_1 = 2$. (9.4.1)

If we condition on the value of Y_2, a different pair of formal models arise, as follows:

$\hat{\theta}$ = -1 or +1 with probabilities $\frac{2-\theta}{3}$ or $\frac{1+\theta}{3}$, respectively, when $Y_2 = 1$, and

$\hat{\theta}$ = -1 or +1 with probabilities $\frac{1-\theta}{3}$ or $\frac{2+\theta}{3}$, respectively, when $Y_2 = 2$. (9.4.2)

Basu's view

All of the reasoning of Example 9.2.1 repeats for the pair of models at (9.4.1), and all of the reasoning of Example 9.2.1 repeats for the pair of models at (9.4.2). Moreover, the present example shows that Fisher's proposed rule, i.e. 'condition on the value of the ancillary', cannot, in general, simply be applied because there might be, as is the case here, different ancillaries involving contradictorily different conditional models. It follows that for Fisher's rule to be applicable, such non-uniqueness must be removed, i.e. it must somehow be made possible to pick out an ancillary that is uniquely 'the correct one' on which to condition. It will prove to be worthwhile to adapt Basu's bent-coin examples in order to

show how the conditional models obtained at (9.4.1) and (9.4.2) might, in a setup different to the biased-die setup, represent performable experiments. So, suppose an investigator has two pairs of bent coins:

a Y_1 pair named (1, 4) and (2, 3), respectively, and
a Y_2 pair named (1, 3) and (2, 4), respectively,

where these coins might be flipped with possible outcomes as follows:

(1, 4): outcome X = 1 or X = 4 with probabilities $\frac{2-\theta}{4}$ and $\frac{2+\theta}{4}$, respectively.

(2, 3): outcome X = 2 or X = 3 with probabilities $\frac{1-\theta}{2}$ and $\frac{1+\theta}{2}$, respectively.

(1, 3): outcome X = 1 or X = 3 with probabilities $\frac{2-\theta}{3}$ and $\frac{1+\theta}{3}$, respectively.

(2, 4): outcome X = 2 or X = 4 with probabilities $\frac{1-\theta}{3}$ and $\frac{2+\theta}{3}$, respectively.

In order to investigate the value of θ, the investigator flips an unbiased coin to choose between the Y_1 pair and the Y_2 pair. If the Y_1 pair is chosen the investigator rolls a six-sided unbiased die, flips coin (1, 4) if the die yields 1, 2, 3 or 4, and flips coin (2, 3) if the die yields 5 or 6. If the Y_2 pair is chosen, the investigator flips the unbiased coin to choose between flipping coin (1, 3) and flipping coin (2, 4). This procedure amounts to a performable experiment with performable subsidiaries named Y_1 and Y_2, respectively. Each subsidiary in turn comprises further subsidiaries performable by flipping coin (1, 4) or coin (2, 3) in the case of Y_1, or else by flipping coin (1, 3) or coin (2, 4) in the case of Y_2, all of which together have the following possible outcomes:

$$\text{X = 1 with probability } \left(\frac{1}{2}\right)\left(\frac{2}{3}\right)\left(\frac{2-\theta}{4}\right)+\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{2-\theta}{3}\right)=\frac{1}{6}\,(2-\theta)$$

$$\text{X = 2 with probability } \left(\frac{1}{2}\right)\left(\frac{1}{3}\right)\left(\frac{1-\theta}{2}\right)+\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1-\theta}{3}\right)=\frac{1}{6}\,(1-\theta)$$

$$\text{X = 3 with probability } \left(\frac{1}{2}\right)\left(\frac{1}{3}\right)\left(\frac{1+\theta}{2}\right)+\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1+\theta}{3}\right)=\frac{1}{6}\,(1+\theta)$$

$$\text{X = 4 with probability } \left(\frac{1}{2}\right)\left(\frac{2}{3}\right)\left(\frac{2+\theta}{4}\right)+\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{2+\theta}{3}\right)=\frac{1}{6}\,(2+\theta) \qquad (9.4.3)$$

In Basu's view any frame of reference is vacuous if non-performable. In the case of the tetrahedral die, Y_1 and Y_2 are indicators of non-performable subsidiaries whose existence must therefore be ignored as being vacuous, and the data must be analysed in terms of the unconditional model. Suppose for instance the data are X = 3. Then:

For X = 3 obtained by rolling the die, the maximum likelihood estimate is given by $\hat{\theta} = +1$, the situations of whose mental correlate can, from Table 9.4.1, be traced at

$$\left[\left(\frac{2-\theta}{6}\right)+\left(\frac{1-\theta}{6}\right), \left(\frac{1+\theta}{6}\right)+\left(\frac{2+\theta}{6}\right), \bullet\right], \text{ i.e. at}$$

$$\left(\frac{3-2\theta}{6}, \frac{3+2\theta}{6}, \bullet\right) \text{ for } -1 \leq \theta \leq +1. \qquad (9.4.4)$$

In case of the bent coins, however, we must condition on the performable initial flip of the unbiased coin choosing a bent-coin pair. We must then condition on the performable choice between the members of the chosen pair, and so must analyse the given data in terms of the performable flipping of that chosen member. Then:

For X = 3 obtained by flipping coin (1, 3), the maximum likelihood estimate is given by $\hat{\theta} = +1$, the situations of whose mental correlate can, from the first of the two conditional distributions given at (9.4.2), be traced at

$$\left(\frac{2-\theta}{3}, \frac{1+\theta}{3}, \bullet\right) \text{for } -1 \leq \theta \leq +1. \qquad (9.4.5)$$

The traces at (9.4.4) and (9.4.5) are distinctly different. For instance, consider testing $\theta = 0$. They differ because they arise from different performable experiments.

Cox's view

Cox holds that 'It is fairly compelling that the probability distributions used in inference should be conditional on the observed value of the ancillary statistic', but recognises that 'there may be alternative ancillary statistics for the same problem', for instance, Y_1 and Y_2 in the present example. He is evidently aware that such 'non-uniqueness' is resolved by Basu's requirement that we may condition in terms of a performable experiment only, but he regards that as 'tentatively suggested' only. He notes that Barnard and Sprott (1970) claim to have 'resolved the non-uniqueness of Basu's main example by appeal to invariance under a natural group of transformations', but he proposes to outline a resolution 'that is more generally applicable than an invariance argument'. (The proposal of Barnard and Sprott, which we consider in a subsequent section, does not resolve the non-uniqueness of Y_1 and Y_2 in the present example.) Turning now to the present example, Cox notes that by conditioning on Y_1 or on Y_2, 'the data are considered as two independent binomially distributed observations, the associated probabilities in the binomial distributions being respectively' as at (9.4.1) and (9.4.2). He considers testing the consistency of any given data with 'an arbitrary value $\theta = \theta_0$ by examining the exact distribution of the efficient score statistic', i.e. of the partial derivative of the log-likelihood at $\theta = \theta_0$. 'In this way,' he says, 'confidence regions for θ can be formed having at least a local optimum property.' He therefore considers the distributions of the score statistic when evaluated conditionally on Y_1 and Y_2, respectively, and he measures the dispersion of each of the two distributions by way of the variance of the information expected in each case. It turns out that

$$\text{the variance} = \frac{2}{[(1-\theta_0^2)\,(4-\theta_0^2)]^2} \quad \text{when conditioning on } Y_1, \text{ and} \tag{9.4.6}$$

$$\text{the variance} = \frac{\theta_0^2}{[(1-\theta_0^2)(4-\theta_0^2)]^2} \quad \text{when conditioning on } Y_2\,. \tag{9.4.7}$$

Cox points out that the more dispersed the distribution of the score statistic is under the ancillary variation, the more informative the conditioning is on that ancillary; that the dispersion, as measured, is greater at (9.4.6) than at (9.4.7); and that this is uniformly true for *all* values of θ_0. So he concludes that we must condition on Y_1. He also shows that these results are generalised to an arbitrary multinomial distribution with a scalar parameter, and in the case of a vector parameter the measures at (9.4.6) and (9.4.7) can be replaced by, for instance, the variance of the information determinant.

But look at what this leads to! There seems to be nothing in Cox's proposal to prevent us from applying it to the bent-coin example. So suppose, for the moment, that we may do so. We saw at (9.4.3) how the bent-coin experiment then leads to the model displayed in Table 9.4.1, which is in fact the model Cox considers, and for which he would have us analyse any given data conditionally on Y_1, rather than on Y_2. Cox would thus have us model the given datum, X = 3, as if it were obtained via the more informative one of the two model-pairs at (9.4.1) and (9.4.2), that is to say, he would in effect have us reason *as if* the given data set had been obtained by flipping coin (2, 3) rather than by flipping coin (1, 3). Before continuing, let us note that flipping (2, 3) is indeed more informative than flipping (1, 3):

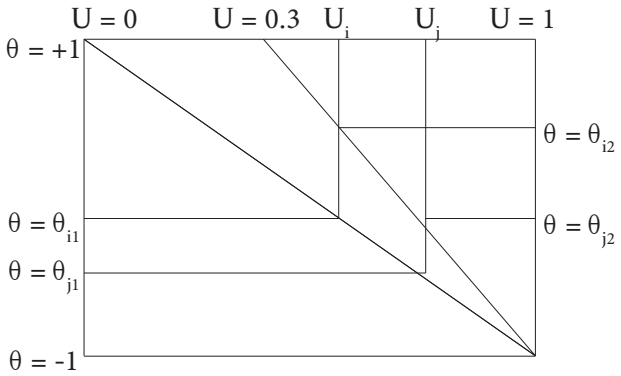
> For X = 3, if it had been obtained by flipping coin (2, 3), the maximum likelihood estimate would have been $\hat{\theta} = +1$, the situations of whose mental correlate would, from the second conditional distribution given at (9.4.1), have been traced at
>
> $$\left\langle \frac{1-\theta}{2}\,,\ \frac{1+\theta}{2}\,,\ \bullet \right\rangle \quad \text{for } -1 \le \theta \le +1. \tag{9.4.8}$$

Figure 9.4.1 overleaf shows that for any two hypothesised values, the trace at (9.4.8) is indeed more separating than the trace at (9.4.5). So one would indeed prefer investigation via flipping (2, 3) rather than flipping (1, 3). In order to then defend the co-ordinates at (9.4.8) in the case of the bent-coin experiment, we would have to reason

> that X = 3 was obtained by flipping coin (1, 3) must be ignored,
> *so that* X = 3 can be viewed unconditionally as in Table 9.4.1,
> *so that* we can condition on $Y_1 = 1$,
> *so that* we can then reason as if X = 3 were obtained by flipping coin (2,3).
> And *that* is circular reasoning.

> Such reasoning is bad enough to be utterly unacceptable. If it were to be argued that the reasoning does not apply in the case of performable subsidiaries such as those in the bent-coin setup, but only to non-performable subsidiaries such as those in the

biased tetrahedral-die setup, the reasoning becomes even worse, since it is then confounded with the arcane notion of a principle of reasoning that does not apply to an experiment that can be performed physically, but that applies only to metaphysical 'experiments' that cannot be physically (actually) performed. (9.4.9)



Trace 1 (lower graph) $(U, \varepsilon, V) = \left[\frac{1 - \theta}{2}, \frac{1 + \theta}{2}, \emptyset\right]$

Trace 2 (upper graph) $(U, \varepsilon, V) = \left[\frac{2 - \theta}{3}, \frac{1 + \theta}{3}, \emptyset\right]$

Given attainable co-ordinates U_i and U_j ($U_i \neq U_j$) the lower graph depicts the more separating tests, as $\theta_{i1} - \theta_{j1} < \theta_{i2} - \theta_{j2}$.

Figure 9.4.1: Two alternative suites of co-ordination tests

Although we have presented the foregoing in terms of a random sample of size $n = 1$, the reasoning repeats for a random sample of any size, because a random sample of any size n comprises n *independent* replicates of size $n = 1$. So the proposal can be applied to the separate replicates, and the results can then be combined.

9.5 Yet another proposal that fails

Barnard and Sprott (1970) propose that, at least in certain cases, any non-uniqueness of ancillaries can be resolved by appeal to invariance principles. They motivate this using Basu's main example, which is reproduced below. We first consider how Basu would have us view the example, then we consider how Barnard and Sprott would have us view the example, and finally we consider how we ought to view the example.

Example 9.5.1

Table 9.5.1 gives the original version of Basu's main example.

Table 9.5.1: A class of models for the outcome of one roll of a biased cubic die

X (value on top)	X = 1	X = 2	X = 3	X = 4	X = 5	X = 6
Probability	$\frac{1-\theta}{12}$	$\frac{2-\theta}{12}$	$\frac{3-\theta}{12}$	$\frac{1+\theta}{12}$	$\frac{2+\theta}{12}$	$\frac{3+\theta}{12}$
Ancillary Y_1	0	1	2	0	1	2
Ancillary Y_2	0	1	2	0	2	1
Ancillary Y_3	0	1	2	1	0	2
Ancillary Y_4	0	1	2	1	2	0
Ancillary Y_5	0	1	2	2	0	1
Ancillary Y_6	0	1	2	2	1	0
Max. like. est. $\hat{\theta}$	-1	-1	-1	+1	+1	+1
$-1 \le \theta \le +1$						

As before, we consider how the model might represent one or the other of two different but substantively meaningful experiments. A first possibility, as indicated in the heading of Table 9.5.1, is one roll of a biased cubic die, where we must ensure that the substance of the action is clearly understood. So we assume that the die was made by covering the six sides of an unbiased cubic die with appropriately weighted squares, and without any distortion of the cubic shape. We also suppose that the resulting biased die is rolled in the usual way by placing the die in a cup, shaking the cup, and tossing the die from the cup. In that case conditioning on any one of the six ancillary statistics, labelled Y_1, Y_2, Y_3, ..., Y_6, in Table 9.5.1, corresponds to a non-performable experiment, because one cannot roll the die such that any of the ancillaries will take on a particular value. A second possibility involves three bent coins with the following outcomes, when flipped:

Coin 1: outcome -1 or +1 with probability $\frac{1-\theta}{2}$ and $\frac{1+\theta}{2}$, respectively.

Coin 2: outcome -1 or +1 with probability $\frac{2-\theta}{5}$ and $\frac{3+\theta}{5}$, respectively.

Coin 3: outcome -1 or +1 with probability $\frac{3-\theta}{5}$ and $\frac{2+\theta}{5}$, respectively.

Note that any outcome is the maximum likelihood estimate of θ. Let the investigator roll an ordinary six-sided die twice in succession. Let the random variables, Z_1 and Z_2, respectively, represent successive outcomes, where $(Z_1, Z_2) = (z_1, z_2)$ represents a particular outcome (z_1, z_2 = 1, 2, 3, ..., 6):

If $z_1 = z_2$ let the investigator flip Coin 1.
If $z_1 < z_2$ let the investigator flip Coin 2.
If $z_1 > z_2$ let the investigator flip Coin 3.

Let X = 1, 2, 3, ..., 6, respectively, label the following, where $(C_j, \theta)\hat{}$ denotes that Coin j is flipped with the indicated outcome:

$$(C_1, -1), (C_2, -1), (C_3, -1), (C_1, +1), (C_3, +1), (C_2, +1). \qquad (9.5.1)$$

Note that at (9.5.1) the order of C_2 and C_3 in the 2nd and 3rd terms are reversed in the 5th and 6th terms. X is a minimal sufficient statistic for θ, and at (9.5.1) we have a one to one transform of X. Hence at (9.5.1) we have alternatively labelled the minimal sufficient statistic for θ, where those alternative labels explicate the following:

$$\Pr(X = 1) \text{ equals } \frac{6}{36} \times \frac{1-\theta}{2} = \frac{1-\theta}{12}, \text{ and } \Pr(X = 4) \text{ equals } \frac{6}{36} \times \frac{1+\theta}{2} = \frac{1+\theta}{12}.$$

$$\Pr(X = 2) \text{ equals } \frac{15}{36} \times \frac{2-\theta}{5} = \frac{2-\theta}{12}, \text{ and } \Pr(X = 4) \text{ equals } \frac{15}{36} \times \frac{2+\theta}{5} = \frac{2+\theta}{12}.$$

$$\Pr(X = 3) \text{ equals } \frac{15}{36} \times \frac{3-\theta}{5} = \frac{3-\theta}{12}, \text{ and } \Pr(X = 4) \text{ equals } \frac{5}{36} \times \frac{3+\theta}{5} = \frac{3+\theta}{12}.$$

This exemplifies a fact worth recalling at this point. The range of any statistic is a set of labels for a mutually exclusive and exhaustive partitioning of a sample space. Any arbitrary one to one re-labelling of the range produces an equivalent statistic. It is often the case that the labels are chosen in some informative way, as done at (9.5.1). However, all the statistical information always derives from the probabilities of whatever labels we use and not from the labels themselves. For instance, in each of Tables 9.2.1, 9.4.1, and 9.5.1 the values of the maximum likelihood estimator derive from the probabilities, not from the values of X.

Basu's view:

As a first possibility, suppose that the given data are X = 5 obtained from one roll of the biased cubic die. Then a model conditional on any of the ancillaries represents a metaphysical experiment that cannot be performed physically. The only performable experiment, which was then in fact performed, is given by the unconditional model. So the maximum likelihood estimator's distribution in this case is

$$\Pr(\theta\hat{} = -1) = \frac{1-\theta}{12} + \frac{2-\theta}{12} + \frac{3-\theta}{12}, \text{ i.e., } \frac{2-\theta}{4}, \text{ and}$$

$$\Pr(\theta\hat{} = +1) = \frac{1+\theta}{12} + \frac{2+\theta}{12} + \frac{3+\theta}{12}, \text{ i.e., } \frac{2+\theta}{4}.$$

Therefore we find:

> For the given data, X = 5 obtained by 'rolling' the die, the maximum likelihood estimate is $\theta\hat{} = +1$, whose mental correlate is traced within the distribution of the estimator to be at

$$\left\langle \frac{2-\theta}{4}, \frac{2+\theta}{4}, \bullet \right\rangle \text{ for } 0 \le \theta \le 1. \tag{9.5.2}$$

As a second possibility, suppose that the given data are X = 5 obtained from the bent coin experiment. Then the model conditional on the outcome of rolling the unbiased die represents the performable subsidiary experiment 'flip Coin 3', which was then in fact performed. So the maximum likelihood estimator's distribution in this case is

$$\Pr(\theta = -1) = \frac{3-\theta}{5} \text{ and } \Pr(\theta = +1) = \frac{2+\theta}{5}.$$

Therefore we find that:

> For the given data, X = 5 obtained by flipping Coin 3, the maximum likelihood estimate is $\theta = +1$, whose mental correlate is traced within the distribution of the estimator to be at
> $$\left\langle \frac{3-\theta}{5}, \frac{2+\theta}{5}, \bullet \right\rangle \text{for } 0 \le \theta \le 1. \tag{9.5.3}$$

The traces at (9.5.2) and (9.5.3) differ; consider for instance testing $\theta = 0$. They differ because they arise from different data, as a statistical datum is not just a number; it is a number arising from a specific sampling scheme.

Barnard's and Sprott's view:

The first two rows of Table 9.5.1 show that the unconditional model is invariant under the transformation X ↠ X+3 (modulo 6) and the induced transformation $\theta \twoheadrightarrow -\theta$. In fact, the first three rows show as follows that [X, Pr(X = x), Y_1] is invariant under the transformation, in that the transformation replaces the given information:

$$[1, \frac{1-\theta}{12}, 0], [2, \frac{2-\theta}{12}, 1], [3, \frac{3-\theta}{12}, 2], [4, \frac{1+\theta}{12}, 0], [5, \frac{2+\theta}{12}, 1], [6, \frac{3+\theta}{12}, 2],$$

with

$$[4, \frac{1+\theta}{12}, 0], [5, \frac{2+\theta}{12}, 1], [6, \frac{3+\theta}{12}, 2], [1, \frac{1-\theta}{12}, 0], [2, \frac{2-\theta}{12}, 1], [3, \frac{3-\theta}{12}, 2],$$

which is merely a rearrangement of the given information. We note in passing that the value X = 5 belongs to the conditional distribution of X given $Y_1 = 1$, as follows:

$$\Pr(X = 2 \text{ given } Y_1 = 1) = \frac{2-\theta}{4} \text{ and } \Pr(X = 5 \text{ given } Y_1 = 1) = \frac{2+\theta}{4}. \tag{9.5.4}$$

With regard to the other ancillary statistics, we find that none of [X, Pr(X = x), Y_j] for j = 2, 3, 4, 5, 6 are invariant under the transformation in question. For instance, in the case of [X, Pr(X = x), Y_2] the transformation replaces the given information, i.e.

$$[1, \frac{1-\theta}{12}, 0], [2, \frac{2-\theta}{12}, 1], [3, \frac{3-\theta}{12}, 2], [4, \frac{1+\theta}{12}, 0], [5, \frac{2+\theta}{12}, 2], [6, \frac{3+\theta}{12}, 1],$$

with

$$[4, \frac{1+\theta}{12}, 0], [5, \frac{2+\theta}{12}, 1], [6, \frac{3+\theta}{12}, 2], [1, \frac{1-\theta}{12}, 0], [2, \frac{2-\theta}{12}, 2], [3, \frac{3-\theta}{12}, 1],$$

where X = 5 then turns out to belong to the conditional distribution of X given $Y_2 = 1$, as follows:

$$\Pr(X = 3 \text{ given } Y_2 = 1) = \frac{3-\theta}{5} \text{ and } \Pr(X = 5 \text{ given } Y_2 = 1) = \frac{2+\theta}{5}. \quad (9.5.5)$$

Barnard and Sprott proceed from the premise that one must condition on an ancillary statistic, and in the event of non-uniqueness, the ambiguity must be removed by way of some or other criterion that picks out 'the correct' ancillary for conditioning. If a model with several ancillary statistics is invariant under any particular transformation, they argue that 'the correct' ancillary for the conditioning will also be invariant under that transformation. So, inasmuch as Y_1 is the only one of the six ancillaries that, in the present example, satisfies this requirement, it must be the correct one. For the given datum, X = 5, the 'correct conditional distribution' is then picked out by conditioning on $Y_1 = 1$, and thus turns out to be the one given at (9.5.4). Hence Barnard and Sprott would have us find:

> For the given data, X = 5, the maximum likelihood estimate is found to be $\hat{\theta} = +1$, whose mental correlate is then traced within the distribution of the estimator given $Y_1 = 1$, to be situated at
> $$\left\langle \frac{2-\theta}{4}, \frac{2+\theta}{4}, \bullet \right\rangle \text{for } -1 \le \theta \le 1. \quad (9.5.6)$$

Reasoning similar to that which leads to Figure 9.4.1, shows that this trace is the more informative of the two traces at (9.5.2) and (9.5.3) respectively.

But look at what this leads to! Suppose the given data, X = 5, was from the bent-coin experiment. Then the given data would have to have been 'X = 5 obtained by flipping Coin 3', whereas the statistical co-ordinates given at (9.5.6) are for 'X = 5 obtained from flipping Coin 1'. So, in order to defend the co-ordinates at (9.5.6) for the bent-coin experiment, we would have to reason that:

> X = 5 was obtained by flipping Coin 3 must be ignored,
> *so that* X = 5 can be viewed unconditionally as in Table 9.5.1,
> *so that* we can condition on $Y_1 = 1$,
> *so that* we can reason as if X = 5 were obtained by flipping a more informative coin.
> And *that* is circular reasoning.

For the same reasons given at (9.4.9), it cannot be argued that the reasoning does not apply in the case of performable subsidiaries, such as those in the bent-coin setup, but only to non-performable subsidiaries, such as those in the biased tetrahedral-die setup.

9.6 Conditioning rules that are bound to fail

The failed proposals in Sections 9.2, 9.4 and 9.5 founder on a common source. Each of them has failed to understand that answers to the investigative question, 'How might these particular data have come about?' can be ruled out by *scientific evidence* only. Each proposal assumes that in case of ancillaries, we must, *as a mathematical principle* for statistical inference, condition on an ancillary, thus forgetting that conditioning rules out certain possibilities, when *as a scientific principle*, such possibilities can be ruled out by virtue of *scientific facts* only. Cox's development shows us how to plan a more informative rather than a less informative experiment, but it cannot retrospectively decide upon the best choice 'because that *would have been* the best choice', without as much as a shred of evidence to show that such a choice had in fact been exercised, the bent-coin examples being cited to show how there *could have been* such evidence. Similar objections apply to the development of Barnard and Sprott. Like Cox they too put the cart before the horse. This must be firmly grasped, as any attempt at explaining 'how these given data might have come about' is bound to be circular if it tries to restrict its answers to 'how one would have *liked* the data to have come about'. Basu's bent-coin examples show us how it *might in fact not be* 'as we *would have liked* it to be'. In other words: Cox (1971) reasons in effect: 'Let us assume without any evidence that the data came about in a subsidiary way, so as then to be able to reason that the data came about in this subsidiary way, rather than in that subsidiary way, as this rather than that would have been the more informative way for it to have come about.' Similarly, Barnard and Sprott (1970) reason in effect: 'Let us assume without any evidence, that the data came about in a subsidiary way, so as then to be able to reason that the data came about in this subsidiary way, rather that that subsidiary way, as this rather than that is mathematically more like the non-subsidiary way in which it came about.' The point here is this: the 'best way' for data to have come about, cannot determine how the data in fact *did* come about, regardless of whether 'best' then refers, as Cox would have it, to an operating characteristic, or refers, as Barnard and Sprott would have it, to an invariance property, or refers to ... . The issue is this: for analysis of given data the statistical analyst must draw the frame of reference as tightly as possible by ruling out whatever can be ruled out *on substantive grounds, and on substantive grounds only.* So, if a frame of reference would then be found to formally involve ancillary statistics that do not convey any substantively meaningful subsidiary models, they must be ignored if the analyst is not to risk falling victim to statistical fictions that arise from circular reasoning.

9.7 Subsidiary models that are not performable yet are not substantively vacuous

Basu (1964, p .13) adjoins a footnote to the example here given in Table 9.5.1, saying:

> 'Of course the experiment of rolling the die repeatedly until, say, either 2 or 5 appears (and then observing only the final score) is essentially equivalent to once tossing a biased coin with probabilities $(2-\theta)/4$ and $(2+\theta)/4$. But who is interested in such a wasteful experiment? The author would classify such experiments under the conceptual (non-performable) category.' (9.7.1)

This seems to be overly restrictive. What if some simpleton has actually carried out such an experiment? For instance, this writer once advised that, in order to compare three rations for dairy cows, cows coming into milk should successively be grouped in blocks of three cows each, and the rations randomly assigned to cows in each block separately; a field plan did not seem to be required. In the event, the outcome was a completely randomised design, with each experimental unit comprising three cows. So if, after all, given data can be fruitfully analysed, the statistician must do so. A more serious interpretation of the statement at (9.7.1) is that together with the ordinary-language interpretation of the term 'performable' it implies that we may condition on an ancillary only if we can *in one fell swoop* fix its value at the value observed, and did so. Consider, however, the analysis of Example 1.31.4, where we employed

> a subsidiary model of a contrast Y_1, obtained by conditioning on the total of three of the 2×2 counts in the body of the contingency table, the first of the two counts on the main diagonal being omitted,
>
> a subsidiary model of a contrast Y_2, obtained by conditioning on the total of two of the 2×2 counts in the body of the contingency table, both of the counts on the main diagonal being omitted, and
>
> a subsidiary model of a contrast Y_3, obtained by conditioning on the total of three of the 2×2 counts in the body of the contingency table, the second of the two counts on the main diagonal being omitted.

None of these three models corresponds to an experiment that is performable in one fell swoop because, given the manner in which data comprising such a contingency table are obtained, one could not, in respect of any one of the three models, have fixed the value of the ancillary involved at the value observed. In fact, the only count that could have been fixed at its observed value is the grand total count of n = 1 301 as in Table 1.31.1. So we either have to reject the analyses advocated in Example 1.31.4 as invalid, or we have to reject the notion that 'performable in one fell swoop' is a necessary requirement for conditioning on an ancillary, where we are then forced by the present example to rule out the one-fell-swoop notion, because the three analyses advocated in Example 1.31.4 are utterly unexceptionable, as follows:

> The conditioning for Y_1 picks out the non-deficient and singly deficient plants as the relevant data for testing the (non-deficient):(singly-deficient) ratio.
>
> The conditioning for Y_2 picks out one and the other of the two kinds of singly deficient plants as the relevant data for testing the (striped):(gold) ratio.
>
> The conditioning for Y_3 picks out the singly deficient and doubly deficient plants as the relevant data for testing the (singly-deficient):(doubly-deficient) ratio.

In each case we are distinguishing the data that are relevant for testing a substantively meaningful model, from data that are irrelevant for that test. So, though the distinction Basu draws between performable and non-performable experiments distinguishes between *substantively meaningful* and *substantively meaningless* models *in the case of his particular examples*, it is the latter distinction that resolves the problem of whether or not

to condition on an ancillary statistic. Note that this also shows that the notion of the conditional model having to correspond to an experiment now also falls away, as a substantively meaningful model need not be that of an experiment. (See Example 9.7.1 below.)

*Exam*ple *9.7.1*

A Mendelian polymorphism in humans involves the distinguishable blood types MM, MN and NN. Genetic polymorphisms are usually caused by heterozygous advantage, which might be of many different kinds. An investigator might for instance wish to test whether the rate of prenatal survival is higher for MN individuals than for NN individuals, and so might determine the blood types of n = 317 new-born infants from MN×NN parents, modelling the data as in Table 9.7.1.

Table 9.7.1: A class of models for the blood type of offspring of MN×NN parents, interest being in whether the value of θ_1 is non-zero; θ_2 is a nuisance parameter

X (blood type)	X = MM	X = MN	X = NN
Number of infants	6	169	142
Probability	θ_2	$\frac{1-\theta_1}{1}(1-\theta_2)$	$\frac{1+\theta_1}{1}(1-\theta_2)$
Ancillary Y	0	1	1
$-1 < \theta_1 < +1$		$0 < \theta_2 < 1$	

We note in passing that the values $\theta_1 = -1, +1$ and $\theta_2 = 0, 1$ are ruled out by the given data, which will serve to remind us that in data analysis, as opposed to forecasting, the model is data dependent. Now in order to test

$$M_0: \theta_1 = 0 \text{ versus } M_1: \theta_1 \neq 0,$$

the investigator will condition on the ancillary, so as to eliminate θ_2, rather than give credence to the idea that it might have been Mendel's principles that had gone astray. The resulting analysis would then be based on the subsidiary model

$$\Pr(X = MN \mid Y = 1) = \frac{1-\theta_1}{2} \text{ and } \Pr(X = NN \mid Y = 1) = \frac{1+\theta_1}{2} \quad (-1 < \theta_1 < +1).$$

The estimated value of θ_1, 1-2[169/(169 +142)], is negative, pointing at heterozygous advantage, and a co-ordination test would measure the strength of that evidence. Such an analysis would be unexceptionable, yet not based on a model that can be described as that of a performable experiment, the numerical data set having been obtained by 'survey' rather than 'experiment'. However, the crux of the matter is that one cannot in one fell swoop ensure that the subsidiary data is fixed at 169+142 relevant observations. But there can be no question that the proposed co-ordination test relies on a substantively meaningful subsidiary model.

9.8 Every-day examples of conditioning on ancillary statistics

In Section 9.3 we saw that owing to the use of randomised experimental design, every-day statistical practice often involves ancillary statistics, indicating that such practice has developed an understanding of the general principles for validly dealing with such statistics. So before trying to formulate such principles, let us consider and discuss the following two paradigmatic examples of that understanding.

Example 9.8.1

Consider a horticulture trial in which a treatment T, aimed at improvement of yield in citrus, is to be compared to a control, C. If each treatment is replicated three times in a completely randomised design, one of 20 equally frequent field plans will be obtained (Table 9.8.1).

Table 9.8.1: The 20 possible field plans for three replications of two treatments in a completely randomised design

Plan 1: T T T C C C	Plan 8: T C C T T C	Plan 14: C T C T T C
Plan 2: T T C T C C	Plan 9: T C C T C T	Plan 15: C T C T C T
Plan 3: T T C C T C	Plan 10: T C C C T T	Plan 16: C T C C T T
Plan 4: T T C C C T	Plan 11: C T T T C C	Plan 17: C C T T T C
Plan 5: T C T T C C	Plan 12: C T T C T C	Plan 18: C C T T C T
Plan 6: T C T C T C	Plan 13: C T T C C T	Plan 19: C C T C T T
Plan 7: T C T C C T		Plan 20: C C C T T T

Let the data be 2, 1, 7, 5, 8, 9, arising from Plan 10. Consider the model

M_0: Treatment, as compared to control, does not improve yield.

In order to test the tenability of this hypothesised model, consider a partial ordering of all the possible samples under this model, according to the value of the test statistic

X = the total yield from the treated plots – the total yield from the control plots.

Under M_0 the value of X equals

-12 with Plan 1, -16 with Plan 2, -10 with Plan 3, ..., -3 with Plan 20.

The test datum arising from Plan 10 is X = +6, whose mental correlate co-ordinates

at (13/20, 1/20, 6/20) in the test distribution. (9.8.1)

Suppose, however, that the investigator has subsequently discovered that, in the case of the first two plots, the soil cover is very shallow on underlying bedrock. The given yields for Plan 10 might be adjusted for this as follows, for an appropriate value of θ:

$2+\theta$, $1+\theta$, 7, 5, 8, 9.

Considered as a source of extraneous variation, the bedrock effect could be eliminated by conditioning on an ancillary statistic, as follows: let Y = 0 for Plans 1, 2, 3, 4, and Plans 17, 18, 19, 20. Let Y = 1 otherwise. Let the treatment effect be modelled as an additive effect τ, so that the expected yields are given by:

$\mu+\theta+\tau$, $\mu+\theta+\tau$, $\mu+\tau$, μ, μ, μ, in the case of Plan 1,
$\mu+\theta+\tau$, $\mu+\theta+\tau$, μ, $\mu+\tau$, μ, μ, in the case of Plan 2,
$\mu+\theta+\tau$, $\mu+\theta+\tau$, μ, μ, $\mu+\tau$, μ, in the case of Plan 3, ...

Y is ancillary because, for any values of θ and τ, the distribution of Y is given by

P(Y = 0) = 8/20 and P(Y = 1) = 12/20.

Y = 1 for the given data set, and conditionally on Y = 1 there are 12 equally frequent field plans, numbered 5, 6, 7, ..., 14, 15, 16. The test datum is X = +6 as before, but its mental correlate now co-ordinates

at (11/12, 1/12, Ø*) in the conditional test distribution. (9.8.2)

Which of the two tests (at (9.8.1) and (9.8.2)) is the correct one? The problem is one of investigative statistics, not one of decision-making under risk and so the test at (9.8.2) is the correct one because the investigative question (as we saw in Definition 1.2.1) is:

How might these *particular* data have come about?, (9.8.3)

where the discovery that the first two plots were on shallow ground, is an informative constituent of those *particular* data. Stated otherwise, the investigator simply cannot in any appropriate manner at all be answering the question at (9.8.3) by pretending not to know about those two plots.

We note in passing that the present paradigm concerns problems that are usually dealt with by analysis of covariance. However, the principles of reasoning that here concern us are more clearly exposed in terms of randomisation theory.

Example 9.8.2

Consider an enlarged version of the previous example by supposing that the original layout called for five replications of each treatment in a completely randomised design, thus generating 252 field plans. As before we suppose that two treated plots, T and C, respectively, are subsequently found to involve very shallow soil owing to underlying bedrock. Additionally, let us suppose that there was some concern that the treatment might discourage visitation by bees, which would have an adverse effect on apiary as a secondary source of income for the citrus farmers. So, one each of ten available bee traps were allocated to the ten experimental plots, respectively. The ten traps comprised five of each of two kinds, A and B. We suppose that it was known from the outset that A traps are more effective than B traps. So the placing of the traps was not randomised. Consider the following field plan, in which the underlining indicates the shallow soil:

<u>(T, A)</u>	<u>(C, A)</u>	(T, B)	(C, B)	(C, B)
(C, A)	(T, A)	(T, A)	(C, B)	(T, B)

For analysis of the fruit yields, an investigator should recognise that fruit yield might be affected by soil differences, but not by the different kinds of traps. So the fruit yield data should be conceived of as originating from two blocks: one of size two units on shallow soil, and the other of size eight units on deep soil. This picks out 2×70 = 140 of the original 252 equally frequent field plans as sampling frame of reference. The analysis would be as in the previous example. For the analysis of the bee counts, the investigator should recognise that bee counts might be affected by yield, and so indirectly also by soil differences and the different kinds of traps. The bee counts should therefore be conceived of as originating from three blocks; one block of size two units on shallow soil with A-like traps, a second block of size three units on deep soil with A-like traps, and a third block on deep soil with B-like traps. This conception picks out 2×3×10 = 60 of the original 252 equally frequent field plans as appropriate frame of reference. Table 9.8.2 shows how those 60 plans are generated by the Cartesian product of the possible patterns within the three blocks.

Table 9.8.2: Cartesian product for generating 2×3×10 field plans

TC CT	×	TTC TCT CTT	×	TTCCC TCTCC TCCTC TCCCT CTTCC CTCTC CTCCT CCTTC CCTCT CCCTT

Let the realised field plan and bee counts be as follows:

Block 1: (T, 3) (C, 4).
Block 2: (C, 6) (T, 8) (T, 7).
Block 3: (T, 7) (C, 9) (C, 8) (C, 7) (T, 9).

We consider as hypothesised model:

M_0: bee counts are not affected by treatment differences.

The effect, if any, of treatment on bee count is estimated within each block as the mean count per plot for treatment minus the mean count per plot for control. Under M_0, the variances of these estimates are proportional to 2, 2/3 and 5/6 for Blocks 1, 2 and 3, respectively. Using an obvious notation for treatment totals per block, the usual weighted combination of the three estimates turns out to be proportional to

$$[(15T_1+10T_2+18T_3)-(15C_1+20C_2+12C_3)]\div 5.$$

For the given data set we find for instance that, under the hypothesised model, the corresponding test statistic, say Z, takes the value

$$Z = [(15(3)+10(15)+18(16))-(15(4)+20(6)+12(24))]\div 5, \text{ which equals } 3.$$

The distribution of the central test statistic is given in Table 9.8.3, where we find that the mental correlate of the datum value, Z = 3, is situated at (36/60, 9/60, 15/60) in the central test distribution, giving no indication of adverse effects on visitation by bees.

Table 9.8.3: Distribution of a test statistic arising from 60 equally frequent cases

Value	-21	-19	-15	-13	-9	-7	-5	-3	-1	1	3	5	7	9	11	13	15	19	21
No. of cases	1	1	3	3	7	6	0	9	6	0	9	3	0	7	1	0	3	0	1

This example involved conditioning on two different ancillary statistics, Y_1 and Y_2, as follows:

$Y_1 = 0$ for those 2×70 amongst 252 equally frequent plans, that have T and C once each in conjunction with shallow soil; $Y_1 = 1$ otherwise.

$Y_2 = 0$ for those $2\times 3\times 10$ amongst 2×70 equally frequent plans, that have T twice and C thrice in conjunction with B-like traps; $Y_2 = 1$ otherwise.

The analysis of yields is conditional on $Y_1 = 0$. The analysis of counts is conditional on $(Y_1, Y_2) = (0, 0)$.

Again, as in the previous example, we have considered analyses using randomisation theory only, because by doing so the reasoning in respect of the ancillary statistics is brought forward untrammelled by other considerations. Whether or not the analyses considered are best in any other sense need not concern us here.

Discussion of Examples 9.8.1 and 9.8.2

These examples show that the randomisation theory of even so mundane a design as a completely randomised one for say 16 replicates of each of eight different treatments, will involve a mind-boggling variety of ancillary partitions, even if limited to those arising from subsidiary designs recognised by Cochran and Cox (1957) or by Hinkelmann and Kempthorne (2007). Yet numerous statisticians analyse such data, day in and day out, using analysis of variance without being aware of having ignored ancillary partitions involved, or using analysis of co-variance without being aware of having conditioned on certain ancillary partitions involved. How then (and this question is a crucial one) do they manage to achieve correct results without awareness of the ancillary statistics involved? There can be only one answer: they choose their model on the ground that it that makes *substantive* sense, for which no insight into the mathematics of ancillary statistics is needed. Looking back at our examples involving ancillary statistics, it will be found that in each example an unexceptionable analysis could be justified on the grounds that it relied on a model that makes *substantive* sense. In each example it was also possible to choose from different ancillary partitions, but in each example we found that such a choice could not

be justified on purely mathematical grounds. On the contrary, for each of several purely mathematical principles considered, that principle could by the use of counter-examples, be shown to fail. Owing to that, we found that such principles in general either fall into circular reasoning, or must resort to arcane metaphysics.

9.9 Conclusion

Our examples provide justification for three different investigative possibilities, as follows:

> Sometimes substantive science must ignore the possibility of conditioning on an ancillary statistic because that would try to introduce a substantively meaningless subsidiary class of models of how given data came about.
>
> Sometimes substantive science must condition on an ancillary statistic because that would introduce a substantively meaningful subsidiary class of models of how given data came about.
>
> Sometimes substantive science must condition on one ancillary statistic in order to model one particular aspect of a given data set, and condition on another ancillary statistic in order to model another aspect of the same data set.

In all three cases *particular* data are involved because we are pointing, not forecasting, and whether or not we condition, depends on *substantive* science, not on mathematical statistics. All this is straightforward enough to make an outsider wonder what the fuss was all about. The answer is revealing: substantive science develops models of physical experience, where such models often involve subsidiary models. So, a statistical model required by substantive science will often involve subsidiary models, where such a subsidiary model will comprise a subsidiary part of the sample space. However, any statistical model comprises a sample space whose probabilities sum to a constant; so any subsidiary model will comprise a subsidiary sample space whose probabilities sum to a constant; and so the result is discernable to mathematical statistics as a formality called 'an ancillary statistic'. The fuss arises simply because mathematical statistics has failed to notice that the converse does not follow, that is to say, it has failed to notice that the formality might be without substance. In fact, it has failed to notice that a randomised design often involves thousands of such formalities without any substance whatsoever.

Chapter 10

LIKELIHOOD INFERENCE

A seminal source of metaphysical views

10.1 Introduction

Likelihood inference concerns the problem of forming opinions about the tenability of alternative values of the index of some or other class of models thought appropriate for the representation of given data. It shares much by way of fundamental ideas with Bayesian inference. In fact almost all, if not all, of the present chapter also applies to Bayesian inference, the main distinction between the two forms of inference being that likelihood inferences do not partake of a knowing subject. That being the case, it is questionable whether likelihood inference is at all appropriately named, because an inference is surely something that has been inferred by a knowing subject. We will develop reasons for rather referring to likelihood inference as *odds-ratio testing*.

Edwards (1972) and Kalbfleisch (1979) have provided introductory accounts of likelihood inference, and we will draw on those accounts. It will be sufficient for the present purposes to consider examples in which the index is a scalar. Recall also that all the fundamental principles of statistical data analysis are capable of exhaustive development in terms of discrete sample spaces only, as real data cannot be recorded in any other way than on a discrete and essentially finite grid of class marks.

10.2 The basic ideas

Edwards's usage of statistical terms is revealingly idiosyncratic; so much so that it is rewarding to introduce the basic ideas of likelihood inference in his terms. Thus, for instance, if a data set S_x is modelled as a sample S_X from some or other member of a class of models indexed by a scalar θ, where $\theta \in \{\theta_1, \theta_2, \theta_3, ...\}$, his definition of *the likelihood function* (1972, p. 9), when expressed in this notation, is:

> 'The *likelihood* of the hypothesis θ_i given data S_x, and a specific model, is proportional to $\Pr(S_x \mid \theta_i)$, the constant of proportionality being arbitrary.' (10.2.1)

When the constant of proportionality is chosen to be 1 (unity) the likelihood is said to be *in kernel form*. At (10.2.1) Edwards's usage of the word 'hypothesis' and of the phrase 'given ... a specific model' implies that likelihood inference is restricted to elimination testing. In fact this appears in so many words (on pp. 3-4) when he states:

> 'A sufficient framework for the drawing of inductive inferences is provided by the concepts of a *statistical model* and a *statistical hypothesis*. ... By *model* we mean that part of the description which is not at present in question, and may be regarded

> as given, and by *statistical hypothesis* we mean the attribution of particular values to the unknown parameters of the model ... these parameters ... being in question, and the subject of the investigation' (original italics). (10.2.2)

He calls the members of a class of models 'a model'. However, he recognises (correctly) that the class characteristic may be 'a matter for dispute' when he says (wrongly):

> 'There is no absolute distinction between the two parts of the model, for what is on one occasion regarded as given, and hence part of the model, may, on another occasion, be a matter for dispute, and hence part of a hypothesis ...' (10.2.3)

and when he continues (even more wrongly):

> 'Every statistical inference is conditional on some model, and the universality with which it is accepted depends upon the general acceptability of the model.' (10.2.4)

So how would he then test the model 'random numbers' in respect of a given data set? Note that whenever he uses the term 'hypothesis', he refers to the possible values of a parameter, and so he begins to go wrong when, as at (10.2.3), he says: 'and hence part of a hypothesis'. The point here is that he assumes (wrongly) that any statistical test concerns the value of some or other parameter. In other words, he is unaware that a commencement test need not involve a parameter. This failing is not peculiar to him; it is a universal failing of all the received theories of statistical inference, and so causes all of them to fall into circular reasoning or incoherence of the kind we met in Sections 4.8 and 4.20. However, before going into that in the present instance, let us at first follow his explanations, so as to grasp what is meant by 'likelihood inference'. To clarify his next definition (1972, p.10) we insert two parenthetic phrases, as follows:

> 'The *likelihood ratio* of two hypotheses on some data (i.e. on the same data) is the ratio of their likelihoods (in kernel form) on that data.' (10.2.5)

Let 'the likelihood of the hypothesis θ_j' as defined at (10.2.1), be denoted by $L(\theta_j)$. If two different hypotheses are denoted by θ_1 and θ_2, respectively, let their likelihood ratio on the same data, as defined at (10.2.5), be denoted by $L(\theta_1/\theta_2)$, the denominator being the likelihood of θ_2.

Example 10.2.1

Let a data set comprising successes and failures bring into mind a sample comprising the number, n, of independent Bernoulli trials needed to obtain a given number, m, of failures, the probability of success, θ, being constant from trial to trial ($0 < \theta < 1$). The probability of a sample involving just x successes then factors as follows, where the first factor in square brackets shows that the class members are negative binomial distributions and the second factor in square brackets conveys the class characteristic:

$$\left[\binom{n\text{-}1}{n\text{-}x\text{-}1}\ \theta^{x}\,(1\text{-}\theta)^{n-x}\right] \times \left[\binom{n\text{-}1}{n\text{-}x\text{-}1}\right]^{-1}. \qquad (10.2.6)$$

According to the definition at (10.2.1) 'the likelihood of θ', given the data, is then any function of θ that is in constant proportion to this probability, and is thus given by:

$$\theta^x (1-\theta)^{n-x} \times C, \text{ where C is } \textit{any arbitrary} \text{ non-zero constant.} \qquad (10.2.7)$$

By choosing C = 1, the likelihood is expressed *in kernel form*, i.e. in the form

$$L(\theta) = \theta^x (1-\theta)^{n-x}. \qquad (10.2.8)$$

We note (and this is crucial) that the phrase *any arbitrary* at (10.2.7) emphasises that by definition the likelihood does not intentionally convey any information on the class characteristic. So one may as well restrict any mention of a likelihood function to its kernel form, and we will tacitly do so. From the expression at (10.2.8) we find that, if θ_1 and θ_2 denote two different hypotheses on the same data, their likelihood ratio, as defined at (10.2.5), is given by:

$$L(\theta_1/\theta_2) = \theta_1^x (1-\theta_1)^{n-x} \div \theta_2^x (1-\theta_2)^{n-x}.$$

For instance, if $n = 2$, $x = 2$, $\theta_1 = 0.50$ and $\theta_2 = 0.25$, we find that $L(\theta_1/\theta_2)$ is given by:

$$(0.50)^2(1-0.50)^0 \div (0.25)^2(1-0.25)^0 = 4. \qquad (10.2.9)$$

In the language of likelihood inference this says that on the given data, '$\theta = 0.50$ is four times "more likely" than $\theta = 0.25$'. We refer to such a comparison of two different index values as *an odds ratio test*.

On p. 30, Edwards formulates the Law of Likelihood (below referred to as the law), as follows, using his italics:

> 'Within the framework of a statistical model, a particular set of data *supports* one statistical hypothesis better than another if the likelihood of the first hypothesis, on the data, exceeds the likelihood of the second hypothesis.' (10.2.10)

He then immediately adjoins the Likelihood Principle (below referred to as the principle) as follows, using his italics:

> 'Within the framework of a statistical model, *all* the information which the data provide concerning the relative merits of two hypotheses is contained in the likelihood ratio of those hypotheses on the data.' (10.2.11)

In passing we again note failure to recognise any need whatsoever for commencement testing. And we note this yet again when he remarks that:

> according to the principle 'there *cannot* be ... facets of the data which are informative, but which the likelihood does not cover'. (Our italics.) (10.2.12)

Likelihood inference would have us reason that for greater support of θ_1 over θ_2, it is *sufficient,* according to The Law, and *necessary,* according to The Principle, that

$$L(\theta_1/\theta_2) > 1.$$

So Edwards joins the Law and the Principle together into a single statement he calls the Likelihood Axiom, as follows, using his italics:

> 'Within the framework of a statistical model, *all* the information which the data provide concerning the relative merits of two hypotheses is contained in the likelihood ratio of those hypotheses on the data, and the likelihood ratio is to be interpreted as the degree to which the data support the one hypothesis against the other.' (Original italics.) (10.2.13)

Here again, the reader should carefully note that when he says 'a statistical model', he refers to the alternative *members* of a class of models, and not to the class as a whole.

Example *10.2.2*

Let a data set comprising successes and failures bring into mind a sample comprising a fixed number, just n, of independent Bernoulli trials, the probability of success, θ, being constant from trial to trial ($0 < \theta < 1$). The probability of a sample comprising x successes and $n - x$ failures then factors as follows, where the first factor in square brackets shows that the class members are binomial distributions and the second factor in square brackets conveys the class characteristic:

$$\left[\binom{n}{x}\theta^{x}(1-\theta)^{n-x}\right] \times \left[\binom{n}{x}^{-1}\right]. \qquad (10.2.14)$$

In kernel form, the likelihood of θ given the data, is:

$$L(\theta) = \theta^{x}(1-\theta)^{n-x}. \qquad (10.2.15)$$

This likelihood is identical to the one at (10.2.8). Thus, should it just happen that the number of trials, n, and the number of successes, x, for different data sets corresponding to Examples 10.2.1 and 10.2.2, respectively, turn out to be identical, likelihood inference would have us draw numerically identical conclusions – despite the two different data sets having originated from two quite different sampling rules.

> The rule in Example 10.2.1 is 'Sample till just m failures (a specified number of failures) have occurred, then stop'. (10.2.16)

> The rule in Example 10.2.2 is 'Sample just n times (a specified number of times), then stop'. (10.2.17)

All else being equal, ought we then to agree to numerically identical conclusions? This question has divided the statistical profession into two irreconcilably different schools of thought.

10.3 Remarks on terminology

The terms 'class of models' and 'class characteristic' have no equivalent counterparts in Edward's usage, none at all. Although he recognises the concept of a statistic that is sufficient for the index of a class of models, he does not recognise the complementary concept of a statistic that is sufficient for the class characteristic. In fact, on p. 24 he defines the concept 'sufficient statistic' as simply 'any contraction of the data which leaves the likelihood unchanged except for a constant'. This views the concept as one whose 'repeated sampling properties' cannot contribute to what we might learn from any given data set – holding that such properties contribute to 'decision procedures' only (p. 38). Thus likelihood inference cannot account for commencement testing. In a subsequent chapter we find that Bayesian inference has the same shortcoming. Henceforth we will dispense with Edwards's idiosyncratic terminology, reverting to the terminology of other chapters.

10.4 Tests of meaning

'Simulation' in the broad sense of 'showing the human body how', is very much the bedrock of science. Geologists, for instance, if challenged to clarify their explanation (model) of say how Table Mountain came about, appeal to laboratory simulation of the mechanical, thermal, chemical and other subsidiary constituents (sub-models) of that explanation. Note also that simulation, thus employed, serves to clarify scientific meaning. It then appears that simulation can also serve as a test for, and of, scientific meaning. Indeed, as shown in Sections 1.34 and 1.36, such tests are especially forceful in the statistical case, as they can be put into the form of operational instructions that might turn out either to be meaningful, or to be meaningless, depending on whether or not they are capable of being carried out. Such tests are of crucial importance when employed as guarantors of scientific meaning, and in the statistical case simulation is particularly simple –certainly much more so than in for instance geology. We must therefore refuse to entertain any explanatory statistical model of how these given data might have come about if it cannot be simulated.

10.5 'Likelihood': What does it *mean*?

Statistical use of the term 'likelihood' originally referred to a well-defined constituent of the discourse of physical (bodily) experience. However, its meaning is not *directly* experiential, as its meaning cannot be put *directly* to the human body by pointing, and saying: See for your self that likelihood there. So, it does not *as such* have a physical meaning, and so cannot *as such* have evidential meaning. Likelihood inference has therefore been criticised as an attempt, by persuasive psychology of the word 'likely', to invest its peculiar usage with some or other metaphysical meaning. Edwards (1972, p. 33) refutes such criticism, by choosing an interpretation in terms of frequencies, as follows (our italics):

> 'There is ... a perfectly simple "operational interpretation" of a likelihood ratio for two hypotheses on some data. It is, of course, the *ratio of the frequencies* with which, in the long run, the two hypotheses will deliver the observed data.' (10.5.1)

This choice of interpretation has three far-reaching consequences:

> Firstly, it must be granted that it does indeed choose to convey physical meaning, as the human body can, by simulation, be forced to grasp that meaning. And so it is possible for scientific reasoning to deal with the consequences of that choice, as will be done in the rest of this chapter. (10.5.2)
>
> Secondly, by choosing to convey physical meanings, likelihood inference is set apart from Bayesian inference, which chooses to convey metaphysical meanings, where metaphysical meanings are difficult to deal with by scientific reasoning. But that is a matter to be dealt with in a later chapter. (10.5.3)
>
> Thirdly, it provides for the physical meaning of any given likelihood value *relative only* to any other value of the same likelihood. So it does not, and indeed cannot, provide for any likelihood to have an *absolute* bodily meaning. (10.5.4)

The point made at (10.5.4) is crucial, as it defines both the scope and limits of odds-ratio testing. This is nicely exemplified by the following adaptation of an example of Kalbfleisch (1979, p.43).

Example 10.5.1

Consider a population comprising a single replicate of each of k denominations and m replicates (m >1) of a further denomination, θ, which is an unknown one of the k + 1 denominations. Let X = x denote the denomination of a random one of the k + m items comprising the population. We wish to use the datum x to develop an opinion about the identity of θ.

First we employ a co-ordination test:

As m > 1, the hypothesised model M_0: $\theta = \theta_0$ ($\theta_0 = x$) is the most likely model, and a likelihood ratio co-ordination test against a given alternative, M_1: $\theta = \theta_1$ ($\theta_1 \neq x$), then compares the statistical co-ordinates of the situations of the mental correlate of X = x under the two alternatives. Let us choose the direction of ordering as:

$O_1 = \{X = x\}$, $O_2 = \{X \neq x\}$, thus pointing to the left.

Then the required co-ordinates are given by:

[*Ø, m/(k+m), k/(k+m)] for M_0, versus
[*Ø, 1/(k+m), (k+m-1)/(k+m)] for M_1.

These are exemplified by:

[*Ø, 0.90, 0.10] versus [*Ø, 0.10, 0.90] if m = 9 and k = 1,
[*Ø, 0.45, 0.55] versus [*Ø, 0.05, 0.95] if m = 9 and k = 11, and
[*Ø, 0.09, 0.91] versus [*Ø, 0.01, 0.99] if m = 9 and k = 91. (10.5.5)

Here the hypothesised model invariably provides a *relatively* better quality of fit than the alternative. However, the *absolute* quality of fit of the hypothesised model, being

excellent for k = 1, satisfactory for k = 11, and awkward for k = 91,

becomes extremely poor for yet larger values of k. All this makes perfectly good sense, as follows: the *relatively* better quality of fit obtained with the hypothesised denomination is an obvious consequence of hypothesising more than just one single replicate for that denomination, as apposed to hypothesising just one single replicate for the alternative denomination. The decline in the *absolute* quality of fit of any particular one amongst the increasingly many denominations is then also an obvious consequence of the 'swamping effect' of such an increase (of course this holds for any fixed sample size, which is just n = 1 in this example).

Next we employ an odds-ratio test:

The probability of the sample is given by:

$1/(m+k)$ if $x \neq \theta$, and $m/(m+k)$ if $x = \theta$.

So, the likelihood ratio is given by:

$L(\theta \neq x/\theta = x) = 1/m$, whatever the value of k.

Thus, corresponding to the three cases at (10.5.5), odds-ratio testing tells us that:

if k = 1 then $\theta = \theta_0$ ($\theta_0 = x$) is '9 times more likely' than $\theta = \theta_1$ ($\theta_1 \neq x$),
if k = 11 then $\theta = \theta_0$ ($\theta_0 = x$) is '9 times more likely' than $\theta = \theta_1$ ($\theta_1 \neq x$), and
if k = 91 then $\theta = \theta_0$ ($\theta_0 = x$) is '9 times more likely' than $\theta = \theta_1$ ($\theta_1 \neq x$). (10.5.6)

Here again the hypothesised model is invariably shown to provide a *relatively* better quality of fit than the alternative model. But these tests give no indication of how the *absolute* quality of fit of either of the two models deteriorates with increasing k.

10.6 Physical (bodily) norms of discrepancy

We must take the position that strictly speaking the explanation at (10.5.1) does not convey a physical meaning for likelihood, or even convey a physical meaning for likelihood ratio; it conveys the physical meaning of the corresponding *frequency* ratio. In order to grasp this we must note that when we conduct odds-ratio tests we must be able to exemplify meanings such as 'extreme ratio' and 'very extreme ratio'. Thus if (0.95, ε, 0.05*) and (0.99, ε, 0.01*) exemplify extreme and very extreme co-ordinates, then by virtue of *the selfsame* physical norms, 1-in-20 and 1-in-100 exemplify extreme and very extreme ratios, respectively. In the case of statistical co-ordinates, a familiar rough-and-ready version of these norms is:

'a discrepancy of two standard error units on the test statistic scale.' (10.6.1)

Edwards (p.35) tries to justify for odds-ratio testing a similar rough-and-ready version of these norms as being:

'a discrepancy of two units on the ln-likelihood-ratio scale'. (10.6.2)

However, a discrepancy of two units at (10.6.2) is physically equivalent to a discrepancy of only 1.1 units at (10.6.1), which discrepancy in turn is equivalent to (0.86, ε, 0.14*) on the co-ordinate scale. Edwards refers to the norm at (10.6.1) as *conventional*, which is beside the point here. What is at issue here is not the *conventionality* of the norm, but the *physicality* of the norm. It is for instance not possible to persuade the human body that if a stick of length 5 cm is short compared to one of length 95 cm, then a chain of length 14 cm is equally short compared to one of length 86 cm. The point here is that norms of physical magnitude have meanings that stand apart from the physical constituency of any objects of such magnitude. For instance, the smallness of a one-twentieth part of a glass of water, is also the smallness of a one-twentieth part of a loaf of bread, and is also the smallness of a one-twentieth part of a population of samples. This must be firmly grasped because *it enables us to set comparable physical norms for the import of certain quite different statistical measures of tenability*. For instance, a 1-in-20 likelihood ratio measures the *ratio of the frequencies* with which, in the long run, two alternative models for the same data would deliver the observed data pattern, whereas a 1-in-20 significance level measures the *absolute frequency* with which, in the long run, the hypothesised model for those same data would deliver a data pattern as extreme or more so than the observed pattern – and where those two measures are *of precisely the same smallness*, despite being the measures of certain quite different constituent parts of two possibly different pools of conceptual samples. That is the case because the smallness involved is to be understood as being a smallness of bodily perception.

10.7 Statistically vacuous likelihood ratios

All the examples considered so far have involved statistically informative likelihood ratios. However, the next three examples show that a likelihood ratio might well convey no statistical evidence – none at all.

Example 10.7.1

Let a real-world datum $x = 0$ be modelled in the human mind as a realisation of X, a random variable whose distribution is a member of the class of models:

$$\Pr(X = x|\theta) = \frac{1+x\theta}{5},$$

with $-2.5 < x < +2.5$ and rounded to $x \in \{-2, -1, 0, +1, +2\}$,

and $\theta \in \{-0.4, 0, +0.4\}$. (10.7.1)

The likelihood of θ, given the datum, is

$$L(\theta) = \frac{1+(0)\theta}{5}, \text{ which equals } 0.2 \text{ identically in } \theta. \quad (10.7.2)$$

So, in respect of the comparative tenability of the three alternative 'hypotheses', this is an utterly uninformative likelihood. In the human mind, however,

the mental correlate of the given datum is being situated at:
(0.64, 0.20, 0.16), (0.40, 0.20, 0.40), (0.16, 0.20, 0.64)
within the members indexed by θ = -0.4, 0, +0.4, respectively. (10.7.3)

In respect of the comparative tenability of the three alternatively explanatory models, these are clearly not uninformative co-ordinates. Let us consider the precise nature of the physical (bodily) evidence reported at (10.7.3). Figure 10.7.1 represents the three alternative models in full detail, where the shaded area in each case displays the situation

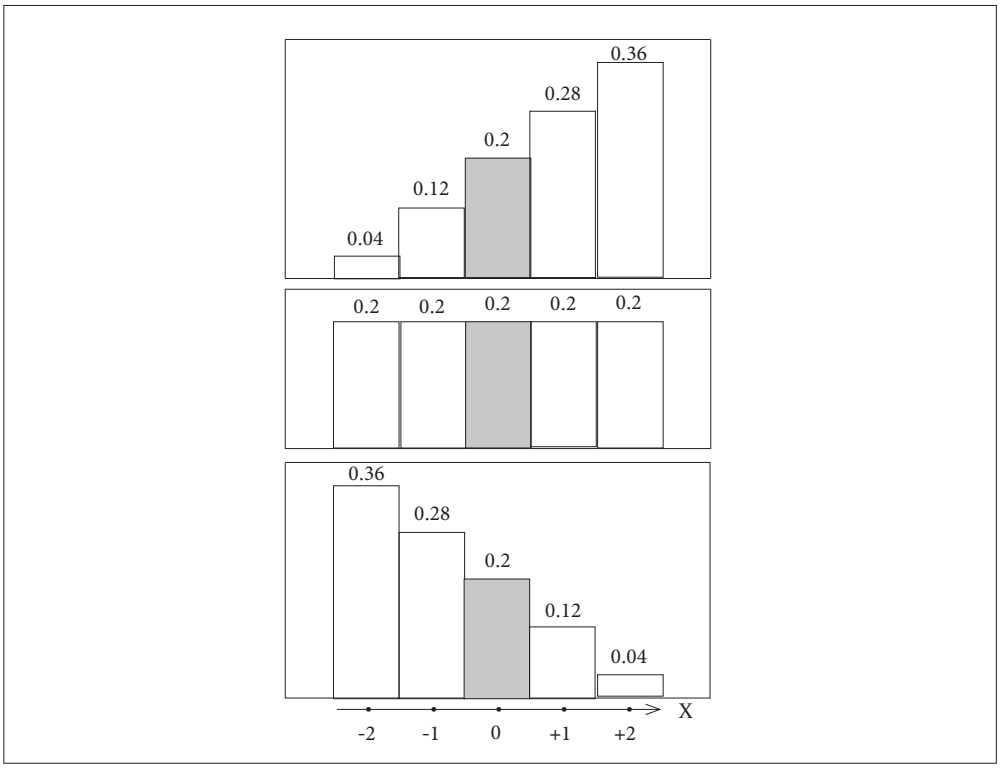


Figure 10.7.1: Testing the quality of fit of three alternative models (depicted with frequency labels) for the source of a given datum X = 0. The models are indexed by θ = -0.4 (top), θ = 0 (middle) and θ = +0.4 (bottom)

of the mental correlate of the given datum. Thus the figure provides all the facts that anyone might require in order to address the question: 'How do these three models compare as alternative explanations of how the given datum might have come about?' We have two competitive answers to this question:

> The answer at (10.7.3) is physically meaningful and clearly appropriate. In effect it points at Figure 10.7.1 for the reader to 'See for yourself that the model indexed by θ = 0 fits the data exceptionally well, and that either one of the other two models also fits the data well, though not quite as well as does the one indexed by θ = 0.'

Figure 10.7.1 represents perceptual facts that could be obtained by simulation. Yet, owing to the likelihood displayed at (10.7.2), likelihood inference would have us defy those facts by insisting instead that the three explanatory models presented in Figure 10.7.1, are equally 'likely'.

Example 10.7.2

An axiom must be *comprehensive*, and so must never be made to rely on a plausibility imparted by the peculiarities of certain examples only. Apart from different notation, the following attempt at such persuasive use of the peculiarity of a particular example is taken *verbatim* from Edwards (1972, p. 35):

> 'A chain has n links, and has been made up by choosing each link at random from a large population of links of which half are made of silver, and half of gold. If it has two silver links, what is n?
>
> $$\Pr(2 \text{ silver} \mid n) = \binom{n}{2}\frac{1}{2^n} = \frac{n(n\text{-}1)}{2^{n+1}}$$
>
> This is therefore the likelihood for n, given two silver links, and at $n = 2, 3, 4, 5, \ldots$ takes the values:
>
> $$\frac{1}{4}, \frac{3}{8}, \frac{3}{8}, \frac{5}{16}, \bullet\bullet\bullet$$
>
> which are in the ratio $4 : 6 : 6 : 5 \ldots$. The solutions $n = 3$ and $n = 4$ are equally well supported; is this reasonable?' (10.7.4)

Edwards is of course confident that we will answer his question in the affirmative, as he evidently reasons that with $X = 2$ silver links from $n = 3$ links all told, the addition of a further link such that we still have just two silver links, determines that the extra link is of gold. So he is pleased to discover that his 'axiom' would have us judge that the models with $n = 3$ and $n = 4$, respectively, fit the given datum $X = 2$ equally well, because $\Pr(X = 2 \mid n)$ equals $6 \div 2^4$ for both $n = 3$ and $n = 4$, as follows:

$$\begin{aligned}\Pr(X = 0, 1, 2, 3, 4, 5, 6, 7, \ldots \mid n = 3) &= (1, 3, 3, 1, 0, 0, 0, 0, \ldots)\div 2^3 \\ &= (2, 6, 6, 2, 0, 0, 0, 0, \ldots)\div 2^4.\end{aligned}$$

$$\Pr(X = 0, 1, 2, 3, 4, 5, 6, 7, \ldots \mid n = 4) = (1, 4, 6, 4, 1, 0, 0, 0, \ldots)\div 2^4.$$

But we must reject such a judgement because the statistical co-ordinates of the mental correlate of the datum, $X = 2$ silver links, are distinctly different for the two models.

> For $n = 3$ the co-ordinates are $(U, \varepsilon, V) = (8, 6, 2)\div 2^4$.
> For $n = 4$ the co-ordinates are $(U, \varepsilon, V) = (5, 6, 5)\div 2^4$. (10.7.5)

Similarly to the result at (10.7.2), the statistical roundings (likelihoods in Edwards's terminology) are equal (both being equal to $6\div 2^4$ in the present case). Similarly also to the result at (10.7.3), however, the statistical co-ordinates at (10.7.5) show that one of the models (here the one with $n = 4$) fits the datum slightly better than does the other. (We

note in passing that, for reasons to be explained in Section 10.9, it will be found that, in the foregoing, the reasoning:

> 'with X = 2 silver links from n = 3 links all told, the addition of a further link such that we still have just two silver links, *determines* that the extra link is one of gold' (10.7.6)

involves a subtle error.)

Example 10.7.3

Let a data set comprising n numbers bring into the human mind a random sample of size n drawn from an infinite population of numbers uniformly distributed from zero to θ ($0 < \theta < \infty$). The largest sample number, $X_{(n)}$, is the minimally sufficient statistic for θ, and so its realisation, $x_{(n)}$, conveys *all the given data* concerning the value of θ. We note that those data comprise two distinctly different parts, which we may call *determinate* and *statistical*, respectively – the determinate part being that no sample value can be larger than θ, and the statistical part being that there then also remains room for statistical variation in the value of $X_{(n)}$, where ignoring that variation would be tantamount to ignoring the *statistical* part of the given data. Now the density function of $X_{(n)}$ is given by:

$$f[x_{(n)}] = \frac{n}{\theta^n}\,[x_{(n)}]^{n-1}, \text{ where } 0 < x_{(n)} \le \theta, \text{ as no sample value can be} > \theta.$$

$$\text{So } \Pr[x_{(n)}-0.5dx_{(n)} < X_{(n)} < x_{(n)}+0.5dx_{(n)}] = \frac{n}{\theta^n}\,[x_{(n)}]^{n-1}dx_{(n)} \text{ for } 0 < x_{(n)} \le \theta.$$

So, the likelihood of θ, for the given datum $x_{(n)}$, is expressed in kernel form as:

$$L(\theta) = \frac{1}{\theta^n}, \text{ where } 0 < x_{(n)} \le \theta < \infty, \text{ as no sample value can be} > \theta. \qquad (10.7.7)$$

Let θ_1 and θ_2 denote two of the different hypotheses that are possible on the same datum, $x_{(n)}$, where $\theta_1 < \theta_2$. Then their likelihood ratio is given by:

$$L(\theta_1/\,\theta_2) = \left\{\frac{\theta_2}{\theta_1}\right\}^n, \text{ where } 0 < x_{(n)} \le \theta_1 < \theta_2 < \infty, \text{ as } x_{(n)} \text{ cannot be} > \theta_1.$$

Thus, for instance, if $\theta_1 = 4$, $\theta_2 = 8$, and $n = 3$, the likelihood ratio *on whatever the value of* $x_{(n)}$ *(apart from not being* > 4*) then happens to be*, is given by:

$$\left\{\frac{8}{4}\right\}^3 = 8, \text{ which, in the language of likelihood inference, would have us infer that:}$$

> $\theta = 4$ is eight times more likely than $\theta = 8$ is.

But, apart from the *determinate* information conveyed by the inequality $x_{(n)} \le \theta$, a co-ordination tester will also consider the *statistical* import of the two alternative models, and thus also consider that the mental correlate of a given $x_{(n)}$ datum is situated within the two alternative test distributions at:

$$(U, \varepsilon, V) = \left[\left(\frac{x_{(n)}}{\theta_J} \right)^n, \varepsilon, 1 - \left(\frac{x_{(n)}}{\theta_J} \right)^n \right], \text{ for } J = 1, 2.$$

So we might have any one of the situations below.

Situation 1**:** The relevant data is $x_{(n)} = 1.7$.

Co-ordination testing knows that $\theta \geq 1.7$, and finds further that the mental correlate of $x_{(n)} = 1.7$ is situated at:

$(0.03, \varepsilon, 0.97)$ in the test distribution for $\theta = 4$, and at

$(0.01, \varepsilon, 0.99)$ in the test distribution for $\theta = 8$.

Co-ordination testing therefore concludes that, as tested, both $\theta = 4$ and $\theta = 8$ fit the data poorly.

Situation 2**:** The relevant data is $x_{(n)} = 2.2$.

Co-ordination testing knows that $\theta \geq 2.2$, and finds further that the mental correlate of $x_{(n)} = 2.2$ is situated at:

$(0.17, \varepsilon, 0.83)$ in the test distribution for $\theta = 4$, and at

$(0.02, \varepsilon, 0.98)$ in the test distribution for $\theta = 8$.

Co-ordination testing therefore concludes that, as tested, $\theta = 4$ fits the data well, and $\theta = 8$ fits the data poorly.

Situation 3**:** The relevant data is $x_{(n)} = 3.8$.

Co-ordination testing knows that $\theta \geq 3.8$, and finds further that the mental correlate of $x_{(n)} = 3.8$ is situated at:

$(0.86, \varepsilon, 0.14)$ in the test distribution for $\theta = 4$, and at

$(0.11, \varepsilon, 0.89)$ in the test distribution for $\theta = 8$.

Co-ordination testing therefore concludes that, as tested, both $\theta = 4$ and $\theta = 8$ fit the data well.

Now consider likelihood inference in the same three situations, as follows, where the information conveyed by the inequality $\theta \geq x_{(n)}$, as such, is *determinate* only.

Situation 1**:** The relevant data is $x_{(n)} = 1.7$.

Likelihood inference notes the *determinate* import of the inequality $\theta \geq 1.7$, but ignores the *statistical* import of the alternative models, and so finds only that:

$L(\theta_1/\theta_2) = 8$.

Likelihood inference therefore concludes only that $\theta = 4$ is eight times more likely than $\theta = 8$.

Situation 2**:** The relevant data is $x_{(n)} = 2.2$.

Likelihood inference notes the *determinate* import of the inequality $\theta \geq 2.2$, but ignores the *statistical* import of the alternative models, and so finds only that:

$L(\theta_1/\theta_2) = 8$.

Likelihood inference therefore concludes only that $\theta = 4$ is eight times more likely than is $\theta = 8$.

Situation 3**:** The relevant data is $x_{(n)} = 3.8$.

Likelihood inference notes the *determinate* import of the inequality $\theta \geq 3.8$, but ignores the *statistical* import of the alternative models, and so finds only that

$L(\theta_1/\theta_2) = 8$.

Likelihood inference therefore concludes only that $\theta = 4$ is eight times more likely than $\theta = 8$.

Discussion of Examples 10.7.1, 10.7.2 and 10.7.3

Any scientific model can be tested empirically, only by comparing predictions derived from it to corresponding empirical data and, in each of our three examples, statistical co-ordinates were used to prove that such tests were both possible and informative. The statistically vacuous nature of the corresponding likelihood ratios involved therefore proves that in each of our examples, those ratios failed to provide information concomitant to the available statistical information. In our Examples 10.7.1 and 10.7.2 that just happens to be the case for particular data. However, in our Example 10.7.3 that would be the case for *any* possible data. The reader will note that for Example 7.4.1 too, that would be the case for *any* possible data.

10.8 A VICIOUS CIRCLE

Axiomatic development is extremely slippery, for two reasons. The first reason is that the human mind tends to cast around for a deduction, whereas the validity of an axiom is not established by deduction; it is established by demonstration because it is a first principle. That involves the second reason: just what constitutes a valid demonstration? The slippery nature of any such development is exemplified by attempts on the part of Birnbaum (1962) and of Basu (1975) to derive the Strong Likelihood Axiom (henceforth referred to as the strong axiom) from more basic principles. This axiom is as follows:

Let datum x_1 be appropriately modelled as having been sampled from a member of a class of models, E_1, and let another datum x_2 be appropriately modelled as having been independently sampled from a member of *another* class of models, E_2, where E_1 and E_2, share a parameter space. Let the likelihoods of x_1 and x_2 have the same kernel form. Then x_1 and x_2 must lead to identical conclusions about the parameter.

A simple example might be that a bent coin is spun just three times and the outcome, x_1, is appropriately modelled as a binomial sample (E_1). The same coin is spun until the first success (as opposed to failure) is obtained, and the outcome, x_2, is appropriately modelled as a geometric sample (E_2). Let Pr(Success) = θ ($0 < \theta < 1$). Then we have:

If x_1 = just 2 failures, then $\Pr(X_1 = x_1) = 3(1-\theta)^2\theta$. The kernel is $L(\theta) = (1-\theta)^2\theta$.
If x_2 = just 2 failures, then $\Pr(X_2 = x_2) = (1-\theta)^2\theta$. The kernel is $L(\theta) = (1-\theta)^2\theta$.

So the strong axiom would have x_1 and x_2 lead to identical conclusions about θ. A co-ordination tester would disagree. For instance, when testing $\theta = 0.25$ as hypothesised model, the situations of the mental correlates of x_1 and x_2 in the test distributions are at:

[••• , $3(0.75)^2(0.25)$, $(0.75)^3$] = (0.16, 0.42, 0.42), and at
[$(0.25)+(0.75)(0.25)$, $(0.75)^2(0.25)$, •••] = (0.44, 0.14, 0.42),

respectively. We must also consider the Weak Likelihood Axiom (henceforth referred to as the weak axiom), as follows:

> Let datum x_1 be appropriately modelled as having been sampled from a member of a class of models E, and let another independent datum x_2 also be sampled from a member of *the same* class of models. Let the likelihoods of x_1 and x_2 have the same kernel form. Then x_1 and x_2 must lead to identical conclusions about the parameter.

A simple example might be that the bent coin is spun just three times and the outcome, x_1, is appropriately modelled as a binomial sample. The same coin is again spun just three times and the outcome, x_2, is appropriately modelled as an independent binomial sample. Then we have:

If x_1 = SFS, then $\Pr(X_1 = \text{SFS}) = \theta\times(1-\theta) \times \theta$. The kernel is $L(\theta) = (1-\theta)\theta^2$.
If x_2 = SSF, then $\Pr(X_1 = \text{SSF}) = \theta\times\theta\times(1-\theta)$. The kernel is $L(\theta) = (1-\theta)\theta^2$.

So the weak axiom would have x_1 and x_2 lead to identical conclusions about θ. When two likelihoods have the same kernel form, Basu refers to them as *equivalent* (his Definition 6). Corresponding to the strong likelihood and weak likelihood principles, Basu formulates two principles called the Invariance Principle and the Weak Invariance Principle, respectively. They differ from the two axioms only inasmuch as they call for *identical* likelihoods instead of *equivalent* likelihoods. Basu also formulates two further principals called the Conditionality Principle and the Weak Conditionality Principle, where only the latter principle need to concern us here. It concerns any class of models E, which is a 'mixture' of two classes of models, E_1 and E_2, which share a parameter space. The mixture is made by using a random device to select one or the other E-like component with known selection probabilities. For instance:

> Roll a true die to choose between:
> E_1 = binomial sampling (just three spins of the bent coin), and
> E_2 = geometric sampling (spinning the bent coin until the first success is obtained),
> with probabilities say 1/6 and 5/6, respectively. (10.8.1)

The Weak Conditionality Principle may be stated as follows:

> If E is a mixture of E_1 and E_2 as described above, and the data
> $[E, (E_i, x_i)]$ for either i = 1 or i = 2, is a realisation of E,
> then the data are equivalent simply to the realisation (E_i, x_i) for the given i.

For instance:

> If at (10.8.1) the die selects geometric sampling, the appropriate model is simply the geometric model (without any reference to the mixture).
>
> Again, in the mixed sampling problem of Section 1.5, the appropriate model according to the Weak Conditionality Principle refers only to the instrument that was actually used.

This brings us to what Basu calls the 'centre-piece' of his essay: a theorem stating that:

> The Weak Invariance Principle and the Weak Conditionality Principle together imply the Strong Likelihood Principle.

Basu's 'proof' of this theorem may be paraphrased as follows:

> Let data (E_1, x_1) and (E_2, x_2) generate equivalent likelihood functions, $L_1(\theta)$ and $L_2(\theta)$, respectively, that is to say, let $L_1(\theta)$ and $L_2(\theta)$ have the same kernel form. Then there exists a known positive constant C such that
>
> $$CL_1(\theta) = L_2(\theta) \text{ for any of the possible values of } \theta. \qquad (10.8.2)$$

Let us then contemplate the mixture E obtained by mixing (E_1, x_1) and (E_2, x_2) in the proportions:

$$\frac{C}{1+C} \text{ and} \frac{1}{1+C} \text{ , respectively.} \qquad (10.8.3)$$

Then $(1, x_1)$ and $(2, x_2)$ are descriptions in the sample space of the mixed class of models, E. In view of the equation at (10.8.2) and of our choice at (10.8.3), the data $[E, (1, x_1)]$ and $[E, (2, x_2)]$ generate *identical* likelihood functions,

$$\frac{C}{1+C} \times L_1(\theta) \text{ and} \frac{1}{1+C} \times L_2(\theta), \text{ respectively.} \qquad (10.8.3)$$

So it follows from the Weak Likelihood Principle that the data $[E, (1, x_1)]$ and the data $[E, (2, x_2)]$ must lead to identical conclusions about the parameter. Now applying the Weak Conditionality Principle to each of these two sets of data, it follows that data $(1, x_1)$ and data $(2, x_2)$ must lead to identical conclusions about the parameter, which is what Basu's theorem says.

Basu notes that Birnbaum advances a slightly different 'proof' in that, instead of appealing to the Weak Likelihood Principle, he appeals to a more general rule implying that principle. In effect this claims to have 'improved' Birnbaum's 'proof' by removing a redundancy. But both of the 'proofs' rely on circular reasoning. The circularity is introduced when we are asked to 'contemplate the mixture E obtained by mixing (E_1, x_1) and (E_2, x_2)'; this mixes *two data sets*, whereas the reasoning must then proceed as if having mixed *two classes of models*. To grasp this very firmly, let us resort to our old standby: simulation. Let E_1 denote the binomial class 'spin the coin just three times' and E_2 denote the geometric class 'spin the coin until the first success'. Let S denote success and F denote failure. Then (E_1, x_1) denotes a data set, possibly as follows:

1st spin: F. 2nd spin: S. 3rd spin: F. Hence (E_1, x_1) denotes (binomial class FSF).

That means that (E_2, x_2) must denote (geometric class FFS). So, let us simulate:

1st spin: F. 2nd spin: S. Ah shucks! We will have to try again.

1st spin: F. 2nd spin: F. 3rd spin: F. Ah shucks! We will have to try again.

1st spin: S. Ah shucks! We will have to try again.

Etc.

This tries to procure a geometric datum whose likelihood has a given kernel form, *so that* its 'likelihood' will have the self-same kernel form as that of the given binomial datum, *so that* an appropriate mixing ration can be selected, *so that* the likelihoods emerging from the mix will be identical, *so that* it can be said that they are as alike as Tweedle Dee and Tweedle Dum, *so that* it can be said that they must then lead to identical conclusions. *That* is circular reasoning.

Concluding remark

The reader will find that odds-ratio tests rely on the Weak Likelihood Axiom. This is in fact implied by the interpretation at (10.5.1).

10.9 Another vicious circle

We are now ready to come to grips with an argument that some statisticians hold to be utterly compelling while others hold it to be utterly silly. The argument in question, which we may refer to as the Stopping-Rule Argument, tries to persuade us that the distinction between the two stop-sampling rules at (10.2.16) and (10.2.17) respectively, is entirely irrelevant for the purposes of data analysis, and thereby tries to motivate the Likelihood Axiom. The argument often abbreviates 'stop-sampling rule' as 'stopping rule' rather than 'sampling rule', which will turn out to be revealing. We now present that argument, then we develop a counter-argument, which we call the Sampling-Rule Argument and which leads to a counter-axiom.

The Stopping-Rule Argument

Let θ denote the probability of outcome S when spinning a bent coin marked S and F on opposite sides ($0 < \theta < 1$). Let an investigator spin the coin repeatedly until that outcome occurs, thus obtaining a data set on the value of θ. Suppose, for the sake of simplicity, that just three repetitions are needed. The following sample space then describes all the outcomes that *could* have occurred, where the underlined description is that of the outcome that *did* occur:

S, FS, <u>FFS</u>, FFFS, FFFFS, ••• (10.9.1)

Now let the investigator consider that if the stopping rule had been 'stop after just three repetitions' instead of 'stop as soon as an S is obtained', then the following sample space would describe all the outcomes that *could* have occurred, where the underlined description would *again* be that of the outcome that *did* occur:

FFF, FFS, FSF, SFF, FSS, SFS, SSF, SSS (10.9.2)

At both (10.9.1) and (10.9.2) the data obtained are described (so it is argued) by FFS, whereas the other descriptions at (10.9.1) and (10.9.2) do not describe any data that were actually obtained. So, it is argued, the investigator should avoid reasoning in terms of outcomes that *could have* occurred, but *did not* occur, and should reason only in terms of the outcome that *did* occur (Birnbaum 1962, p. 271). The persuasive thrust of this argument is encapsulated in the question:

> How can a stopping rule the investigator had *in mind only,* be a constituent of the actual evidence that *as such* can arise *in the real world only*? (10.9.3)

The argument represents the beginnings of a metaphysical view that has, by way of Bayesian inference been able to mount a mathematically sophisticated and hugely influential attack on scientific reasoning – an attack against which our profession has been remarkably inept at defending itself.

The Sampling-Rule Argument

Let us recall how Example 10.7.2 exemplifies the pitfalls of trying to derive a general principle from the peculiarities of a particular example. Recall, especially, the phrase at (10.7.6) showing how the peculiarity of Example 10.7.2 tricks one into *deterministic* reasoning where *statistical* reasoning is actually required. Then note that the reasoning leading up to the question at (10.9.3) relies on a similarly deceptive peculiarity in that it tricks one into thinking in *mechanical* terms – in this particular instance of how, at (10.9.1), the bent coin topples over to produce the first F, the second F, and then, toppling over to the other side, the S. We are thereby tricked into thinking (entirely correctly) that those events could just as well have occurred at (10.9.2). In this our thinking is not wrong; it might in fact be entirely correct in some other context, but it is irrelevant in the present context because it makes us forget that the investigator, when spinning the coin, does not exercise a deterministic control over the outcome. Thus, entirely correctly simulated *mechanics* of how FFS might have come about would nevertheless be irrelevant, because here simulation of a *statistical* event is required. So, consider instead how we might, somewhat naïvely, try to perform the appropriate simulation by stopping and restarting whenever one of the outcomes, S, FS, FFS, and FFF, occurs. Then the likelihood of FFS would be given by:

$(1\text{-}\theta)^2\theta \div [\theta+(1\text{-}\theta)\theta+(1\text{-}\theta)^2\theta+(1\text{-}\theta)^3]$, which is $\neq (1\text{-}\theta)^2\theta$, the required likelihood.

The reader will note that we here arrive at the wrong likelihood because of employing the wrong stop-sampling rule. More importantly, however, the reader should note that the mechanical simulation and the statistical simulation described above *are both* inappropriate because of trying, in *both* attempts, to simulate *a particular sample pattern*, whereas only the simulation of *a particular population of samples* is possible. So it dawns on us that we cannot simulate 'a sample'. And so it dawns on us that we can simulate 'a population' only. And so it dawns on us that we cannot stage the requisite simulation without a sampling rule. Thus Axiom 10.9.1 appears.

Axiom 10.9.1 (The sampling-rule axiom):

> Any purported statistical model that is incapable of simulation is incapable of a scientific defence and, as it is impossible to simulate how a purported statistical sample pattern might have come about without simulating how the entire statistical population to which it belongs comes about, it is impossible to simulate the model without involving its particular sampling rule.

It is in the nature of axiomatic reasoning that an axiom cannot be deduced by logical reasoning; it can be defended – or refuted – by examples only. Therefore, in order to compel the reader to accept Axiom 10.9.1, we must challenge the reader to try, by way of examples, to refute the axiom. It will then be found that one is incapable of devising such examples. This is hardly surprising in view of Section 1.4, where the discourse of statistical science was seen to originate in an *inseparable* pair of ultimate concepts called 'sampling' and 'long-run frequency', respectively, and where we now see that Axiom 10.9.1 is just a reassertion of the *ultimate nature*, and of the *inseparability*, of the two members of that pair. Thus we can see that Axiom 10.9.1 may appropriately be called '*The Fundamental Axiom of Statistical Reasoning*'. It shows why willy-nilly appropriate simulations for the matters in hand must necessarily involve the sampling rules given at (10.2.16) and (10.2.17). It can also now be seen that the term 'sampling rule' is more appropriate than 'stopping rule', as different sampling rules need not be distinguished by different stopping rules. In other words, the two 'stopping rules' at (10.2.16) and (10.2.17) are peculiar to Examples 10.2.1 and 10.2.2, respectively, and we must be exceedingly leery of trying to draw a general principle from anything that is peculiar to particular cases only. Clearly then our point of departure must here be to invoke the sampling rules that bring to the human mind the two different populations whose sample spaces are indicated at (10.9.1) and (10.9.2), respectively, and then to establish how the data are alternatively modelled in the two cases, noting, of course, that in the light of the fundamental axiom*, one or the other pseudo-sampling rule is inexorably part and parcel of those 'data'*. For the present purposes it will be sufficient to summarise each model by giving the statistical co-ordinates of the mental correlate of the given datum (just one success) thus modelled. Those co-ordinates are:

$$[U_1(\theta), \varepsilon_1(\theta), \bullet] = [\theta+(1-\theta)\theta, (1-\theta)^2\theta, \bullet] \qquad (10.9.4)$$

in the negative binomial case, and

$$[U_2(\theta), \varepsilon_2(\theta), \bullet] = [(1-\theta)^3, 3(1-\theta)^2\theta, \bullet] \qquad (10.9.5)$$

in the binomial case. (At (10.9.5) the data patterns FFS, FSF and SFF must be taken to be evidentially equivalent, as the correct likelihood always requires derivation from a statistic that is sufficient for the index.) The likelihood ratio for θ_1 versus θ_2 is then obtained from:

(10.9.4) as $\varepsilon_1(\theta_1) \div \varepsilon_1(\theta_2) = (1-\theta_1)^2\theta_1 \div (1-\theta_2)^2\theta_2$, or from

(10.9.5) as $\varepsilon_2(\theta_1) \div \varepsilon_2(\theta_2) = (1-\theta_1)^2\theta_1 \div (1-\theta_2)^2\theta_2$,

being of course the same in either case, where the Likelihood Axiom would then have us discover a general principle. However, whenever particular examples seem to lead to a general principle, we must carefully test that principle on further examples from the

general domain in which that principle would have to apply. So, apart from the two stopping-rule examples, consider also Example 10.7.1, letting the data for that example be denoted by $\underline{x = 0}$, where underlining again denotes an outcome that *did* occur. Thus the notations $\underline{\text{FFS}}$ and $\underline{x = 0}$ denote constituents of the real world (out there) whereas one reasons (up here) in terms of the constituents of the human mind only. It follows that a 'principle' that tells us to avoid reasoning in terms of any outcomes that *could have* occurred but that *did not* occur, thus to reason in terms only of outcomes that *did* occur, does not make sense. For instance, in order to reason about certain real-world bananas (out there) we can hardly bring the real bananas to mind (up here) by stuffing them up our ears; instead, it is their mental correlates that must then be brought to mind. Similarly, $\underline{\text{FFS}}$ and $\underline{x = 0}$ denote real-world objects whose mental correlates must be brought to mind. Axiom 10.9.1 shows how, and how *only*, that can be accomplished by way of a model that is capable of simulation. Thus the best we can possibly do to accommodate The Stopping-Rule Argument, is to bring to mind, not $\underline{\text{FFS}}$, but its mental correlate, and necessarily in some way indicated by Axiom 10.9.1. We then find that the mental correlate is not a specifiable one of the infinitely many samples that comprise the statistical rounding, as we cannot for instance specify it as sample number 1 733 from the left. We can but describe the rounding it belongs to. And so, willy-nilly the outcome that *did* occur is replaced by an infinite pool of samples called the statistical rounding. Moreover, for any present purposes the only useful description of that rounding is its measure, and that measures the rounding as a proportion of the *entire* population of samples. So, willy-nilly we are then reasoning in terms of *all* the samples comprising the population, both those comprising the rounding, and those comprising the two co-ordinates. Now, taking stock of our situation, for each of the two data sets denoted by $\underline{\text{FFS}}$ and $\underline{x = 0}$, respectively, we find that *in both cases*:

> We have a given datum in the data record, for which we have brought a variety of explanatory models into the human mind.
>
> For each model, we have to judge quality of fit of that model as a possible explanation of how the given datum might have come about.
>
> For each judgment, the present purposes will find that sufficient information on which to base that judgement is conveyed by a set of statistical co-ordinates.
>
> Each of those sets of statistical co-ordinates comprises just three numbers, but involving at most only two independent variables (because $U+\varepsilon+V = 1$).
>
> Likelihood inference would have us replace those two variables with just a single variable called the 'likelihood ratio'.
>
> We must then ask: 'Why only one variable?'

It then appears that The Stopping-Rule Argument is circular, because it would have us

> draw on example $\underline{\text{FFS}}$ to motivate, as a binding principle, that '*all* the relevant information that the data provide ... is contained in the likelihood ratio', and then have us apply that principle also to example $\underline{x = 0}$,

where we must then surely ask (proceeding the other way round): Why not

> draw on example x = 0 to motivate, as a binding principle, that '*none* of the relevant information that the data provide ... is contained in the likelihood ratio', and then have us instead apply that principle to example FFS.

In fact, neither motivation would be correct because, in order to arrive at the correct principles we must draw on that which is common to *both* of the two examples, as well as to *all other* examples where the question 'How might these data have come about?' calls for a statistical answer.

Resumé

The investigative question of science in general is always: 'How might these data have come about?', and a satisfactory answer must necessarily take the form of a model, or several alternative models, in the human mind, where mental correlates of the given data describe how, in the real world, those data might have come about. So we *must* be able to explicate the models we have in mind by simulation, or otherwise we might find ourselves reasoning, as in Sections 1.34 and 1.36, about purported 'experiences' incapable of actually being experienced. This is absolutely crucial and so it must be firmly grasped that *simulation*, and its close relative *experimentation*, are of overwhelming importance in science. It simply will not do to try and evade such importance. So, in order to ensure that a statistical model is scientifically meaningful it must be capable of simulation, and so necessarily comprise a sampling rule and a population of samples arising from that rule. As asserted by Axiom 10.9.1, it is impossible to separate a population from its sampling rule. An investigator must therefore necessarily reason in terms of a population of infinitely many samples that a sampling rule has brought into the human mind. A single one of those many samples provides the mental correlate of the given data (of what *did* take place). That correlate is, however, an unidentified member of a pool called 'the rounding', and so that correlate, being of measure zero, can be dealt with only via the measure of the rounding. However, the measure of the rounding is its measure relative to the entire population, and so willy-nilly we have to reason in terms of *all* the samples the model brings to mind. The many samples can then be partitioned into three disjointed sample pools called 'the left co-ordinate, 'the rounding', and 'the right co-ordinate', respectively, being of measures U, ε, and V, respectively. But, inasmuch as U+ε+V = 1, those measures can at most involve just two independent variables. So we find that the Likelihood Axiom, now stripped of its psychologically persuasive language and 'tricky' supportive examples, simply asserts that:

> only one of the two independent variables, namely the likelihood ratio, conveys all the informative variation.

This assertion is refuted by countless examples such as those given in sections 10.5 and 10.6, and yet further examples to be given in the sequel. The assertion is in fact refuted as follows by the favourite example of advocates of the Likelihood Axiom.

Example 10.9.1

Consider negative binomial sampling as developed in Example 10.2.1, and binomial sampling as developed in Example 10.2.2, with, in each case, $n = 6$ trials comprising $x = 5$ failures and n-x = 1 success. Let θ denote the probability of failure, where we wish

to test the quality of fit of alternative models indexed by θ. Using the number of failures in the sample, X, as our test statistic, the trace of the mental correlate of the datum, X = 5, in the case of negative binomial sampling is found to be given by:

$$[(1-\theta)(1+\theta+\theta^2+\theta^3+\theta^4),\ (1-\theta)\theta^5,\ (1-\theta)(\theta^6+\theta^7+\theta^8+\ldots)],$$

and in the case of binomial sampling by:

$$[1(1-\theta)^6\theta^0+6(1-\theta)^5\theta^1+15(1-\theta)^4\theta^2+20(1-\theta)^3\theta^3+15(1-\theta)^2\theta^4,\ 6(1-\theta)^1\theta^5,\ 1(1-\theta)^0\theta^6].$$

These traces are, for various values of θ, evaluated in Table 10.9.1, with very different results.

Table 10.9.1: Tracing the mental correlate of five failures, in case of negative binomial sampling till just one success, or binomial sampling till just six trials

Pr(Success)	X co-ordinates negative binomial	X co-ordinates regular binomial	L(θ, 0.1 6)
0.75	(0.999, 0.001, 0.000)	(0.995, 0.004, 0.000)	0.01
...	...	...	...
0.55	(0.981, 0.010, 0.008)	(0.931, 0.061, 0.008)	0.15
0.50	(0.969, 0.016, 0.016)	(0.891, 0.094, 0.016)	0.23
0.45	(0.950, 0.023, 0.028)	(0.836, 0.136, 0.028)	0.34
0.40	(0.922, 0.031, 0.047)	(0.767, 0.187, 0.047)	0.46
0.35	(0.884, 0.041, 0.075)	(0.681, 0.244, 0.075)	0.61
0.30	(0.832, 0.050, 0.118)	(0.580, 0.302, 0.118)	0.75
...	...	...	
0.1 6	(0.598, 0.067, 0.335)	(0.263, 0.402, 0.335)	1.00
...	...	...	
0.10	(0.410, 0.059, 0.531)	(0.114, 0.354, 0.531)	0.88
0.05	(0.226, 0.039, 0.735)	(0.033, 0.232, 0.735)	0.58
...	...	...	
0.01	(0.049, 0.010, 0.941)	(0.002, 0.057, 0.941)	0.14

The differences are especially forceful in the range from about $\theta = 0.50$ up to about $\theta = 0.40$, where the negative binomial models fit the data poorly, whereas the binomial models fit the data well, and even very well. As the test statistic, X, is in both cases a most separating test statistic for comparing alternative values of θ, a co-ordination tester cannot possibly agree that the differing importance of the different sampling rules is vacuous. On the contrary, as the different explanatory models described in Table 10.9.1 are easy to simulate – we could point at such simulations and say to our fellow scientists: 'Here see for yourself the alternative explanatory processes of how these given data might have come about'. And we could ask: 'Do you have other explanatory processes in

mind? If so, please go ahead and simulate those processes for us, so that we can understand what you have in mind. How is that? Are you saying that you cannot simulate the processes you have in mind? In that case we cannot take you seriously.'

10.10 Likelihood inference and the *idée fixe*

It is worth noting that we have now met three different theories of inference, each of which try to develop an epistemology that uses just one informative variable amongst U, ε and V. Significance tests try to base an epistemology on the sum of the rounding and the pointing co-ordinate, where that might be defended by arguing that both measures must be small in order to be significant. However, that overlooks the fact that such tests cannot avoid the notion of simultaneous statistical inferences on the part of their knowing subject. Hypothesis tests fall into the same difficulty, inasmuch as they differ from significance tests only in that, instead of having the significance level vary according to the data, they interpret the significance level as a potential error rate, which is made to vary as specified by *their* knowing subject. In both kinds of tests the notion of the probability of a mistaken inference is inspired by the *idée fixe*, and a probability is of course just one number, not two. Inasmuch as a likelihood is not a probability, likelihood inference might seem to escape the *idée fixe*. However, its usage of the term 'likely' betrays psychology that derives from the *idée fixe*, even though such usage does not explicitly involve a knowing subject.

10.11 Absolute quality of fit as opposed to relative quality of fit

In this section we consider how likelihood inference approaches the idea that such and such values of the parameter of interest will by investigation be found to be those that are consistent with the data. Kalbfleisch (1979) develops the idea by evaluating the likelihood of different models, relative to the likelihood of the model most likely in view of the given data. Thus, by solving for values of the model index θ from:

$$L(\theta/\text{the most likely value of } \theta \text{ given the data}) \geq \alpha \text{ for various specified } \alpha,$$

Kalbfleish typically obtains an array of intervals of the form:

$$\theta_1(\alpha) \leq \theta \leq \theta_2(\alpha) \text{ for various specified } \alpha\ (0 < \alpha < 1).$$

He calls such an interval 'a 100α % likelihood interval for θ'. We will now apply this approach to three different examples, and then discuss the results. It turns out that the distinction between absolute quality of fit and relative quality of fit (as previously explained in Example 10.5.1) has awkward implications for likelihood inference.

Example 10.11.1

The likelihood function for the data FFFFFS of Example 10.9.1, both for the negative binomial class of models and the binomial class of models, is given by:

$(1-\theta)^5\theta^1$, where $\theta = \Pr(S)$. This likelihood is maximised by $\theta = 1 \div (5+1) = 1.\ 6.$

So the right-most column of Table 10.9.1 gives the ratio of the likelihood for different values of θ, relative to the likelihood for the most likely value of θ. For instance,

$L(\theta/\theta = 0.1\ 6) = 0.75$ for $\theta = 0.30$.

This tells us that, for the data in hand, the lower bound of the 75% likelihood interval for θ is given by $\theta_1(0.75) = 0.30$. The upper bound of the same interval is given by the other root of the equation:

$(1 - \theta)^5\theta^1 \div (0.8\ 3)^5(0.1\ 6)^1 = 0.75$, which turns out to be $\theta_2(0.75) = 0.07$.

Thus, for the given data, $0.07 \leq \theta \leq 0.30$ is the 75% likelihood interval for θ, meaning that the likelihood of any value in the interval is, *relatively* speaking, at least 75% as likely as the most likely value of θ. But what does that mean in *absolute* terms? On the one hand, if Marie is at least 75% as swift as Sarie and Sarie is slow, Marie is slow. On the other hand, if Sarie is an Olympic champion, Marie might not be slow.

Example 10.11.2

Let an $N(\theta, 1^2)$ random variable, X, where $\theta \in \{\ldots, -6, -3, 0, +3, +6, \ldots\}$, be brought to mind as a class of models that model how a real-world datum, x, might have come about. Let the datum be x = 1.4. We wish to measure the quality of fit of the models indexed by various values of θ. The appropriate elimination pivot is then X-θ, whose distribution is $N(0, 1^2)$, and whose values in standard error units are:

$\ldots, +4.4, +1.4, -1.6, -4.6, \ldots$ for $\theta = \ldots, -3, 0, +3, +6, \ldots$, respectively.

The situations of these test values within the $N(0, 1^2)$ test distribution are given by:

$(1.000, \varepsilon, 0.000)$ for $\theta = -3$,
$(0.919, \varepsilon, 0.081)$ for $\theta = 0$,
$(0.055, \varepsilon, 0.945)$ for $\theta = +3$,
$(0.000, \varepsilon, 1.000)$ for $\theta = +6$, etc. (10.11.1)

Thus co-ordination tests using X, the minimal sufficient statistic for θ, as test statistic, show that none of the various models provide a satisfactory quality of fit. The reader should note in particular that the model indexed by $\theta = 0$ might be described in everyday language as 'a somewhat unlikely model', where the descriptive term 'unlikely' is not being used in the terminological sense used by likelihood inference. Now consider likelihood inference also proceeding from the minimal sufficient statistic. The most likely amongst the possible values of θ, is $\theta = 0$, as that is the value closest to x = 1.4. So, consider the relative likelihood of any hypothesised θ by considering:

$L(\theta/0)$ for $\theta = \ldots, -3, 0, +3, +6, \ldots$, respectively.

$L(\theta)$ is given in kernel form by $\exp[-0.5(1.4-\theta)^2]$. It is often convenient to consider the natural logarithm of the likelihood ratio, given in this case by $-0.5[(1.4-\theta)^2-(1.4-0)^2]$, which equals -8.7 when $\theta = -3$. Thus $L(-3/0) = \exp(-8.7)$, which equals 0.0002. By proceeding in this way we obtain, below, an array of odds-ratio tests of the form: hypothe-

sised θ versus most likely θ. Note that one of these tests is necessarily vacuous because it is not possible for an odds-ratio test to test a hypothesised model against itself:

$$\begin{aligned}&L(\theta/0) = 0.0002 \text{ for } \theta = -3,\\ &L(\theta/0) = 1 \text{ for } \theta = 0 \text{ (the necessarily vacuous test)},\\ &L(\theta/0) = 0.7408 \text{ for } \theta = +3,\\ &L(\theta/0) = 0.0001 \text{ for } \theta = +6, \text{ etc.}\end{aligned} \qquad (10.11.2)$$

At (10.11.2), likelihood inference judges the model indexed by $\theta = +3$ as being 74% as likely as the most likely model. But how likely is the most likely model? At (10.11.1) a co-ordination test in fact has shown that the most likely model is one that every-day language described as a somewhat unlikely model. But an advocate of likelihood inference is not supposed to be capable of knowing that. In such cases the advocates of likelihood inference and Bayesian inference sometimes complain that critics imply they are stupid. 'We are not so stupid,' they are inclined to complain, 'as to overlook discrepancies of magnitude 1.4 and 1.6 standard error units.' However, such complaint is beside the point; no statistician worth his while overlooks discrepancies of such magnitude. The point here is simply that stupidity is in this case being avoided at the cost of incoherence, as advocacy of the sweeping claims made at (10.2.10), (10.2.11), (10.2.12) and (10.2.13), cannot coherently serve reasoning that simultaneously relies on reasons that by those sweeping claims are held to be deluded.

Example 10.11.3

Consider a data set on the presence or absence of certain bacteria in a number, n, of test tubes, each tube containing a given volume of river water, where prior experience shows that the numbers of bacteria in different tubes may be represented as a random sample from a Poisson population with mean θ ($0 < \theta < \infty$). The probability of a 'negative' (i.e. a tube containing no bacteria) is the probability of a zero Poisson count. So, the likelihood of just x positives with just n tubes is obtained from the expression at (10.2.15) as:

$$L(\theta) = \left(1-e^{-\theta}\right)^{x}\left(e^{-\theta}\right)^{n-x}, \text{ because Pr(a zero count)} = e^{-\theta}.$$

Now suppose that there happened to be no negatives (i.e. suppose x = n). Then:

$$L(\theta) = \left(1-e^{-\theta}\right)^{n}, \text{ which} \rightarrow 1 \text{ as } \theta \rightarrow \infty. \qquad (10.11.3)$$

So, loosely speaking, we might say that the most likely value of θ is ∞, although, strictly speaking, this value does not exist, as it is not a value in the parameter space. Nevertheless, the corresponding likelihood is well defined, where (10.11.3) gives its value as unity. So, the likelihood ratios of interest are given by:

$$L(\theta/'\infty') = \left(1-e^{-\theta}\right)^{n} \div (1), \qquad (10.11.4)$$

which, as in previous examples, is interpreted as the likelihood of the θ value under test, relative to the likelihood of the most likely θ value (see Kalbfleisch, p. 27). We note in passing that our expression, *the most likely value,* refers to the value that is conventionally called *the maximum likelihood estimate*, but our expression – though unconventional – better serves the epistemological interpretations that likelihood inference places upon the

likelihood of such and such. Returning to the expression developed at (10.11.4) we note that for any specified percentage, say 75%, likelihood inference would have us interpret a value of θ such that:

$$\left(1-e^{-\theta}\right)^n = 75\%,$$

as a value of θ that is 75% *as likely* as the most likely value of θ. Thus for instance, recalling that 75% is simply a way of denoting 75÷100, we find that if:

$$\text{if } n = 1, \text{ then } \left(1-e^{-\theta}\right)^n = 75\% \text{ for } \theta = 1.4,$$
$$\text{if } n = 10, \text{ then } \left(1-e^{-\theta}\right)^n = 75\% \text{ for } \theta = 3.6, \text{ and}$$
$$\text{if } n = 100, \text{ then } \left(1-e^{-\theta}\right)^n = 75\% \text{ for } \theta = 5.9.$$

We thus obtain the following three 75% likelihood intervals for θ:

$$\text{If } n = 1, \text{ the 75\% likelihood interval is } 1.4 \le \theta < \infty.$$
$$\text{If } n = 10, \text{ the 75\% likelihood interval is } 3.6 \le \theta < \infty.$$
$$\text{If } n = 100, \text{ the 75\% likelihood interval is } 5.9 \le \theta < \infty. \qquad (10.11.5)$$

Kalbfleisch (p. 23) describes any parameter value within a 10% likelihood interval as fairly plausible, and any parameter value within a 50% likelihood interval as quite plausible. So he would be compelled to describe each of the three values, 1.4, 3.6 and 5.9 found at (10.11.5), as quite plausible, and in fact as equally plausible, whereas that cannot possibly be correct. Each of the three values is 75% as likely as a most likely value, but those three most likely values arise from n = 1, n = 10 and n =100 observations, respectively, and so those three most likely values cannot possibly be three equally likely values. The difficulty is insurmountable because the expressions '10% as likely', '50% as likely' and '75% as likely' have *relative* meaning only, whereas the expressions 'fairly plausible' and 'quite plausible' must convey *absolute* meaning, or else be found vacuous. Telling us that a horse named Lucky is almost as swift as a horse named Patch, does not tell us whether or not Lucky is swift.

Discussion of Examples 10.11.1, 10.11.2 and 10.11.3

Consider the concept 'energy'. The latent energy stored in a bag of coal is converted into heat energy by burning the coal, and then converted into kinetic energy when used to propel a locomotive. Our bodies can see the coal, can feel the heat, can ride in the locomotive, but cannot experience energy *as such*. This shows that whilst the world of science is the world as experienced through the human body, science has also to resort to concepts that do not *directly* describe bodily experience. In other words, scientific language employs two different kinds of terms: on the one hand are terms such as 'yellow', 'loud' and 'bitter' used to describe an experienced colour, sound and taste, respectively; they are terms whose meanings are *directly* experiential. On the other hand are terms such as 'energy', 'long-run frequency' and 'statistical independence', whose meanings are not limited to sight, or hearing, or taste, or to any other particular form of bodily perception; they are terms whose meanings are *indirectly* experiential. Any minimally sufficient scientific language requires both kinds of terms, where that leads inexorably to the problem of demonstrating the *necessity* of a proposed term, especially one whose meaning is to be indirectly experiential. In modern physics for instance, the term 'ether' has fallen away as being no

longer needed, whereas the term 'energy' remains indispensable. Along these lines, what can we then say of statistical usages of the term 'likelihood'? That the term is needed as an indirectly experiential one is beyond reasonable contest. We have but to consider the alternative versions of the Neyman-Pearson lemma in order to realise that if the term were to be discarded, the lemma would simply resurface in other, equivalent language, as the necessity of the concept in the context of the predictions, predications and forecasts with which the lemma deals, is demonstrable by simulation. Opposed to that, 'likelihood inference', as expounded by Edwards in the initial 32 pages of his book, tries to persuade us that 'likelihood' can be employed as having evidential meaning. But evidential meaning in science ultimately has to be put to the human body, which compels Edwards to retreat to an explanation in terms of frequencies, which explanation, quoted at (10.5.1) and clarified by way of Example (10.5.1), then limits 'likelihood inference' to *relative* evidential meanings, to comparison of the quality of fit of this model *relative* to that model, and of that model *relative* to the next model. It cannot deal with questions like: 'Do any of these models fit the data well? Or does each and every one of these models fit the data poorly?' This makes it incapable of commencement testing, and therefore inherently susceptible to circular reasoning.

10.12 Can odds-ratio testing usefully supplement other methods of data analysis?

We can now dispense with the Likelihood Axiom because we have developed numerous examples that refute the sweeping nature of its claims. However, the explanation at (10.5.1) implies that in those cases where the likelihood ratio is not statistically vacuous, an odds-ratio test produces scientifically meaningful evidence. Instead of the Likelihood Axiom we must therefore consider a far more modest possibility, namely that an odds-ratio test might informatively add to what a data analyst can learn by other methods. So, to begin with, we note that an odds-ratio test can exist if and only if a corresponding likelihood-ratio ordering exists. An odds-ratio test therefore exists if and only if a corresponding likelihood-ratio co-ordination test exists. Consider any pair of models, M_1 and M_2, capable of being tested by each of the members of such a pair of tests. Let the outcome of the co-ordination test be denoted by:

$$(U_1, \varepsilon_1, V_1) \text{ for } M_1, \text{ and } (U_2, \varepsilon_2, V_2) \text{ for and } M_2. \qquad (10.12.1)$$

The outcome of the corresponding odds-ratio test is then given by $\varepsilon_2 \div \varepsilon_1$, which gives rise to Theorem 10.12.1.

Theorem 10.12.1:

For every given odds-ratio test, a corresponding likelihood-ratio co-ordination test exists; and whatever can be learned from the odds-ratio test, can also be learned from the co-ordination test, but not *vice versa*.

There is worse to come.

Let $O_1, O_2, O_3, \ldots, O_k$, denote the likelihood-ratio ordering in question. Let the mental correlate of the test datum, $X = x$, be situated in ordinal class O_j. Let M_1 and M_2 be

indexed by $\theta = \theta_1$ and $\theta = \theta_2$, respectively. We may, without loss of generality, suppose that pointing is to the right. Then consider the following:

Ordinal class:	O_j	O_{j+1}	O_{j+2}	O_{j+3}	...	O_k
$Pr(X = x \mid \theta = \theta_1)$:	$CL_j(\theta_1)$	$CL_{j+1}(\theta_1)$	$CL_{j+2}(\theta_1)$	$CL_{j+3}(\theta_1)$	...	$CL_k(\theta_1)$
$Pr(X = x \mid \theta = \theta_2)$:	$CL_j(\theta_2)$	$CL_{j+1}(\theta_2)$	$CL_{j+2}(\theta_2)$	$CL_{j+3}(\theta_2)$	...	$CL_k(\theta_2)$

The ordering is a likelihood ratio ordering and, as pointing is to the right, we have:

$$L_j(\theta_2/\theta_1) \leq L_{j+1}(\theta_2/\theta_1) \leq L_{j+2}(\theta_2/\theta_1) \leq L_{j+3}(\theta_2/\theta_1) \leq \ldots \leq L_k(\theta_2/\theta_1). \qquad (10.12.2)$$

Let O_j and O_{j+1} be rounded into a single class, thus placing the mental correlate of the test datum in $O_j \cup O_{j+1}$, where the likelihood ratio for a datum in $O_j \cup O_{j+1}$ is given by:

$$\begin{aligned}
&[CL_j(\theta_2)+CL_{j+1}(\theta_2)] \div [CL_j(\theta_1)+CL_{j+1}(\theta_1)] \\
&= [L_j(\theta_2)+L_{j+1}(\theta_2)] \div [L_j(\theta_1)+L_{j+1}(\theta_1)] \\
&= [L_j(\theta_2/\theta_1)L_j(\theta_1)+L_{j+1}(\theta_2/\theta_1)L_{j+1}(\theta_1)] \div [L_j(\theta_1)+L_{j+1}(\theta_1)]. \qquad (10.12.3)
\end{aligned}$$

The inequalities at (10.12.2) show that this likelihood ratio is bounded from below by:

$$\begin{aligned}
&[L_j(\theta_2/\theta_1)L_j(\theta_1)+L_j(\theta_2/\theta_1)L_{j+1}(\theta_1)] \div [L_j(\theta_1)+L_{j+1}(\theta_1)] \\
&= L_j(\theta_2/\theta_1).
\end{aligned}$$

This shows that the rounding has produced a more discriminatory odds-ratio test. The inequalities at (10.12.2) also show that the likelihood ratio for the rounded datum, i.e. the ratio at (10.12.3), is bounded from above by:

$$\begin{aligned}
&[L_{j+1}(\theta_2/\theta_1)L_j(\theta_1)+L_{j+1}(\theta_2/\theta_1)L_{j+1}(\theta_1)] \div [L_j(\theta_1)+L_{j+1}(\theta_1)] \\
&= L_{j+1}(\theta_2/\theta_1).
\end{aligned}$$

So the reasoning repeats. Thus increasingly discriminatory odds-ratio tests arise when:

O_j is replaced by $O_j \cup O_{j+1}$,
$O_j \cup O_{j+1}$ is replaced by $O_j \cup O_{j+1} \cup O_{j+2}$,
$O_j \cup O_{j+1} \cup O_{j+2}$ is replaced by $O_j \cup O_{j+1} \cup O_{j+2} \cup O_{j+3}$,
... ,

and so on, until the original order class, O_j, is ultimately replaced by:

$$O_j \cup O_{j+1} \cup O_{j+2} \cup O_{j+3} \cup \ldots \cup O_k.$$

The odds ratio produced by this ultimate test is a ratio of significance levels, given by:

$$\begin{aligned}
&Pr(X \in O_j \cup O_{j+1} \cup O_{j+2} \cup \ldots \cup O_k \mid \theta = \theta_2) \div Pr(X \in O_j \cup O_{j+1} \cup O_{j+2} \cup \ldots \cup O_k \mid \theta = \theta_1) \\
&= (\varepsilon_2+V_2) \div (\varepsilon_1+V_1) \text{ in the notation used at (10.12.1).}
\end{aligned}$$

This brings us to Theorem 10.12.2.

Theorem 10.12.2:

> For every given odds-ratio test, a corresponding likelihood-ratio significance test exists; and a more discriminative odds-ratio test is obtained when the given odds ratio is replaced by the corresponding ratio of significance levels.

Note that Theorems 10.12.1 and 10.12.2 concern tests directed at the problem of measuring the quality of fit of alternative statistical models in respect of just one solitary real-world data set. They do not for instance concern the comparison of such tests to hypothesis tests, as the latter kind of test concerns the very different kind of problem of bringing populations made to specification into the real world.

10.13 Which variable is the concomitant?

In an analysis of covariance the concomitant variable is usually of no interest in itself, having been drawn into the analysis as a predictor variable only; our interest is in the predicted variable. The likelihood ratio fulfils a similar role in the Neyman-Pearson lemma, as we are then not interested in the values of the ratio itself; our interest is in the predicted consequences of the ordering it produces. Similarly, in the method of maximum likelihood, interest is in the predicted properties of the resulting estimator, and not in the values of the likelihood itself. The point here is that the uncontroversial role of the likelihood function in statistical theory is reasonably described as that of a concomitant variable that is of no interest in itself. By contrast, likelihood inference controversially tries to elevate likelihood to a position of primary interest. In Example 10.7.3 the likelihood ratio turns out to be a defective concomitant, and in the present chapter we have met a variety of such examples. Those examples do not imply that we should never use the likelihood as a concomitant variable; they imply only that we cannot always rely on the likelihood as a concomitant variable. By way of comparison, we might find an example where last year's wheat yields turn out to be uninformative concomitants for covariance adjustment of this year's turnip yields; but that would not imply that last year's wheat yields are *always* poor concomitants. Turning then to likelihood inference, we might well ask: 'Does that not mistake the concomitant variation for the variation of primary interest? In other words, is that not trying to reverse the roles of the predicted variation and the concomitant variation?' There is only one way in which likelihood inference can escape the criticism implied by these questions, and that is for it to counter our examples where such inference falls short when compared to co-ordination testing, with examples where the opposite would be the case. But there appears to be no such examples.

10.14 A definition of statistical inference

In the usage of present-day statistics, the term 'statistical inference' expresses a vague notion. Despite common usage to the contrary, the notion of repetitive decision-making under statistical risk, when not mistaken for data analysis, may, however, be ruled out, in which case statistical inference refers to a variety of received theories about how conclusions are to be drawn from given data, for which theories the reader will find Definition 10.14.1 useful.

Definition 10.14.1:

> The term *statistical inference* refers to any epistemology that posits a constituent in the role of *a knowing subject*, held to be necessary for drawing from given data, statistical conclusions in the form of 'inferences' that the knowing subject 'infers'.

This definition has in effect already been used in Chapters 4 and 6. In Section 10.12 we found that in current usage the term 'likelihood inference' names a form of test that should more appropriately be named an *odds-ratio test*. Definition 10.14.1 does not apply to either co-ordination tests, or to odds-ratio tests, as such tests do not involve the notion of a knowing subject who infers the inferences and so might or might not infer mistakenly. The definition does not define the term 'inference', and thereby accommodates hypothesis tests, where an inference is a decision by the knowing subject, and significance tests, where an inference is a potential decision by the knowing subject. Chapter 12 will show that the definition also accommodates Bayesian inference, where an inference is a metaphysical belief arrived at by the knowing subject.

Chapter 11

BAYES'S THEOREM

A formula in frequency physics

11.1 Introduction

In this chapter we develop a theorem of Bayes (1763). It has become associated with a long-standing controversy that has nothing at all to do with the theorem *as such*. So, the purpose of this chapter is to make that entirely clear, i.e. to make it clear that Bayes's theorem is a perfectly respectable part of frequency physics. The development will rely largely on examples.

11.2 An explanatory example

Consider three cards: one red on both sides, one red on one side and white on the other side, one white on both sides. One card is drawn at random and laid down on one side. The visible side is red. With what frequency in such cases is the other side white? The answer is not one half; it is one third, as we will now show.

Label the sides as follows:

Card 1: (Red 1, Red 2). Card 2: (Red 3, White 3). Card 3: (White 2, White 1).

Then the appropriate sample space comprises six equally frequent cases, as follows:

Top side:	Red 1	Red 2	Red 3	White 3	White 2	White 1	
Bottom side:	Red 2	Red 1	White 3	Red 3	White 1	White 2	(11.2.1)

Three cases have 'top red', of which one case has 'bottom white'. The problem can also be solved using Bayes's theorem, which we now derive. Take A = 'red top', B = 'white bottom', and consider the elementary form:

$$\Pr(A \,\&\, B) = \Pr(A)\Pr(B|A), \text{ where } B|A \text{ denotes B given A.} \qquad (11.2.2)$$

We want to evaluate Pr(B|A), and we can do that by solving from the equation at (11.2.2) as follows, where the numerical information is obtained by counting 'favourable cases out of six' at (11.2.1):

$$(1\div 6) = (3\div 6)\Pr(B|A), \text{ i.e., } \Pr(B|A) = 1\div 3, \text{ as before.}$$

Interchanging the roles of A and B at (11.2.2), we have:

$$\Pr(A \,\&\, B) = \Pr(B)\Pr(A|B), \qquad (11.2.3)$$

and by inserting the right-hand side of the equation at (11.2.2) instead of P(A & B) in the equation at (11.2.3), we obtain the simplest form of Bayes's theorem, namely:

$$\Pr(B|A) = \Pr(B)\Pr(A|B) \div \Pr(A). \qquad (11.2.4)$$

Denote 'white bottom' as B_1 and 'red bottom' as B_2. Then the equation at (11.2.4) refers to B_1, and the reasoning leading to that equation repeats with B_2 instead of B_1. The equation at (11.2.4) can thus be made more explicit as:

$$\Pr(B_j|A) = \Pr(B_j)\Pr(A|B_j) \div \Pr(A) \text{ for } j = 1, 2. \qquad (11.2.5)$$

Here Pr(A) can be expressed in terms of the other quantities involved at (11.2.5), as:

$$\begin{aligned}\Pr(A) &= \Pr(A \,\&\, B_1) + \Pr(A \,\&\, B_2) \\ &= \Pr(B_1)\Pr(A|B_1) + \Pr(B_2)\Pr(A|B_2).\end{aligned}$$

So, the result at (11.2.5) shows that for discrete sample spaces the theorem takes the form:

$$\Pr(B_j|A) = \Pr(B_j)\Pr(A|B_j) \div [\Pr(B_1)\Pr(A|B_1) + \Pr(B_2)\Pr(A|B_2)] \text{ for } j = 1, 2. \qquad (11.2.6)$$

Interest in Bayes's theorem arises largely because of applications where an A-like event is preceded by an unknown B-like event, such that for each one of the different B-like events the theorem enables us to find the probability that that particular one was the predecessor. Its attraction in the pursuit the statistical inference will-o'-the-wisp is therefore obvious. However, the premises of the theorem must be provided, and there lies the rub. Thus for instance, the following example from Reichenbach (1949) shows the attraction of the theorem for statistical inference, the difficulty of supplying its premises, and that it is sometimes difficult to make appropriate sense out of the answer it provides.

Example 11.2.1

'Mr Smith's gardener is not dependable; the probability that he will forget to water the rosebush during Smith's absence is 2/3. The rosebush is in a questionable condition, anyhow; if watered the probability of its withering is 1/2; if it is not watered, the probability of its withering is 3/4. Upon returning, Smith finds that the rosebush has withered. What is the probability that the gardener did not water the rosebush?'

Let B_1 represent the event that the rosebush is watered, B_2 the event that it is not watered, and A the event that it withers. We want to evaluate $\Pr(B_2|A)$, which the theorem says is:

$$\frac{\Pr(B_2)\Pr(A|B_2)}{\Pr(B_1)\Pr(A|B_1)+\Pr(B_2)\Pr(A|B_2)} = \frac{(2/3)(3/4)}{(1/3)(1/2)+(2/3)(3/4)} = 3/4.$$

This example has a certain charm typical of many textbook examples that are supposed to help us come to grips with Bayes's theorem. And, like many of its kind, one needs to be alert to its premises. How are the given probabilities obtained? We must also be alert as to the possible lack of sense made by the answer obtained. What population of cases does it belong to? In the present instance, a stereotypic array would seem to be the only real-world possibility.

11.3 Some general forms

In the previous section, a datum, A, is used to form an opinion on which B-like unknown was involved. So, for notation more in line with that of previous chapters, let us envisage A as a randomly sampled value, X = x, and B as a parameter, $\theta \in \{\theta_1, \theta_2, \theta_3, ...\}$, which in the case of Bayes's theorem is of course also a random variable. Then a general form corresponding to the special case obtained at (11.2.6) is given by Theorem 11.3.1.

Theorem 11.3.1: Bayes's theorem in case of a discrete parameter space

If $\theta_1, \theta_2, \theta_3, ..., \theta_k$, denotes a mutually exclusive and exhaustive partition of a sample space, such that $\Pr(\theta = \theta_j) \neq 0$ for j = 1, 2, 3, ..., k, and if X denotes a random variable, such that $\Pr(X = x) \neq 0$, then

$$\Pr(\theta = \theta_j | X = x) = \frac{\Pr(\theta = \theta_j)\Pr(X = x | \theta = \theta_j)}{\sum_{i=1}^{k} \Pr(\theta = \theta_i)\Pr(X = x | \theta = \theta_i)} \quad \text{for } j = 1, 2, 3, ..., k.$$

The denominator on the right-hand side of the equation that expresses the conclusion in the theorem is just a normalising constant, say C. So, for practical purposes the recipe given by the theorem can be expressed as:

$$\Pr(\theta = \theta_j | X = x) = C^{-1}\Pr(\theta = \theta_j)\Pr(X = x | \theta = \theta_j) \text{ for } j = 1, 2, 3, ..., k,$$

or simply in mnemonic form as:

$$h(\theta|x) = C^{-1}f(\theta)g(x|\theta), \tag{11.3.1}$$

where h(•), f(•), and g(•) denote the appropriate density functions, and where by summing or integrating over the parameter space, C is obtained as:

$$\sum[f(\theta)g(x|\theta)] \text{ if } \theta \text{ is discrete, and } \int[f(\theta)g(x|\theta)]d\theta \text{ if } \theta \text{ is continuous.} \tag{11.3.2}$$

We refer to f(θ) as *the prior distribution of θ*, and to h(θ|x) as *the posterior distribution of θ*. Since any constant arising from g(x|θ) is absorbed by C^{-1}, g(x|θ) is interpretable as *the likelihood of θ, given the data*.

Example 11.3.1

Player A rolls a billiard ball at random with respect to the length of a billiard table taken to be one unit in length, and removes the ball. Player B must try to discover how A's ball bisected the table, θ:(1-θ), using a geometric sampling model for the number of attempts required when repeatedly also rolling the ball at random, till the ball first comes to rest beyond θ (0 < θ < 1). Let the number of unsuccessful attempts be denoted by X = x. By using the forms at (11.3.1) and (11.3.2), the distribution of θ given x is found to be:

$$h(\theta|x) = C^{-1}(1)[\theta^x(1-\theta)], \text{ and } C = \int_0^1 \theta^x(1-\theta)d\theta \text{ which equals } \frac{1}{(x+1)(x+2)} \quad .$$

So, the distribution is given by:

$h(\theta|x) = (x+1)(x+2)\,\theta^{x}\,(1-\theta)$, whose modus is at $\theta = x \div (x+1)$.

For $x < 1$ the distribution is positively skewed, for $x = 1$ it is symmetric, and for $x > 1$ it is negatively skewed. This was Thomas Bayes's favourite example, except that he took the sampling to be binomial rather than negative binomial. The reader should note that if the given data (x failures and one success) had been obtained by binomial sampling, the same posterior would have been obtained – a fact that has prompted much learned confusion.

Example 11.3.2

Mood, Graybill, and Boes (1974, p. 347) represent $x_1, x_2, x_3, \ldots, x_n$, as the realisation of a random sample from an $N(\mu, 1^2)$ population, where μ is an $N(x_0, 1^2)$ random variable in the first place. Bayes's theorem shows that the posterior distribution of μ is that of an

$N[(x_0+x_1+x_2+x_3+\ldots+x_n) \div (n+1),\ (1 \div \sqrt{n+1})^2]$ random variable.

Thus the contribution of the prior is equivalent to one additional observation.

For a more general result, consider an observation, x, which may be represented as:

$N(\mu, \sigma^2)$ with σ^2 known,

and where μ may be represented as:

$N(\nu, \tau^2)$ with ν and τ^2 both known.

Then the posterior distribution of μ (Kempthorne and Folks, 1971, p. 300) turns out to be:

$$N(\lambda, \phi^2) \text{ where } \lambda = \frac{\left(\frac{x}{\sigma^2} + \frac{\nu}{\tau^2}\right)}{\left(\frac{1}{\sigma^2} + \frac{1}{\tau^2}\right)}, \text{ and } \phi^2 = \frac{1}{\left(\frac{1}{\sigma^2} + \frac{1}{\tau^2}\right)} \quad . \qquad (11.3.3)$$

Remark 1: Apart from the foregoing, if x and ν are uncorrelated estimates of the same quantity μ, the minimum variance linearly combined estimate and its variance are given by the two equations at (11.3.3), respectively. So, the prior contributes information that in this case can formally be expressed as if obtained by sampling. (11.3.4)

Remark 2: If τ^2 takes on some very large value corresponding to an uninformative prior (a so-called diffuse prior), the forms at (11.3.3) allow us to pass from:

X being $N(\mu, \sigma^2)$ to μ being $N(x, \sigma^2)$. (11.3.5)

We will return to these remarks in Chapters 12 and 13.

11.4 Bayes's Theorem in decision-making under risk

For all its mathematical simplicity Bayes's theorem has little to contribute to real-world statistical data analysis. The reason for that is simply that the notion of a prior distribution seldom corresponds to reality. (However, the extent of the theorem's contribution to ivory-tower statistics is an entirely different matter – as we will see later.) The following examples indicate that the theorem, as a formula in frequency physics, can be useful in special circumstances.

Example 11.4.1

Suppose that large consignments of components are regularly received for manufactory. Let 1-θ denote the proportion of defective items in such a consignment. Let *n*-X denote the number of defective items in a pseudo-random sample of size *n*. If X is represented as a Bn(n, θ) random variable, and if the historical record of consignments indicates that θ can be represented as a random variable with the Beta_1(a, b) distribution

$$g(\theta) = [\Gamma(a+b) \div \Gamma(a)\Gamma(b)]\theta^{a-1}(1-\theta)^{b-1}$$

for a and b determined from the historical record, it follows from the recipes at (11.3.1) and (11.3.2) that the posterior distribution of θ is also a Beta_1 distribution given for X = x by:

$$h(\theta|x) = [\Gamma(n+a+b) \div \Gamma(x+a)\Gamma(n-x+b)]\theta^{x+a-1}(1-\theta)^{n-x+b-1}. \qquad (11.4.1)$$

From this it can be shown to follow that:

$$\left(\frac{1-\theta}{\theta}\right)\left(\frac{n-x+b}{x+a}\right) = \text{Snedecor's } F \text{ on } 2(x+a) \text{ and } 2(n-x+b) \text{ df.} \qquad (11.4.2)$$

By inserting, in turn, the upper and lower 5% critical values of *F* into the form at (11.4.2) and solving for θ , lower and upper one-sided 95% confidence limits for θ are obtained. An inspector of consignments, who wants to accept only consignments with $1 - \theta < 0.10$ apart from a 1-in-20 chance of erroneous acceptances, could try to accomplish that by rejecting any particular consignment if the lower limit for θ turns out to be < 0.90 for that consignment, because

1-the lower limit for θ = the upper limit for 1-θ.

Thus the inspector must reject those consignments for which the calculated confidence limit for 1-θ, given by:

$$1 - \frac{(n-x+b)}{(n-x+b)+(x+a)F}$$ for the appropriate critical value of *F*, exceeds 1-θ = 0.10.

For instance, if (x, n) = (5, 100) and (a, b) = (10, 105),

the 5% critical value of *F* on 2(100-5+105) = 400 and 2(5+10) = 30 df. is 1.65,

and so the requisite upper bound is given by:

$$1 - \frac{(100-5-105)}{(100-5+105)+(5+10)(1.65)} = 1-0.89, \text{ i.e. } 0.11, \text{ which is} > 0.10.$$

Such a consignment would therefore be rejected. Alternatively, when inserting $1\text{-}\theta = 0.10$ as the hypothesised value at (11.4.2), the observed $F = 1.48$ on 400 and 30 df., which falls short of $F = 1.65$, thus indicating that $1\text{-}\theta$ larger than 0.10 is required for significance at the 5 % level. So we find again that such a consignment would be rejected. It might seem that for fixed n =100, the inspector's decision rule should be:

Accept consignments with X = 0, 1, 2, or 3. Reject those with X = 4, 5, 6, ...

The presentation of a scientific theory should, however, always exhibit its full potential. So consider a decision-maker, who computes the following significance levels:

$$X = 3, \text{ observed } F = \left\langle \frac{0.10}{1\text{-}0.10} \right\rangle \times \left\langle \frac{202}{13} \right\rangle = 1.7265 \text{ on } 404 \text{ and } 26 \text{ df. SL} = 0.047, \text{ and}$$

$$X = 4, \text{ observed } F = \left\langle \frac{0.10}{1\text{-}0.10} \right\rangle \times \left\langle \frac{201}{14} \right\rangle = 1.5952 \text{ on } 402 \text{ and } 28 \text{ df. SL} = 0.067, \text{ and}$$

who realises that an auxiliary device that produces the numbers Y = 1, 2, 3, ... 20 with equal frequency would enable specs to be met more precisely by means of the following decision rule:

If X = 0, 1, 2, 3, accept the consignment.
If X = 4, draw Y.
If Y = 1, 2, 3, accept the consignment.
If Y = 4, 5, 6, ..., 20, reject the consignment.
If X = 5, 6, 7, ..., 100, reject the consignment. (11.4.3)

It would be absurdly silly for an inspector who wishes to please his masters by realising a 5% error rate as they specified, to shy away from using such a decision rule. So let us proceed to the decision-maker's workplace, where we witness repetitive sampling being used to bring a population (a host of individuals) into the real world, as follows:

Consignment 1: (n, X) = (100, 5). The consignment is rejected.
Consignment 2: (n, X) = (100, 4), Y = 3. The consignment is accepted.
Consignment 3: (n, X) = (100, 4), Y = 7. The consignment is rejected.
Consignment 4: (n, X) = (100, 2). The consignment is accepted.
... (11.4.4)

In respect of this host, the inspector forecasts (correctly):

'Mark my words, less than one in twenty accepted consignments will turn out to be defective.' (11.4.5)

Example 11.4 2

Let us recall that the decision-maker of Example 3.2.2 wanted assurance that the average content of the active ingredient in certain roots containing insecticide would be at least eight parts per hundred apart from a 1-in-100 chance of error. Roots were available in batches and the decision-maker had to accept or reject, per batch, on the basis of the mean amount of active ingredient in nine bundles of roots drawn from that batch. The decision-maker used a historical record showing that the means can be satisfactorily

represented as realisations of independent normal random variables whose expectations represented the amounts of insecticide for corresponding batches, and whose variances were from that record known to be given (in parts per 100 squared) by:

$$\sigma^2 = (3.30 \div \sqrt{9})^2. \qquad (11.4.6)$$

It now appears that an extension of the same historical record might also show that the distribution of the batch means is satisfactorily represented as:

an $N(\nu, \tau^2)$ population with ν and τ^2 both known.

Then the forms at (11.3.3) can be brought to bear by way of achieving the same operating characteristics with fewer than nine bundles per batch, as follows:

Suppose that the values $\nu = 9.120$ and $\tau^2 = 3.63$ are obtained from the historical record. Then the forms at (11.3.3) and the information at (11.4.6) show that the decision-maker can meet specs drawing only n roots per bindle, for n such that

$$\frac{n}{3.30^2} + \frac{1}{3.63} = \frac{9}{3.30^2}, \text{ that is to say, for } n = 6 \text{ bundles,}$$

and where the decision rule is now that any given batch of insecticide roots is classified as acceptable if and only if the weighted linear combination of the mean of six bundles drawn from that batch, ν, and the mean of the population of batches, $\overline{X}$, exceeds

8+Z×(the standard error of the weighted combination), where Z will be exceeded by $N(0, 1^2)$ random variables only one in a hundred times, i.e. Z = 2.330. (11.4.7)

The weights for $\overline{X}$ and ν are given respectively by:

$$\frac{1}{(3.30 \div \sqrt{6})^3} = 0.5510 \text{ and } \frac{1}{3.63} = 0.2755, \text{ totalling } 0.5510+0.2755 = 0.8265.$$

So the weighted combination of the two sources of information is given by:

$$(0.5510\,\overline{X}+0.2755\,\nu) \div 0.8265$$
$$= 0.667\,\overline{X}+3.040$$
when $\nu = 9.120$ is inserted, (11.4.8)

and its standard error is given at (11.3.3) as the square root of the reciprocal of the total weight, which turns out to be 1.100. Therefore the decision rule at (11.4.7) is to accept a given batch when and only when:

$0.667\,\overline{X}+3.04 > 8+2.330(1.100)$, i.e when and only when $\overline{X} > 11.28$.

Let us then proceed to the decision-maker's workplace, where we find that by repetitive pseudo-sampling the decision-maker is bringing a population (a host of individuals) into the real world, as follows:

$\{\overline{X} = 12.22, \text{accept}\}, \{\overline{X} = 9.32, \text{reject}\}, \{\overline{X} = 14.12, \text{accept}\}, \ldots$ (11.4.9)

In respect of this host, the decision-maker forecasts (correctly):

> 'Mark my words, less than one in a hundred accepted batches will turn out to be defective.' (11.4.10)

In principle the forecast can be improved by a randomised decision rule, as X has been measured on a discrete grid of class marks spaced 0.01 units apart.

11.5 Bayes's theorem in data analysis

We now give some examples of the use of Bayes's theorem in data analysis. We remark in advance that the examples will seem to be contrived. There is a reason for that, which is to be discussed in the next section.

Example 11.5.1

With reference to Example 11.4.1, Ms Spare Parts is delighted to learn that, owing to the data representing her consignment, (n, x) = (100, 2), the consignment was accepted, but she wonders whether she might have been rather lucky. So let her ask us to investigate which values of the mean of her particular batch, θ, are consonant with the data. We are of course aware that as a matter of public knowledge the batch mean can be represented as having been sampled at random from a $\text{Beta}_1(a, b)$ population with (a, b) = (10, 105). This brings into the human mind the class of models given at (11.4.1). In fact, the class of present interest is given more specifically by:

$$h(\theta|x = 2) = [\Gamma(n+a+b) \div \Gamma(2+a)\Gamma(n-2+b)]\theta^{2+a-1}(1-\theta)^{n-2+b-1}. \qquad (11.5.1)$$

Even more specifically, inserting n = 100 and (a, b) = (10, 105) at (11.4.2), the posterior distribution of θ for the particular batch of interest is represented by:

$$\left(\frac{1-\theta}{\theta}\right)\left(\frac{203}{12}\right) = \text{Snedecor's } F \text{ on 24 and 406 df.} \qquad (11.5.2)$$

Inserting appropriate F values into the equation at (11.5.2) and solving for the index θ, the position of the mental correlate of the given datum, x = 2, within the correspondingly indexed members of the class of models as hypothesised at (11.5.1) is traced as follows:

The correlate is situated respectively at:

> (0.08, ε, •) and (•, ε, 0.08) within the members indexed by θ = 0.038 and 0.094,
> (0.04, ε, •) and (•, ε, 0.04) within the members indexed by θ = 0.034 and 0.104, and
> (0.02, ε, •) and (•, ε, 0.02) within the members indexed by θ = 0.030 and 0.113.

The facts conveyed by this trace can, if needs be by simulation, be forced upon the human body, where we could then point and say:

> 'See for yourself that any member indexed by a value in excess of θ = 0.10 fits the given data poorly.' (11.5.3)

We note in passing that this does not address Ms Spare Parts's problem. In fact, science in general, statistical or otherwise, can recognise an instance of misleading data, only by further investigation, either via comparable data that might directly expose the misleading data as such, or via further data that might indirectly expose the misleading data as not making sense. (Note that this also applies to an outlier.) In short, scientific investigation can never accomplish more that to establish whether or not a particular model of interest is consonant or dissonant with the accumulated data (or of course with theory arising from the accumulated data).

Comparison of Examples 11.4.1 and 11.5.1

In Example 11.4.1 repetitive sampling takes place in the real world; in Example 11.5.1 repetitive sampling takes place in the human mind only. In Example 11.4.1 the decision-maker brings a population (a host comprising many individuals) into the real world; in Example 11.5.1 the investigator addresses just a single, solitary real-world individual called the 'given data set'. In Example 11.4.1 the decision-maker is not so silly as to shy away from using an auxiliary random device when that helps to meet specs; in Example 11.5.1 the investigator is not so silly as to try to augment the given data with any vacuous 'data' obtained *post hoc* by using a random device. In Example 11.4.1 the decision-maker *forecasts* by saying: 'Mark my words ...'; in Example 11.5.1 the investigator *points* at the relevant facts by saying: 'See for yourself ...'. In Example 11.4.1 the decision-maker tries to realise a specified error rate; in Example 11.5.1 the concept of an error rate does not apply – in fact cannot even be sensibly defined – as there is no sensible room for such a rate when we are pointing at facts. We challenge the reader to note very carefully that each of these five comparisons employs a diagnostic of the distinction between, on the one hand, *the pursuit of statistical knowledge* (data analysis) and, on the other hand, *the use of statistical knowledge* (decision-making under risk), as defined by Definitions 1.2.1 and 1.2.2. Moreover (and this is not to be overlooked), the diagnostics also warn us when the statistical literature fails to draw that distinction, as much of the literature indeed fails to do, with confusing consequences.

Example 11.5.2

Consider the second consignment referred to at (11.4.9) and let us suppose that it came from a particular one amongst various producers of insecticide roots. Call that producer Mr Bane, and let us suppose that he is rather upset about his batch having been rejected, and so decides to investigate the matter by drawing anew six bundles from that particular batch. The new mean that then arises, say $\bar{x} = 12.11$, is a single, solitary real-world item, which brings into the human mind a class of models, as follows:

> The given mean, $\bar{x}$, can be represented as a solitary realisation of an $N[\mu, (3.30 \div \sqrt{6})^2]$ random variable, where μ can in turn be represented as an earlier solitary realisation of an $N(\nu, \tau^2)$ random variable, it being known by virtue of an analysis of the earlier (and very extensive) data set that $\nu = 9.120$ and $\tau^2 = 3.63$.

Referring to the result at (11.4.8), there is thus brought to mind the following elimination pivot for the value of μ in Mr Bane's particular case:

$$Z = \mu - (0.667\bar{x} + 3.040) = \mu - 11.12, \text{ is an } N(0, 1^2) \text{ random variable.} \qquad (11.5.4)$$

For instance, if Z = 1.96, i.e. if the hypothesised μ = 1.96+11.12 = 13.08, then the mental correlate of the given datum, 0.667*x*+3.040 = 11.12, is being placed at (0.025, ε, 0.975) in the hypothesised member of the class of models. Thus, inserting appropriate Z values at (11.5.4) and then solving for μ, the situation of the mental correlate of the given datum within correspondingly indexed members of the class of models, is traced respectively at:

> (0.03, ε, •) and (•, ε, 0.03) within the members indexed by μ = 9.2 and 13.0,
> (0.06, ε, •) and (•, ε, 0.06) within the members indexed by μ = 9.6 and 12.7,
> (0.09, ε, •) and (•, ε, 0.09) within the members indexed by μ = 9.8 and 12.5,
> and on these facts μ = 8 parts per 100 is hardly tenable.

The facts conveyed by this trace can, if needs be by simulation, be forced upon the human body, where we could then point and say:

> 'See for yourself that any member indexed by a value less than μ = 8 fits the given data very poorly.'

What Mr Bane could do with this information is moot, but that need not concern us here.

Comparison of Examples 11.4.2 and 11.5.2

All that was previously stated in the comparisons between Examples 11.4.1 and 11.5.1 also applies to Examples 11.4.2 and 11.5.2, respectively.

11.6 Of how much practical use is Bayes's theorem?

In the case of data analysis the answer to this question is 'very little'. The reason for this is indicated by the contrived nature of Examples 11.5.1 and 11.5.2. They are contrived, as we must explain how the historical record that is requisite for the prior might have come about. It then appears at once that such a record is unlikely to arise otherwise than as an array of recorded activity aimed at some or other form of ongoing quality control. And from that it appears that an investigative activity aimed at trying to form an opinion about how *just one particular term* of such an array might have come about, is rather difficult to motivate. So in fact the practical uses of Bayes's theorem (apart from purely theoretical development) seem to be found mainly in forms of industrial process control. That of course does not concern us; our concern is data analysis, not statistical technology. We note, however, that refined methods for the accumulation and ongoing feed-back of prior information for such technology have been developed under the name *empirical Bayes procedures*. For a definitive account of those procedures, the interested reader is referred to Maritz (1989).

Chapter 12

INVESTIGATION MISTAKEN FOR THE METAPHYSICS OF BELIEF

The Bayesian vicious circle

12.1 Introduction

The term 'Bayesian inference', as conventionally understood, does not refer to the use of Bayes's rule in frequency physics; it refers to a metaphysical view of how an investigator must respond to statistical data. We must begin by explaining the term 'metaphysical' as it is used here, which will show how Bayesian inference has divided the statistical profession into two irreconcilably different schools of thought. In itself that would not necessarily persuade the reader to reject Bayesian inference. Otherwise why would there be the two different schools of thought? However, we also show that Bayesian inferences cannot be simulated, rest on a ramshackle foundation, are incoherent despite claims to the contrary, and cannot avoid circular reasoning. We note in advance that the literature on Bayesian inference, like the rest of present-day statistical literature, often fails to draw a distinction between informative data analysis (the pursuit of knowledge) and decision-making under risk (the use of knowledge). In this book, Bayesian inference concerns the problem of data analysis, unless explicitly stated otherwise.

12.2 Personal probabilities

Consider the proposition 'South Africa's next state president will be a Zulu.' Suppose that if you answer 'yes' or 'no' and your answer turns out to be correct, you receive R100 in prize money. If your answer is 'no', an interpretation of that would be that your *personal probability* of the next president being a Zulu is < 0.50. Let a further offer then be R20 if 'no' proves to be correct and R80 if 'yes' proves to be correct. Then if, in respect of the second offer, your answer is 'yes', the interpretation of that would be that your *personal probability* of the next president being a Zulu is > 0.20. Continuing in this way your *personal probabilities* of yes and no are measured as say 0.35 and 0.65, respectively. Such probabilities are metaphysical; that is to say, they are incapable of being forced onto the human body. There is a simple proof of this, as follows: let a cocktail consist in equal proportions of vermouth, lemon juice and gin. Then inasmuch as mathematical probability is simply the mathematics of proportional constituency we may describe the composition of the cocktail as:

$$\Pr(\text{vermouth}) = \tfrac{1}{3}. \quad \Pr(\text{lemon juice}) = \tfrac{1}{3}. \quad \Pr(\text{gin}) = \tfrac{1}{3}. \tag{12.2.1}$$

And inasmuch as vermouth and gin are alcoholic drinks, which lemon juice is not,

$$\Pr(\text{alcoholic}) = \tfrac{2}{3}\,. \quad \Pr(\text{non-alcoholic}) = \tfrac{1}{3}\,. \tag{12.2.2}$$

Similarly,

$$\Pr(\text{gin}|\text{alcoholic}) = \tfrac{1}{2}. \tag{12.2.3}$$

The forms at (12.2.1, 12.2.2 and 12.2.3) express physical facts, that is to say, express facts that can be forced upon the human body. Now opposed to that, consider a plant whose flowers might turn out to be scarlet, magenta or white, and consider also a person who has no inkling which colour they might turn out to be. Such a person's personal probabilities would then necessarily seem to be:

$$\Pr(\text{scarlet}) = \tfrac{1}{3}\,. \quad \Pr(\text{magenta}) = \tfrac{1}{3}\,. \quad \Pr(\text{white}) = \tfrac{1}{3}\,. \tag{12.2.4}$$

Otherwise that person's probabilities would seem to be idiosyncratic, that is to say, to be incapable of explanation. However, inasmuch as scarlet and magenta are shades of red, the choice at (12.2.4) implies:

$$\Pr(\text{red}) = \tfrac{2}{3}\,. \quad \Pr(\text{white}) = \tfrac{1}{3}\,. \tag{12.2.5}$$

Whereas whatever considerations lead to the personal probabilities at (12.2.4), it would seem that by the self-same considerations, the personal probabilities at (12.2.5) instead should be:

$$\Pr(\text{red}) = \tfrac{1}{2}. \quad \Pr(\text{white}) = \tfrac{1}{2}. \tag{12.2.6}$$

The point here is that when the mathematics of proportional constituency expresses a *physical* constituency, contradictions like that at (12.2.5) and (12.2.6) cannot arise. No matter as to whether such physical constituency is that of mixed drinks, or urns with scarlet, magenta and white chips, or long-run frequencies, or different kinds amongst Imelda Marcos's shoes. *Hence a 'personal probability' is a metaphysical notion*. The foregoing proof can be expressed in different ways. For instance, a non-Zulu president might nevertheless be a Nguni, such as a Swazi or a Xhosa, or might be a non-Nguni, such as an Afrikaner, or a Basuto, or a Shangaan, or a Venda, or ... . Again, consider a random variable whose distribution is uniform, as given by the density

$$f(\theta) = 1 \text{ for } 0 < \theta < 1, \text{ and zero otherwise.} \tag{12.2.7}$$

Let $\theta = \eta(1-\eta)\div 4$. Then the distribution of η is triangular, as given by the density

$$g(\eta) = (1-2\eta)\div 4 \text{ for } 0 < \eta < \tfrac{1}{2}\,, \text{ and zero otherwise.} \tag{12.2.8}$$

So if the density at (12.2.7) is supposed to characterise ignorance about the value of θ, the corresponding ignorance about the value of η would have to be characterised, not by the density at (12.2.8), which would then be expressing knowledge about the value of η, but by

$$g(\eta) = 2 \text{ for } 0 < \eta < \tfrac{1}{2}, \text{ and zero otherwise.}$$

To have thus mistaken physics for metaphysics has inescapable consequences because metaphorically speaking it hangs an albatross from the neck of a Bayesian inference. It

has for instance been proposed that for sampling from an $N(\mu, 1^2)$ population, prior ignorance about the value of μ is to be expressed by the *improper prior* according to which the density of μ is 'zero from $-\infty$ to $+\infty$'. As such a prior is in effect 'infinitely diffuse', as explained at (11.3.5), it thēn follows that

> x from an $N(\mu, 1^2)$ population, amounts to μ from an $N(x, 1^2)$ population. (12.2.9)

But the notion of *a constant* having a density of 'zero from $-\infty$ to $+\infty$' is physically meaningless. So, despite all its mathematical elegance, a physically meaningless prior will result in (and this is the albatross) *a posterior that is also physically meaningless; a posterior that is incapable of being simulated; a posterior whose meaning cannot be forced upon the human body*. It will be found that the albatross is inescapable. So let us underscore this, by introducing the notation bPr(•) for personal probability (i.e. for 'belief Pr(•)') as opposed to Pr(•) for long-run frequency. At (12.2.9) for instance,

$$\Pr(x > 1.645) = 0.05, \text{ and } b\Pr(\mu > 1.645) = 0.05,$$

with widely different meaning, despite formal similarity.

12.3 A METAPHYSICAL VIEW

In Bayesian inference the index of a class of models *is treated* as if it were a random variable, when it *is*, by universal agreement, a fixed constant. This might seem self-contradictory, yet *technically* its justification is extremely simple, as follows:

> 'Yes we know full well that θ is indeed a constant; however, the probabilities we attach to its different possible values express our beliefs in terms of our personal probabilities that those would be the correct values.' (12.3.1)

A Bayesian can thus introduce *prior* personal probabilities and, given suitable data, can convert prior personal probabilities to *posterior* personal probabilities by means of Theorem 12.3.1, which is often attributed to Thomas Bayes on dubious grounds.

Theorem 12.3.1: (possibly due to Bayes, and possibly not)

Let $\theta_1, \theta_2, \theta_3, \ldots, \theta_k$, denote the values of the index, θ, of a class of models.
Let $b\Pr(\theta = \theta_j) \neq 0$ for j = 1, 2, 3, ..., k.
Let X denote a random variable such that $\Pr(X = x) \neq 0$.
Then

$$b\Pr(\theta = \theta_j | X = x) = \frac{b\Pr(\theta = \theta_j)\Pr(X = x | \theta = \theta_j)}{\sum_{i=1}^{k} b\Pr(\theta = \theta_i)\Pr(X = x | \theta = \theta_i)} \text{ for } j = 1, 2, 3, \ldots, k.$$

The reader might find it revealing to consider Example 11.2.1 in the present context, and to interpret the probabilities of the gardener forgetting (or not) to water the rosebush, and of the rosebush withering (or not), as the personal probabilities of Mr Smith. The point to note is that such an interpretation is psychologically persuasive, especially when the

means of replacing those personal probabilities with apodictic probabilities is daunting, or impractical, or not readily available, or worse. And that precisely is the attraction that Theorem 12.3.1 has for those decision-makers who must rely largely on archival resources for requisite empirical information. Consider for instance how an agricultural economist must decide on the alternatives of stock farming with goats, or springbuck, or ostriches in the Three Sisters area of the Great Karoo. Much archival information is available, *but not all that is needed*. What would a realistic stocking rate, θ, for springbuck be? The available empirical data might be insufficient, and any prospect of further empirical determination not an option. So, resorting instead to the extraction of expert belief is attractive. The argument in favour of this is that such belief, albeit based on vague knowledge, should in the absence of facts be *used* – the word 'used' is crucial, as it signals the use of knowledge, as opposed to the pursuit of knowledge. So, whilst we may tolerate these ideas in respect of a decision-maker's needs, we must firmly resist the notion that belief in the sense that here concerns us, can be interpreted as a kind of knowledge rather than of conjecture. We must also resist the notion that such a conjecture could sensibly refer to a distribution other than one with physical meaning. To that end let us revisit Mr Smith's gardener, and let us suppose that Mr Smith has recorded that on 12 of 35 occasions the gardener forgot to water the rose bush. And let us, for the sake of argument, also suppose that Mr Smith obstinately tries to grow a rose bush in an unsuitable place, despite a record showing that previously four of nine rose bushes perished there, despite having been watered, and six of seven rose bushes perished there, perhaps owing to not having been watered. Then, by using estimated probabilities (frequencies) in place of the purported probabilities used in the calculation at (11.2.7), we can estimate the probability (the long-run frequency) with which the gardener would have forgotten to water the rosebush in those cases in which it withered, as follows:

$$\frac{\Pr(B_2)\Pr(A|B_2)}{\Pr(B_1)\Pr(A|B_1) + \Pr(B_2)\Pr(A|B_2)} = \frac{(12/35)(6/7)}{(23/35)(4/9) + (12/35)(6/7)} = 1/2. \qquad (12.3.2)$$

Let us now ask our Bayesian friends:

> 'With reference to your Theorem 12.3.1, can we point at our calculation at (12.3.2) as an exemplification of the kind of interpretation you would have us make of your use of that theorem?'

It will be found that the answer is 'No, not at all!' On this Bayesians are adamant. So it would seem that though we are dealing with a familiar logic, we were not supposed to interpret it as at (12.3.2). Let us then ask Miss Evelyn Rosenthall (1965) to refresh our understanding of the incompleteness of any logic. She considers (p. 205) the system of logic displayed in Table 12.3.1, and explains: 'Mathematicians would love to be able to prove that the sets of axioms they use are consistent and that no contradictions can ever arise ... However, in 1931 ... Kurt Goedel proved that this ambition can never be realized: it is impossible to prove that a set of axioms for a system is consistent without going outside the system into a more complex one whose consistency is equally doubtful.' By 'system' she means system of logic. However, by going outside *logic* itself into *science*, it can be shown that the foregoing system is consistent. Miss Rosenthall does this by pointing at the model reproduced in Figure 12.3.1(a), where rasks are lines and syrls are points. The model is

Table 12.3.1: A logic to which different scientific meanings can be adjoined

Undefined terms: *rask, syrl*

Axioms: (1) Each syrl has exactly one rask in common with every other syrl.
(2) Each rask is on exactly two syrls.
(3) There are exactly four syrls.

Among the theorems that can be deduced are:

Theorem 1: There are exactly six rasks.

Theorem 2: Each syrl contains exactly three rasks.

Theorem 3: For each rask, there is exactly one other rask not on the same syrl.

a triangular pyramid. We note that here, as in all science, any proof must ultimately appeal to the human body as the overriding arbiter of truth. Miss Rosenthall underscores this by way of a second, quite different, physical interpretation she obtains by pointing at the model reproduced in Figure 12.3.1(b), where rasks are points and syrls are lines. From our point of view the crux of the matter is that no matter how persuasive a mathematical argument might be, ultimately *scientific* meaning and truth rely on *pointing*. And so, no matter how persuasive the metaphysics of a Bayesian use of Bayes's theorem might be, ultimately the test of whether or not it has value for investigative science must necessarily be sought in its answer to the question 'What are you *pointing* at?' Miss Rosenthall has taught us that a formal system can be made to point at different physical (bodily) experiences, and we have clearly understood that our Bayesian friends are adamant that their use of mathematical probability does not point at experiences of the kind we introduced at (12.3.2). So let us accept that the axioms of Kolmogoroff (1933) lead to a purely formal logic called 'mathematical probability', which logic makes no reference to the physical world, but can, as we have seen by way of different applications to the physics of proportional constituency, be made to point at physical experiences other than frequencies, experiences such as the proportional constituency of mixed drinks. So let us ask our

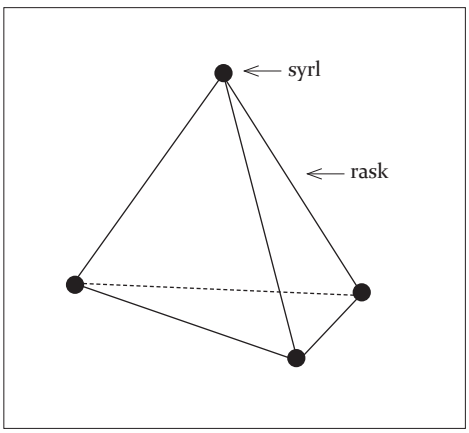


Figure 12.3.1(a): In this figure, syrls are points and rasks are lines

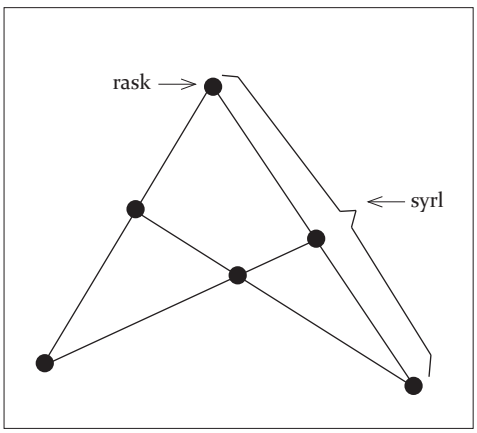


Figure 12.3.1(b): In this figures, rasks are points and syrls are lines

Bayesians friends: 'As you are not pointing at frequencies, what *are* you pointing at?' Alas! After more than two centuries of debate and development, Bayesian inference fails to produce an answer that is understandable to the majority of statisticians. In fact (and this is the albatross) the metaphysical nature of Bayesian inferences cannot be shaken off: try to replace the belief probabilities by resulting expected values, then those are belief expectations; try to replace those expectations with resulting losses or gains, then those are belief losses or belief gains. And so on. No matter what, the albatross keeps on surfacing whenever we ask: 'How is that to be put to the human body?'

Let us note in passing that also, no matter how persuasive the metaphysics of a Bayesian use of 'Bayes's theorem' might be, ultimately the test of whether or not it has value for scientific technology, that is to say for the *use* of scientific knowledge, must necessarily be sought in its answer to the question 'What are you *forecasting*?

12.4 On the notion of 'vague knowledge'

It is well worth noting that the notion of 'vague knowledge' is an offshoot of the *idée fixe*, as the latter has conditioned some of us to accept statements such as '... and there is all of a 95% probability that this conclusion is correct', where that is but a short step removed from statements of the kind '... my personal odds are 19:1 in favour of that'. Thus the *idée fixe* has lead to a general failure to recognise that metaphysical odds cannot be explained in terms of belief; such an 'explanation' is a cop-out, because a scientific explanation, even when rejected, must at the very least try to explain physical (bodily) experience. In any case, it is meaningless to imply that one can 'know something 95%'; one might know something, or one might not know something, but the idea of one's having a degree of knowledge, is utter nonsense. That does not imply that knowledge cannot be fragmentary in the sense of not knowing *all* of certain requisite facts. After all, an investigator is someone who is short of facts, that is to say, who is in a state of partial ignorance.

12.5 A ramshackle foundation

The odds on $\theta = \theta_j$ as opposed to $\theta = \theta_k$ is given according to Theorem 12.3.1 by:

$$\frac{b\ \Pr(\theta = \theta_j)}{b\ \Pr(\theta = \theta_k)} \times \frac{\Pr(X = x \mid \theta = \theta_j)}{\Pr(X = x \mid \theta = \theta_k)} = \frac{\mathrm{b}\ \Pr(\theta = \theta_j \mid X = x)}{\mathrm{b}\ \Pr(\theta = \theta_k \mid X = x)}\ . \tag{12.5.1}$$

This well-known form states:

> The prior odds × The likelihood ratio = The posterior odds.

It appears at once that Bayesian inference is subject to the shortcomings of odds-ratio testing as exemplified by the statistically vacuous likelihood ratios of Section 10.7. Let us consider two further examples of such vacuous likelihood ratios.

Example 12.5.1

Let [x, Pr(X = x | θ = 1)] and [x, Pr(X = x | θ = 2)] be given by the singletons

(-3, 0.01), (-2, 0.01), (-1, 0.01), (0, 0.02), (+1, 0.15), (+2, 0.35), (+3, 0.45),
and
(-3, 0.45), (-2, 0.35), (-1, 0.15), (0, 0.02), (+1, 0.01), (+2, 0.01), (+3, 0.01),

respectively. Let the given data be X = 0. Then co-ordination testing finds the mental correlate of the given datum situated at

(0.03, 0.02, 0.95) and at (0.95, 0.02, 0.03)

within the singletons indexed by θ = 1 and θ = 2, respectively, whereas Bayesian inference, using the form at (12.5.1) finds

$$\frac{b\ \Pr(\theta = \theta_1)}{b\ \Pr(\theta = \theta_2)} \times \frac{0.02}{0.02} = \frac{\text{b}\ \Pr(\theta = \theta_1)}{\text{b}\ \Pr(\theta = \theta_2)}$$

.That is to say, finds: 'The data are vacuous.'

Co-ordination testing therefore learns that either model fits the given data poorly, whereas Bayesian inference is incapable of learning anything at all from those same data.

Example 12.5.2

Basu (1975, p. 2) considers an urn that contains N tickets numbered consecutively as θ+1, θ+2, θ+3, ..., θ+N, where N is a known integer, θ is an unknown integer, and a random sample of *n* tickets is drawn without replacement. Then the minimal sufficient statistic for θ is the order-statistical pair $[X_{(1)}, X_{(n)}]$ which will presently be found to convey both determinate as well as statistical information about the possible value of θ. Let the given data be

N = 10, n = 4, and $[x_{(1)}, x_{(4)}] = (13, 17)$.

What can these data tell us about θ? First we consider how co-ordination tests address this question, next how Basu would have likelihood inference address the question, and then how Bayesian inference would have us address the question.

Co-ordination testing: As $x_{(1)}$ cannot be less than θ+1, and $x_{(n)}$ cannot be more than θ+10, we have (this conveys *determinate* information) that

$13 \geq \theta+1$ and $17 \leq \theta+10$, i.e. $12 \geq \theta$ and $7 \leq \theta$.

So, owing to the given circumstantial details, there comes into the human mind a class comprising just six possible models, indexed respectively by

θ = 7, 8, 9, 10, 11, 12, (12.5.2)

and comprising frequencies (this conveys *statistical* information) as in Table 12.5.1.

Table 12.5.1: A class of models for an order-statistical datum $[x_{(1)}, x_{(4)}] = (13, 17)$, in a random sample of n = 4 tickets drawn without replacement from N = 10 tickets numbered θ+j for j = 1, 2, 3, ..., 10, where $7 \le \theta \le 12$. The frequencies of possible outcomes are displayed here as numbers of favourable cases out of 210 possible cases.

	$X_{(4)}$	θ+4	θ+5	θ+6	θ+7	θ+8	θ+9	θ+10	Total
$X_{(1)}$									
θ+1		1/210	3/210	6/210	10/210	15/210	21/210	28/210	84/210
θ+2			1/210	3/210	6/210	10/210	15/210	21/210	56/210
θ+3				1/210	3/210	6/210	10/210	15/210	35/210
θ+4					1/210	3/210	6/210	10/210	20/210
θ+5						1/210	3/210	6/210	10/210
θ+6							1/210	3/210	4/210
θ+7								1/210	1/210
Total		1/210	4/210	10/210	20/210	35/210	56/210	84/210	1

The table shows, for instance, that for a co-ordination test of θ = 9 as a hypothesised model, and using $X_{(4)}$ as the test statistic, the situation of the mental correlate of the test datum, i.e. of $x_{(4)} = 17$, is to be found

at [(1+4+10+20)/210, 35/210, (56+84)/210] in the test distribution.

Using such calculations, we develop the pair of traces displayed in Table 12.5.2.

Table 12.5.2: A pair of traces giving the co-ordinates of the mental correlates of a test datum pair $[x_{(1)}, x_{(4)}] = (13, 17)$ in test distributions indexed by θ ($7 \le \theta \le 12$)

Trace for the suite of $X_{(1)}$ tests	Trace for the suite of $X_{(4)}$ tests
[θ = 7, (0.976, 0.019, 0.005)]	[θ = 7, (0.600, 0.400, ∅)]
[θ = 8, (0.929, 0.048, 0.024)]	[θ = 8, (0.333, 0.267, 0.400)]
[θ = 9, (0.833, 0.095, 0.071)]	[θ = 9, (0.167, 0.167, 0.667)]
[θ = 10, (0.667, 0.167, 0.167)]	[θ = 10, (0.071, 0.095, 0.833)]
[θ = 11, (0.400, 0.267, 0.333)]	[θ = 11, (0.024, 0.048, 0.929)]
[θ = 12, (∅, 0.400, 0.600)]	[θ = 12, (0.005, 0.019, 0.976)]

The table shows that the $X_{(1)}$ test is a unilateral right-sensitive test for small θ, and the $X_{(4)}$ test is a unilateral left-sensitive test for large θ. The $X_{(1)}$ test shows that the singleton for θ = 8 fits the data poorly, and that the singleton for θ = 7 fits the data very poorly. The $X_{(4)}$ test shows that the singleton for θ = 11 fits the data poorly, and that the singleton for θ = 12 fits the data very poorly. These findings are not qualified as being in any sense probable or likely, as they are simply facts. They can by simulation be forced upon the human body,

and as such they are beyond all reasonable contest. Furthermore, they cannot appropriately to be described as inferences, because they do not partake of a knowing subject, as no knowing subject has inferred them.

Likelihood inference: Basu (1975, p. 1), in part one of his essay (the part that concerns us) tries to formulate 'the notion of "statistical information generated by a data (set)" in terms of some intuitively appealing principles of data analysis'. He remarks that he 'comes out very strongly in favour of the unrestricted likelihood principle' (he means, as we saw in Section 10.8, 'the strong likelihood axiom') and his principle example (one could almost say his only example) is the present one, in respect of which he remarks that 'we know without any shadow of a doubt' that the true value of θ belongs to the set

$\{x_{(1)}\text{-}1, x_{(1)}\text{-}2, x_{(1)}\text{-}3, \ldots, x_{(1)}\text{-}N+[x_{(n)}\text{-}x_{(1)}]\}$, i.e. (12.5.3)

{13-1, 13-2, 13-3, ..., 13-10+[17-13]} when N = 10 and n = 4, i.e.

{12, 11, 10, ..., 7} as we have already seen at (12.5.2).

So he asserts the likelihood axiom by denoting the set at (12.5.3) as A, and saying that in the present case:

> 'the likelihood function ... is "flat" over the set A and is zero outside (a situation that is typical of all survey sampling set-ups) and this means that the sample ... "supports" each of the points in the set A with equal intensity.' (12.5.4)

Basu re-asserts the axiom, by denoting $x_{(n)}\text{-}x_{(1)}$ at (12.5.3) as *m*, and saying:

> 'Once ... the sample is recorded, the magnitude of the information obtained depends on the integer *m* (which varies from sample to sample) rather than on the constant *n*.' (12.5.5)

The parenthetic remark at (12.5.4) is significant, however, for the moment that need not concern us. What must concern us here, is that the assertions at (12.5.4) and (12.5.5) are demonstrably fallacious. Let us begin with the curious notion that, once the data are before us, the information obtained does not depend on the sample size. Let N = 10 and $[x_{(1)}, x_{(n)}]$ = (13, 17) just as before, but now with *n* = 3. Then as before 'we know without any shadow of a doubt' that the true value of θ belongs to the set

$\{x_{(1)}\text{-}1, x_{(1)}\text{-}2, x_{(1)}\text{-}3, \ldots, x_{(1)}\text{-}N+[x_{(n)}\text{-}x_{(1)}]\}$, i.e.

{13-1, 13-2, 13-3, ..., 13-10+[17-13]} when N = 10 and n = 3, i.e.

{12, 11, 10, ..., 7} precisely as before.

And the likelihood is 'flat' over this self-same set, precisely as before. So for Basu to uphold his axiom, he must conclude precisely as before at (12.5.4) and (12.5.5). But we must disagree, because a distinctly different class of models, as displayed in Table 12.5.3, has now been brought into the human mind.

Table 12.5.3: A class of models for an order-statistical datum $[x_{(1)}, x_{(3)}] = (13, 17)$, in a random sample of n = 3 tickets drawn without replacement from N = 10 tickets numbered θ+j for j= 1, 2, 3, ..., 10, where $7 \leq \theta \leq 12$. The frequencies of possible outcomes are displayed here as numbers of favourable cases out of 120 possible cases

	$X_{(3)}$	θ+3	θ+4	θ+5	θ+6	θ+7	θ+8	θ+9	θ+10	Total
$X_{(1)}$										
θ+1		1/120	2/120	3/120	4/120	5/120	6/120	7/120	8/120	36/120
θ+2			1/120	2/120	3/120	4/120	5/120	6/120	7/120	28/120
θ+3				1/120	2/120	3/120	4/120	5/120	6/120	21/120
θ+4					1/120	2/120	3/120	4/120	5/120	15/120
θ+5						1/120	2/120	3/120	4/120	10/120
θ+6							1/120	2/120	3/120	6/120
θ+7								1/120	2/120	3/120
θ+8									1/120	1/120
Total		1/120	3/120	6/120	10/120	15/120	21/120	28/120	36/120	1

The pairs of co-ordination tests based on the respective elements of the minimal sufficient statistic for θ, that is to say, based on the respective elements of the pair $[X_{(1)}, X_{(3)}]$, as displayed in Table 12.5.4, differ from those previously obtained with $n = 4$.

Table 12.5.4: A pair of traces giving the co-ordinates of the mental correlates of a test datum pair $[x_{(1)}, x_{(3)}] = (13, 17)$ in test distributions indexed by θ ($7 \leq \theta \leq 12$)

Trace for the suite of $X_{(1)}$ tests	Trace for the suite of $X_{(3)}$ tests
[θ = 7, (0.917, 0.050, 0.033)]	[θ = 7, (0.700, 0.300, ∅)]
[θ = 8, (0.833, 0.083, 0.083)]	[θ = 8, (0.467, 0.233, 0.300)]
[θ = 9, (0.708, 0.125, 0.167)]	[θ = 9, (0.292, 0.175, 0.533)]
[θ = 10, (0.533, 0.175, 0.292)]	[θ = 10, (0.167, 0.125, 0.708)]
[θ = 11, (0.300, 0.233, 0.467)]	[θ = 11, (0.083, 0.083, 0.833)]
[θ = 12, (∅, 0.300, 0.700)]	[θ = 12, (0.033, 0.050, 0.917)]

We note for instance that if the norm for 'a poor fit' is set at

(the statistical rounding)+(the pointing co-ordinate) < 0.100,

then

θ = 7, 8, 11, 12 are eliminated in Table 12.5.2, whereas only
θ = 7, 12 are eliminated in Table 12.5.4.

To summarise:

At (12.5.4) we are being told (correctly) that the likelihood axiom would have us conclude (wrongly) that on the given data, the values θ = 7, 8, 11, ..., 12, are equally well supported – wrongly so, as shown in Table 12.5.2.

At (12.5.5) we are told (correctly) that the likelihood axiom would have us conclude (wrongly) that, on the given data, correct conclusions would not depend on the sample size – wrongly so, as shown in tables 12.5.2 and 12.5.4.

Bayesian inference: It follows from the form at (12.5.1) that Bayesian inference cannot rectify wrongs that arise from the likelihood axiom.

12.6 A DOUBLY INCOHERENT DISCOURSE

Scientists tend to create jargon by adopting everyday words – sometimes with misleading consequences. In this section we consider the usage of the word 'coherent' in the literature of Bayesian inference, and we give reasons for being very critical of that usage. In fact, the following example shows that usage to be downright misleading.

Example 12.6.1

The data shown in Table 12.6.1 are taken from Snedecor and Cochran (1989, p. 205).

Table 12.6.1: Frequency distribution of run lengths in 207 runs of diseased plants

Length of run	0	1	2	3	4	5	6	Total
Observed frequency	164	33	9	1	0	0	0	207
Expected frequency	164.17	33.97	7.03	1.45	0.30	0.06	0.01	206.99
(Obs.-Exp.)²÷Exp.	0.00	0.03	0.55	0.14	0.30	0.06	0.01	1.09
(When pooling data for run-lengths > 2, chi-square has 4-1-1 degrees of freedom.)								

They model the data as originating from a series of independent Bernoulli trials, and refer to an array of consecutive successes preceded and followed by a failure as a run of successes. They refer to the number of successes that comprise the run as its length. Each run is bracketed by the form FS ... F. Thus, FFSFSSSFSSF involves three runs of lengths 0, 2 and 1, respectively. The probability of success, θ, is taken to be constant from trial to trial. The probability of a run of length z is thus

$$\theta \times \theta \times \theta \times \ldots \times \theta \times (1-\theta) = \theta^{z}(1-\theta) \ldots .$$

So, if a data set consists of x_0, x_1, x_2, ..., runs of lengths 0, 1, 2, ..., respectively, in a particular order of occurrence, the probability of the data set is given by

$$[\theta^0(1-\theta)]^{x_0}[\theta^1(1-\theta)]^{x_1}[\theta^2(1-\theta)]^{x_2}\ldots = \theta^{x-n}(1-\theta)^n, \qquad (12.6.1)$$

where x denotes total number of successes, and n the number of runs ($x \geq n$). The maximum likelihood estimate of θ, i.e. the value of that maximises the probability at (12.6.1), is given by

$$\hat{\theta} = \frac{x-n}{x}, \text{ which equals } \frac{261-207}{261} \text{ for Snedecor and Cochran's data.} \qquad (12.6.2)$$

As $\hat{\theta}$ is a minimal sufficient statistic for θ, the class characteristic is obtained when we condition on the value of that statistic. Hence, in the notation of Table 12.6.1, the residual quantities

$$\text{(Obs.-Exp.)}\div\sqrt{\text{Exp.}}, \text{ calculated by inserting } \theta = \frac{261-207}{261}, \qquad (12.6.3)$$

represent the class characteristic. We now consider at first how co-ordination testing proposes to develop models to represent the data, and then how Bayesian inference proposes to develop such models.

Co-ordination testing

Under Snedecor's and Cochran's hypothesised class of models, the sum of the squares of the residuals defined at (12.6.3) provides an appropriate statistic to test the quality of fit of the class characteristic. Under that hypothesised model, it is distributed approximately as a chi-square random variable on 2 df. The test datum equals 1.09 as calculated in Table 12.6.1, and its mental correlate is situated at

$$(0.42, \varepsilon, 0.58^*) \text{ in the test distribution.} \qquad (12.6.4)$$

The class characteristic, as tested, fits the given data well. Next we need elimination tests to weed out untenable values of θ. Solving for x from the equation at (12.6.2), we find

$$x = \frac{n}{1-\hat{\theta}}, \text{ which shows that } x \text{ is a one to one transformation of } \hat{\theta}.$$

So x is an alternative form of the minimal sufficient statistic for θ. The co-ordinates of the mental correlate of x for different θ thus provide a uniformly most separating suite of co-ordination tests for the comparison of any pairs of alternative values of θ. The distribution of x for any θ is a member of the negative binomial class

$$\Pr(X = x \mid \theta) = \binom{x-1}{n-1}\theta^x(1-\theta)^n \text{ for } x = 0, 1, 2, \ldots (0 < \theta < 1). \qquad (12.6.5)$$

So, the expected value and the variance of the mean number of successes per trial are

$$E\left(\frac{x}{n}\right) = \left(\frac{\theta}{1-\theta}\right) \text{ and } Var\left(\frac{x}{n}\right) = \frac{1}{n}\left(\frac{\theta}{(1-\theta)^2}\right), \text{ respectively.}$$

Using the maximum likelihood estimate of θ to estimate the standard error of the mean number of successes per trial, we obtain by $N(0, 1^2)$ approximation

$$\left[\left(\frac{x}{n}\right) - \left(\frac{\theta}{1-\theta}\right)\right] \div \sqrt{\frac{1}{n}\left(\frac{\hat{\theta}}{(1-\hat{\theta})^2}\right)} \text{ as an elimination pivot for } \left(\frac{\theta}{1-\theta}\right).$$

Using Theorem 2.5.1, we then obtain pairs of elimination bounds for θ. For instance

$\theta = 0.17$ and 0.24 at $(0.10, \varepsilon, 0.90)$ and $(0.90, \varepsilon, 0.10)$, respectively,
$\theta = 0.16$ and 0.25 at $(0.05, \varepsilon, 0.95)$ and $(0.95, \varepsilon, 0.05)$, respectively,
$\theta = 0.15$ and 0.26 at $(0.01, \varepsilon, 0.99)$ and $(0.99, \varepsilon, 0.01)$, respectively,
where each pair straddles the maximum likelihood estimate, $\hat{\theta} = 0.21$. (12.6.6)

At (12.6.4) and (12.6.6) we point at physical facts of fit – facts that, if needs be, can by simulation be forced upon the human body. The facts pointed out at (12.6.6) stand entirely apart from those pointed out at (12.6.4). When we point at a spoor in the veld saying 'The shape of this looks like the spoor of a nyala rather than a kudu, but if it were the spoor of a nyala, it would be that of an unusually large one', the phrase 'if it were' shows that the datum of size stands apart from the datum of shape, and that our predication of the datum of size *does not assume* we have correctly predicated the datum of shape. Similarly, the facts pointed out at (12.6.6) *do not assume* anything at all about the facts pointed out at (12.6.4). That is so because the class of models was analysed into the two entirely separate entities called the 'class characteristic' and the 'array of members', and the data were analysed into correspondingly separate entities, thus enabling two entirely separate investigations culminating in evidence separately cited at (12.6.4) and (12.6.6), respectively. Moreover, statistical co-ordinates are not in any sense intended to convey the notion of 'probable truth'. On the contrary, if we must *perforce* attach a probability of truth to the fact of fit that is conveyed at for instance (12.6.4), we would declare:

> 'The "probability" that that fact of fit is true = 1, because it is after all *a fact'*.

But we would do so under protest, as there is nothing whatsoever to gain by dragging an irrelevant concept into our development. We must rather heed William of Occam when he says: 'It is vain to try to do with more, what can be done with fewer.'

Bayesian inference

One has to begin with a commencement test, and for the present purposes the chi-square test introduced in Table 12.6.1 will do as well as any other. We have previously seen that such a test presents an insurmountable problem for 'frequentist inference' as any acceptable commencement is inexorably open to the possibility of a Type II error of unknown kind, and hence of unknown 'rate'. It might seem that Bayesians can escape the problem, as their probabilities are derived from introspection. So it might seem that just as at (12.3.1) here, too, they could say:

> 'Yes of course we know full well that the class characteristic, whatever it might be, was not sampled from a population of alternative class characteristics. However,

the probabilities we attach to different possibilities express our beliefs in terms of our personal probabilities that those might amount to the actual state of affairs.'

So, consider how the investigator, after introspection, might express belief as follows:

Class characteristic	A = Negative binomial	$\overline{A}$ = Not negative binomial	
Personal probability	$\mathrm{bPr}(A) = 0.85$	$\mathrm{bPr}(\overline{A}) = 0.15$	(12.6.7)

Turning now to inferences about θ, suppose the investigator's prior is given by the $\mathrm{Beta}_1(50, 300)$ distribution on $0 < \theta < 1$, where the likelihood of θ given the data is of course given at (12.6.5). Then

the prior is proportional to $\theta^{50-1}(1-\theta)^{300-1}$,
the likelihood is proportional to $\theta^{54}(1-\theta)^{207}$, and
the posterior is proportional to $\theta^{54+50-1}(1-\theta)^{207+300-1}$, (12.6.8)

where the posterior is a $\mathrm{Beta}_1(54+50, 207+300)$ distribution on $0 < \theta < 1$. A personal probability 'elimination pivot' for θ is thus given by

$$\left(\frac{1-\theta}{\theta}\right)\left(\frac{54+50}{207+300}\right) = \text{Snedecor's } F \text{ on } 2(54+50) \text{ and } 2(207+300) \text{ df.} \qquad (12.6.9)$$

For instance, the upper 10% point of F here equals 1.143, and by inserting this value at (12.6.9) and solving for θ, we find $\theta = 0.152$ as the lower bound of a 90% one-sided credibility interval for θ. Continuing like this, we obtain the following:

(90% one-sided lower bound, 90% one-sided upper bound)
$= (\theta = 0.152, \theta = 0.191)$.
(95% one-sided lower bound, 95% one-sided upper bound)
$= (\theta = 0.147, \theta = 0.198)$.
(99% one-sided lower bound, 99% one-sided upper bound)
$= (\theta = 0.139, \theta = 0.210)$.

Maximising the quantity at (12.6.8) we obtain a point estimate, $\hat{\hat{\theta}} = 0.170$, which is straddled by each pair of bounds. (12.6.10)

Bayesian inference is clearly an attempt at implementing the notion of 'probability inference', and so cannot avoid the notion of 'simultaneous statistical inference'. For instance, if any pair of bounds obtained at (12.6.10) is considered *simultaneously*, it becomes a pair of *two-sided* credibility bounds with altered coverage probability. So the credibility bounds at (12.6.10) can also be interpreted as follows:

The 'unbiased' 80% two-sided bounds for θ are given by $(\theta = 0.152, \theta = 0.191)$.
The 'unbiased' 90% two-sided bounds for θ are given by $(\theta = 0.147, \theta = 0.198)$.
The 'unbiased' 98% two-sided bounds for θ are given by $(\theta = 0.139, \theta = 0.210)$.

Maximising the quantity at (12.6.8) we obtain a point estimate, $\hat{\hat{\theta}} = 0.170$, which is straddled by each pair of bounds. (12.6.11)

The developments at (12.6.6), (12.6.10) and (12.6.11) lead to revealing questions. To begin with, let us ask of each of these developments in turn:

> Are these results correct? (12.6.12)

At (12.6.6) we can answer 'Yes of course. We can, if needs be, simulate them for you'. At (12.6.10) and at (12.6.11) such an answer is not possible, because those results are physically meaningless; they cannot be put to the human body; they are presumably to be put to the human psyche. But how is that done? Consider the bound $\theta = 0.152$, as it appears at (12.6.10) and (12.6.11); is it a case of 'one-sided knowing' versus 'two-sided knowing'? It cannot be a case of different interest, because that belongs to the use of knowledge, not to the pursuit of knowledge; interest might be in an assurance against too small a value only, but one cannot call that one-sided knowing, as the idea that an investigator gains advantage by turning a blind eye to certain possibilities is arrant nonsense. Again, the question at (12.6.12) presupposes 'a correct prior'. But how can that be? And again, consider the question: 'Is the (0.050, ε, 0.95) bound at (12.6.6) comparable to the 95% one-sided lower bound given at (12.6.10), or to the 90% two-sided lower bound given at (12.6.11)?' It cannot be answered; physics cannot be compared to metaphysics. The issue here is simply this: in the discourse of physics, the question at (12.6.12) is a meaningful question. In the discourse of metaphysics it is evidently meaningless.

There is more to come.

The word 'coherent' has two quite different meanings; it might mean 'sticking together' or 'making sense'. Bayesians hold that their inferences are 'coherent'. If that refers to using Bayes's rule in order to 'stick together' belief probabilities and likelihoods, then that is so in a trivial sense. If, however, one is supposed to understand that the posterior outcomes of the process of 'sticking together' makes sense, that is fallacious, because if that were to be the case, the question at (12.6.12) would be easily answerable. In case of the results at (12.6.6) it is indeed easily answerable. But as we have seen above, in the case of the results at (12.6.10) and (12.6.11) the question leads to an epistemological mare's nest. So let us note carefully that the term 'coherent' in the epistemologically important sense means 'to communicate', 'to make sense', 'to convey meaning' as in

> 'Despite his traumatic experience little Oliver was able to give a perfectly coherent account of what had happened',

as opposed to

> 'The poor child was quite incoherent; nobody could make out what had happened.'

We assert, using the term 'coherent' in this important sense, that Bayesian inference is incoherent. In fact, it is incoherent in two distinctly different ways, which we might identify as *endogenous incoherence* and *exogenous incoherence*, respectively.

The endogenous incoherence of Bayesian inference: Consider the 'unbiased' 80% two-sided bounds for θ given by ($\theta = 0.152$, $\theta = 0.191$) at (12.6.11). The unconditional personal probability that these bounds bracket 'the true value of θ', with reference to the personal probabilities expressed at (12.6.7), is given by

$$\begin{aligned}&\text{bPr}(A)\text{bPr}(0.152 \le \theta \le 0.191|A)+\text{bPr}(\bar{A})\text{bPr}(0.152 \le \theta \le 0.191|\bar{A})\\&= (0.85)(0.80)+(0.15)\text{bPr}(0.152 \le \theta \le 0.191|\bar{A}),\end{aligned} \qquad (12.6.13)$$

where the second factor of the second term depends on an unknown possibility, $\bar{A}$. The point here is simply this: Bayesian inference cannot be coherently adjoined to its own commencement. The incoherence is endogenous, as it arises within the frame of statistical reference. We note that co-ordination tests do not fall victim to any such difficulty, because commencement tests of co-ordination and elimination tests of co-ordination do not have to be 'stuck together' in order for them to 'make sense'. After all, this is true of any investigative reasoning of science. When for instance we point at a fossil saying 'This seems to be the fossilised skull of a three-toed horse, and if so it would be of more ancient origin than has so far been thought possible', we say 'if so' indicating that the two observations are not to be 'stuck together'; should the datum of species fall away, the datum of radiometric dating could still be standing.

The exogenous incoherence of Bayesian inference

Any reasoning that must resort to an input by Bayesian inference will necessarily be incoherent to science, because the arcane nature of the 'personal probabilities' that are required for such inference, are then made into a characteristic of that reasoning. No matter how remotely the reasoning might progress beyond that input, the questions

'What does the reasoning contribute to the discourse of physical experience?'

or (equivalently)

'How can an understanding of that contribution be forced upon the human body?'

or (equivalently, and quite simply)

'How is it capable of simulation?'

cannot be escaped. Therefore, such reasoning leads back inexorably to the metaphorical albatross of being unable to reply to the question: 'What on earth does that mean?' Some Bayesians have tried to suppress the influence of 'personal belief' by resorting to diffuse priors. But, as we saw at (12.2.9) the albatross remains. In short, some 200 years of debate and development has failed to remove the exogenous incoherence of Bayesian inference, i.e. failed to make such inference understandable to the discourse of substantive science. This must be firmly grasped, because there is a great deal of literature in which all kinds of clever mathematical development might make us forget that if it started from Bayesian inference, it retains the incoherence of that source. So, to be brutally frank, what derives from the arcane is arcane, and what derives from the arcane that derives from the arcane is arcane, and so forth.

12.7 A VICIOUS CIRCLE

In this section we develop proof that the use of Bayesian inference for informative data analysis, that is to say, for the discourse of the pursuit of knowledge, inescapably leads to circular reasoning. We do so by means of examples.

Example 12.7.1

Recall the development at (12.6.13) where we found that for a purportedly 'unbiased' 80% two-sided credibility interval for θ to be given by

$$(\theta = 0.152,\ \theta = 0.191),$$

the unconditional personal coverage probability should in fact be calculated as

$$\mathrm{bPr}(A)\mathrm{bPr}(0.152 \le \theta \le 0.191|A)+\mathrm{bPr}(\overline{A})\mathrm{bPr}(0.152 \le \theta \le 0.191|\overline{A})$$

$$= (0.85)(0.80)+(0.15)\mathrm{bPr}(0.152 \le \theta \le 0.191|\overline{A}), \text{ which is not calculable.}$$

Clearly, the only way in which $0.152 \le \theta \le 0.191$ can be defended as 'an "unbiased" 80% two-sided credibility interval for θ', is by setting

$$\mathrm{bPr}(\overline{A}) = 0,$$

so that

$$\mathrm{bPr}(A) = 1,$$

so that the coverage probability can be calculated as

$$(1)\mathrm{bPr}(0.152 \le \theta \le 0.191|A) + (0)\mathrm{bPr}(0.152 \le\ \le 0.191|\overline{A}),$$

*so that t*he answer would be

$$(1)(0.80) + 0(\bullet) = 0.80 \text{ as required,}$$

and *that* is circular reasoning.

Example 12.7.2

Bliss (1967) reproduces from Campbell (1926) the survival times, in log minutes -2.3, of 140 individual fourth-instar silkworm larvae following a dose of 0.10 mg arsenic per gram of body mass. The variance of the survival times, 0.0022884, can for all practical purposes be treated as the population variance. Bliss reports (p. 105) that the data are '... clearly lognormal without major deviations ... in line with a wide range of tests of reaction time to insecticides'. But, in order to simulate a number of investigations, we pretend in the following that the value of the population variance is all that we can learn from Campbell's historical record. From Bliss's Table A1 reading downward in columns of three and using the first 18 positive numbers less than 141 (thus counting from 001 to 140) we draw 18 survival times, as follows in pseudo-random order:

0.174	0.108	0.127	0.038	0.098	0.167	} Data Set 1
0.127	0.133	0.091	0.123	0.082	0.100	
0.115	0.139	0.263	0.156	0.067	0.158	

Adding consecutive pairs of survival times, we reduce the data as follows:

0.282	0.165	0.265	} Data Set 2
0.260	0.214	0.182	
0.254	0.419	0.225	

Adding consecutive triplets in Data Set 1, we reduce the data as follows:

0.409	0.303	} Data Set 3
0.351	0.305	
0.517	0.381	

Co-ordination tests applied to Data Sets 1, 2 and 3, in turn

Using Shapiro-Wilk's W to test, in turn, Data Sets 1, 2 and 3 for non-normality, we obtain respectively:

W = 0.939 with mental correlate at (*0.34, ε, 0.66) in the test distribution.
W = 0.874 with mental correlate at (*0.18, ε, 0.82) in the test distribution.
W = 0.898 with mental correlate at (*0.38, ε, 0.62) in the test distribution. (12.7.1)

Next we note that for the class of normal models a minimal sufficient statistic for the population mean may be taken to be the sample mean, whose distribution is that of a normal random variable. For Data Set J that distribution is

$N[J\mu, J\sigma^2 \div (18 \div J)]$ where $\sigma^2 = 0.0022884$, J = 1, 2, 3, respectively. (12.7.2)

Co-ordinate bounds for Jμ at (0.05, ε, 0.95) and (0.95, ε, 0.05), are thus for J = 1, 2, 3,

$$\mu = \frac{2.266}{18} \pm 1.645 \sqrt{\frac{0.0022884}{18}}$$, i.e. 0.126±0.00185 in -2.3 code,

$$2\mu = \frac{2.266}{9} \pm 1.645 \sqrt{\frac{2(0.0022884)}{9}}$$, i.e. 2(0.126±0.00185 in -2.3 code), and

$$3\mu = \frac{2.266}{6} \pm 1.645 \sqrt{\frac{3(0.0022884)}{6}}$$, i.e. 3(0.126±0.00185 in -2.3 code),

respectively. We remark on three points:

(1) The three pairs of co-ordinate bounds are one to one transforms of each other because the minimal sufficient statistics and the data it addresses are one to one transforms of each other. In short, we have derived the same bounds in superficially different ways.

(2) The three tests at (12.7.1) differ from each other, because the data differ and the class characteristics differ correspondingly by way of different sample sizes.

(3) In all three cases the normal class characteristic, as tested, fits the data very well. However, the co-ordinate bounds are not based on an *assumption* of normality. This is so because co-ordinate tests measure quality of fit only. The reader must recall that if we should point at a spoor in the veld saying, 'This does not resemble the spoor of an aardvark; but *if it were* the spoor of an aardvark, it would seem to be that of a juvenile', then *we have not assumed* that the spoor is that of an aardvark. And if we point at a spoor in the veld saying, 'This resembles the spoor of an aardvark; and *if it were* the spoor of an aardvark, it would seem to be that of a juvenile' then, too, *we have not assumed* that the spoor is that of an aardvark.

Bayesian inference applied to Data Sets 1, 2 *and* 3, *in turn*

The minimal sufficient statistic for μ is identical in the three cases, and is given by

$$\overline{X} = 0.1259, \text{ whose distribution is N } [\mu, (\sigma \div \sqrt{18})^2] \text{ with } \sigma^2 = 0.0022884.$$

Let the prior distribution of μ be N(ν, τ^2) with $\nu = 0.3$, and $\tau^2 = 0.0025000$. Then the posterior distribution of μ, according to the forms at (11.3.3) is N(λ, ϕ^2), where

ϕ^2 is given by $\dfrac{1}{\left(\dfrac{1}{\sigma^2 \div 18} + \dfrac{1}{\tau^2}\right)} = 0.00012$, so that $\phi = 0.0110$, and

λ is given by $\dfrac{\left(\dfrac{0.1259}{\sigma^2 \div 18} + \dfrac{0.3}{\tau^2}\right)}{\left(\dfrac{1}{\sigma^2 \div 18} + \dfrac{1}{\tau^2}\right)} = 0.1222.$

So the 'unbiased' 90% two-sided 'credibility interval' for μ is given by

$$0.1222\text{-}1.645(0.0110) \le \mu \le 0.1222\text{+}1.645(0.0110).$$

Hence the 'unbiased' 90% two-sided 'credibility interval' for μ is given by

$0.104 \le \mu \le 0.140$ for Data Set 1,
$0.104 \le \mu \le 0.140$ for Data Set 2, and
$0.104 \le \mu \le 0.140$ for Data Set 3, (12.6.3)

each of these being conditional on the population being normal, and we cannot close our minds to the possibility that the population might be non-normal. So let us consider how by introspection an investigator expresses belief, perhaps as follows:

Class Characteristic	A = Normal	$\overline{A}$ = Non-normal	
Personal Probability on Data Set 1	bPr(A) = 0.85	bPr($\overline{A}$) = 0.15	
Personal Probability on Data Set 2	bPr(A) = 0.75	bPr($\overline{A}$) = 0.25	
Personal Probability on Data Set 3	bPr(A) = 0.65	bPr($\overline{A}$) = 0.35	(12.6.4)

Here bPr(A) for Data Set 1 > bPr(A) for Data Set 2 > bPr(A) for Data Set 3, might arguably be seen as the natural response to the tests at (12.7.1), as three equal values for bPr(A) in the three cases would seem laughable. Be that as it may, we note that the unconditional coverage probability is in each of the three cases given by

$$\text{bPr}(A)\text{bP}(0.104 \le \mu \le 0.140 \mid A)+\text{bPr}(\overline{A})\text{bP}(0.104 \le \mu \le 0.140 \mid \overline{A}),$$

which is unknown. So, the only way in which $0.104 \le \mu \le 0.140$ can be defended as 'an "unbiased" 90% two-sided credibility interval for μ', is by setting

$\text{bPr}(\overline{A}) = 0,$
so that
$\text{bPr}(A) = 1,$
so that the coverage probability can be calculated as
$(1)\text{bPr}(0.104 \le \mu \le 0.140 \mid A)+(0)\text{bPr}(0.104 \le \mu \le 0.140 \mid \overline{A}),$
so that the answer would be
$(1)(0.90)+0(\bullet) = 0.90$ as required,
and *that* is circular reasoning.

Concluding remark

The *idée fixe* invariably leads to a vicious circle because its commencement requires us to know for sure something that cannot possibly be known for sure. So, Bayesian inference is inexorably caught on the horns of an inescapable dilemma. If it tries (this is the first horn of the dilemma) to provide for a commencement test, it turns out to be impossible to achieve a coherent transition from that commencement to the elimination tests at which Bayesian inference is directed. So it is forced (this is the second horn of the dilemma) to proceed from a purportedly reasonable assumption about the class characteristic, whereby its reasoning then becomes circular. It is well worth noting that the proof we developed in the present section is in principle the same as the proof we developed in Section 4.8 in respect of the *bête noire* of Bayesian inference, namely frequentist inference. More than any others amongst the various silly theories of statistical inference, these two theories have cluttered the statistical literature with endless arguments and counter-arguments in which both of the two theories are so busy scoring points off each other that they are blind to the fatal circularity they have in common.

12.8 On a method of commencement testing

Certain remarks made by Daniel and Wood (1971, p. 29) might prompt the following approach to commencement testing of the quality of fit of a given class characteristic. Consider for instance whether or not a data set comprising say n = 16 measurements can satisfactorily be represented as a random sample from an N(0, 1^2) population. For such samples, the standardised residuals may be taken to be the corresponding samples from the normal class characteristic. So, simulate say 100 independent samples of such residuals, and plot for each simulated sample, the quantiles of its empirical cumulative distribution (e.c.d.) against the quantiles of the N(0, 1^2) distribution for the percentage points

(j-0.5)÷n, where j = 1, 2, 3, ..., n. The plotted points will tend to scatter round a straight line (examples are given by Daniel and Wood on pp. 34-43). Similarly, plot the data e.c.d. for comparison to the hundred simulated e.c.d.-s Does the data e.c.d. blend snugly into that crowd? Perhaps not, the data e.c.d perhaps being in a particular way atypical, such as involving an oversized residual, or being curved rather than straight, or seeming to arise from a discontinuity. In such a case, sort the simulated samples into three groups with Group U, on the left, comprising those simulated e.c.d.-s that are more atypical in the particular way that the data e.c.d. is atypical; Group V, on the right, comprising those simulated e.c.d.-s that are less atypical in that way; and Group ε, in the middle, comprising the remaining e.c.d.-s. Count the e.c.d.-s in each group. Typical relative counts might be

$$(U, \varepsilon, V) = (0.07, 0.02, 0.91), \text{ or } (0.11, 0.06, 0.83), \text{ or } (0.74, 0.12, 0.14), \text{ or } \ldots ,$$

where a low relative count on the left points at a poor quality of fit. This procedure can be facilitated by computing for each e.c.d. the value of a measure of the particular atypicality of the data e.c.d. The measure might be the Shapiro-Wilk measure, or Pearson's measure of skewness, or Pearson's measure of kurtosis, or whatever measure might be retrospectively devised to describe the atypicality of the data e.c.d. It will in fact be found that certain computer facilities in effect do just that, thereby producing printouts that give the results for a variety of such measures, usually along with a plot of the data e.c.d. What do advocates of Bayesian inference want a data analyst to do with such a printout? Is such a printout a snare and a delusion? Is it best ignored? Is it best, with eyes averted to be on the safe side, quickly torn off and thrown away? An answer in the affirmative would be absurd. And an answer in the negative must prompt us to ask how Bayesian inference then avoids incoherence or (even worse) circularity.

12.9 There cannot be 'a parametric commencement'

Tukey's test for non-additivity can be interpreted as a test for the value of a parameter that represents a multiplicative effect. It might therefore seem that commencement testing can be avoided by introducing a suitable parameter (or parameters) and then proceeding to elimination testing. However, that is not the case. Examples given by Tukey (1949a) make it clear that he tests for a multiplicative *term*, rather than a *parameter*. So the test detects non-additivity of different kinds, such as non-additivity removable by arcsine transformation, or by power transformations of various kind, or yes also by logarithmic transformation, or by the removal of outliers. Again, if γ denotes the parameter for which Tukey ostensibly tests, the test is a commencement test because we are not interested in what variety of γ-values are tenable on the given data; we are interested in whether just one particular γ-value (usually zero) is tenable, where that value is often suggested by circumstantial evidence above and beyond the data to be statistically analysed. Again, it is arguably the case that non-normality of a data set of say ten observations can often be dealt with by means of a transformation such as

$$X \rightarrow (\alpha+X)^{\beta}, \text{ where } \alpha \text{ and } \beta \text{ are parameters additional to the usual } \mu \text{ and } \sigma^2.$$

That simply means that instead of having 10-2 degrees of freedom amongst the residuals to test for non-normality, we have 10-4 degrees of freedom amongst the residuals, and we

still have to test for non-normality. If we introduce a further six parameters, there will be no residual slack, and we will be unable to test for anything. Similar remarks apply to the generalisation of Tukey's test developed by Milliken and Graybill (1970, 1971). The crux of the matter is that there must be slack if we are to test the adequacy of the class characteristic.

12.10 An unrealistic outlook

There is a community of ivory-tower individuals who have confidently declared that 'the 21st century will be the Bayesian century' and who therefore seem to believe it possible for current practices of investigative statistics to be advantageously replaced by Bayesian inference. It would seem they have either failed to notice that their ideas find no place at all in the handbooks by Snedecor and Cochran (1989), Mead and Curnow (1983), Steel and Torrie (1980), Bliss (1967), and so forth, or else they must believe these handbooks to be badly out of date. So, let us consider what a successful palace revolution under their leadership would inflict upon our customers in the substantive sciences. Consider how the examiners of a PhD candidate in for argument's sake agronomy would then have to be persuaded to approve his/her thesis. It would, if properly organised, comprise two parts that the candidate would then have to defend as follows:

Defence of Part 1: In this part of my thesis I had to use certain commencement tests to defend the class characteristics that define various models that I need to represent my data. I realise that these tests rely on an outdated form of reasoning, but my statistical advisors tell me that inasmuch as Bayesian improvements on such reasoning have yet to be developed, the outdated tests are currently the best that can be done. So ...

Defence of Part 2: Thanks to my statistical advisors, I have, for this part of my thesis, been able to use the modern methods of Bayesian inference that have recently been introduced to replace most of the old-fashioned Snedecor-style statistical methods. I realise that it is awkward, and in fact incoherent, to have modern methods rest on findings that in Part 1 were arrived at by reasoning that, by the modern methods, are held to be outdated if not downright fallacious, but on that point I must plead for your indulgence. After all, if I cannot assume that the findings obtained by the Snedecor-style methods I am forced to rely on in Part 1 are correct, I would be quite unable to proceed. So ...

It might be countered that such incoherence already exists, as we have seen in Chapter 4. We must, however, note that the current incoherence arises within the framework of frequency physics, owing to which it is demonstrable that statisticians and substantive investigators together somehow manage to avoid its worst consequences. As opposed to that, the palace revolutionaries would have that incoherence greatly aggravated, as one would then have a 'Part 1' in frequency physics followed by a 'Part 2' in belief metaphysics. In any case, the theory of co-ordination tests shows how to eliminate the current incoherence.

12.11 'Bayesian inference' and the problem of ancillary partitions

It is widely (and wrongly) believed that the problem of finding the correct ancillary partition does not arise in Bayesian inference. Such belief is wrong, as the ancillary-partition conundrum in respect of a solitary data set to be analysed can be resolved by, and only by, an epistemology whose findings are expressed in terms of populations of samples that are capable of showing, by simulation, how those data may, or may not, have come about. Frequentist inference cannot resolve the conundrum because, although it understands that a physical resolution is required, it fails to grasp that the resolution must necessarily address just one particular real-world data set; instead it opportunistically tries to adjoin its rules for physical discourse to whatever the correct resolution might be. Bayesian inference also cannot resolve the conundrum because, although it understands that a particular data set is to be addressed, it fails to grasp that the resolution needs to do so in physical terms; instead it opportunistically tries to adjoin its rules for metaphysical discourse to whatever the correct resolution might be. In fact, no theory of statistical inference could resolve the conundrum because the conundrum does not arise in elimination testing; it *precedes* elimination testing. Consider for instance the examples in Sections 9.4 and 9.5, where the conundrum for frequentist inference takes the form: 'Which one of these different arrays of statistical models indexed by θ is the correct array?', and the conundrum for Bayesian inference takes the form: 'Which one of these different arrays of likelihoods indexed by θ is the correct array?' The conundrum in fact also precedes commencement testing, because its general form is: 'Which one of these different *classes* of statistical models is the correct class?' and, as pointed out in Section 9.5, its resolution can be achieved by, and only by, involving substantive science.

12.12 Subjective versus objective – a silly dispute

An infernal nuisance in the statistical literature arises from frequentist claims that it employs probability in the 'objective' sense of long-run frequency only, as opposed to the 'subjective' probabilities employed by the Bayesians. A counter-argument by the Bayesians claims that frequentist inference and Bayesian inference both rely on assumptions that cannot 'objectively' be proven to be true, and so both these forms of inference involve 'subjective' constituents anyway. Because of this, the statistical literature pervasively draws a distinction between objective and subjective matters. But that is not the important distinction to be drawn between, for instance, the trace of a mental correlate and a Bayesian posterior of belief probabilities. The important distinction is that the trace expresses *physical* meanings; no one can say of the trace 'I cannot grasp its meanings', because one can be invited into the statistical laboratory where those meanings can by simulation be forced onto the human body. Opposed to that, the Bayesian posterior tries to convey *metaphysical* meanings; that is to say, it tries to convey meanings that cannot be forced onto the human body. So, Bayesians have little use for the statistical laboratory; some even hold that the randomisation of an experimental design is unnecessary. The issue is this: often, even usually, usage of the terms 'objective' and 'subjective' distracts the attention from the distinction between physics and metaphysics in cases where the latter distinction is the important one. We can therefore understand why Oscar Kempthorne often called for the terms 'subjective' and 'objective' to be banned from the statistical literature.

Chapter 13

THE MULTIPLE COMPARISON MUDDLE

A profession in denial

13.1 Introduction

In previous developments we underscored the analytic nature of scientific investigation. Consider for instance what to expect in a book on geology. A naïve understanding might expect chapters describing the Cango Caves, Baviaans Kloof, Table Mountain, and so on. Instead one finds chapters on wind erosion, sedimentation, metamorphic processes, and so on – chapters on the *analytic* models of geology, where the Cango Caves, Baviaans Kloof and Table Mountain serve merely as examples. Much, if not most, of the literature on statistical inference fails to grasp this. Instead of developing the analytic questions that substantive science might ask, and developing statistics that are *minimally sufficient* for *separately* addressing each question in turn, our literature on inference falls victim to irrelevant notions about the possible errors of a knowing subject. This inexorably leads to statistical views in which the activities of investigators in substantive science are envisaged as 'data snooping' (Scheffé 1959, p. 80). According to such views statisticians must reconcile themselves with the fact that substantive investigators simply cannot be prevented from recklessly 'looking at the data' (Cox and Hinkley 1974, p. 241) and so (horror of horrors) arriving at effects 'suggested by the data', that is to say, arriving at 'unplanned comparisons' (Steel and Torrie 1980, p. 174). This, such views would have us hold, cannot be dealt with in the same way that we would deal with a limited number of 'planned comparisons' formulated in advance (Snedecor and Cochran 1989, p. 226). So, willy-nilly, our profession has set itself the task of 'protecting' science from the outcomes of such 'unplanned' investigations, by providing recipes for such investigations to be subject to 'controlled error rates'. The outcome is predictably incoherent because comparisons that are unplanned are also unforeseeable. So we are supposed to provide for forecasting error rates of unforeseeable conclusions on the part of the knowing subject. But as the extent of the unforeseeable is infinite, it turns out to be impossible to ensure that such attempts at foreseeing the unforeseeable make scientific sense. In fact, we will find that all multiple comparison procedures have logical implications that do not make sense. The procedures are of two different types named *simultaneous-confidence-region procedures* and *multiple range procedures*, respectively. Their beginnings are essentially to be found in Student's two-tailed *t* test as we have previously explained in Section 1.27. We will recall that explanation after briefly introducing the following definitions and notations that will also be required for examples to be used in subsequent sections:

> *m* treatments A, B, C, ... , are replicated *n* times in a completely randomised design.
> The treatment means are denoted by, $\overline{x}_A$, $\overline{x}_B$, $\overline{x}_C$, ..., for A, B, C, ..., respectively.
> $\overline{X}_A$, $\overline{X}_B$, $\overline{X}_C$, ..., denote independent homoscedastic normal random variables.

$\bar{x}_A$, $\bar{x}_B$, $\bar{x}_C$, ..., also denote realised values of $\bar{X}_A$, $\bar{X}_B$, $\bar{X}_C$, ..., respectively.

μ_A, μ_B, μ_C, ... = E($\bar{X}_A$, $\bar{X}_B$, $\bar{X}_C$, ...), denote corresponding population means.

σ^2 and s^2 denote the error variance and the pooled error estimate, respectively.

s^2 is a realisation of S^2 whose distribution is $\sigma^2[\chi^2$ on m(n-1) df$]\div$m(n-1). (13.1.1)

If interest is expressed in μ_A-μ_B the investigator might consider Student's *t*, given by

$$\frac{\bar{X}_A\text{-}\bar{X}_B}{s \times \sqrt{2\div n}} = t\text{(A versus B) say.} \qquad (13.1.2)$$

If this turns out to be unusually large in the positive direction, the investigator might well consider that to be evidence that μ_A-$\mu_B > 0$. Cox and Hinkley (1974) would then have us envisage other cases in which such a value is unusually large in the negative direction, and therefore considered to be evidence that μ_A-$\mu_B < 0$. The possibility persuades Cox and Hinkley to make 'a correction for selection' (p. 106) which amounts to

replacing ($\bar{X}_A$-$\bar{X}_B$, S^2) with (| $\bar{X}_A$-$\bar{X}_B$ |, S^2). (13.1.3)

There is more to come. If interest is also expressed in μ_C-μ_D the investigator would consider Student's *t* as given by *t*(C versus D) in the notation defined at (13.1.2), where 'correction for selection' would then amount to

replacing ($\bar{X}_C$-$\bar{X}_D$, S^2) with (| $\bar{X}_C$-$\bar{X}_D$ |, S^2). (13.1.4)

A further correction for selection is then required to account for

(| $\bar{X}_A$-$\bar{X}_B$ |, S^2) and (| $\bar{X}_C$-$\bar{X}_D$ |, S^2) *simultaneously*. (13.1.5)

Cox and Hinkley sometimes refer to such reasoning as making 'allowance for selection' (p. 123) or by saying in effect that the 'knowing subject's error rate' must be 'adjusted for selection' (p. 124). The reasoning is very persuasive, as follows:

If the error rate of findings like μ_A-$\mu_B > 0$ would be α,
then the error rate of findings like μ_A-$\mu_B < 0$ would also be α.
So, the error rate of findings like μ_A-$\mu_B > 0$ or μ_A-$\mu_B < 0$ would be 2α,
and the error rate of findings like μ_C-$\mu_D > 0$ or μ_C-$\mu_D < 0$ would also be 2α.
So, the error rate of findings like
μ_A-$\mu_B > 0$ or μ_A-$\mu_B < 0$, and μ_C-$\mu_D > 0$ or μ_C-$\mu_D < 0$,
would be $2\alpha+2\alpha-(2\alpha)(2\alpha) = 1-(1-2\alpha)^2$.

But look at what this amounts to! At (13.1.3) we replace the minimally sufficient statistic for μ_A-μ_B with a statistic that is insufficient for μ_A-μ_B. At (13.1.4) we replace the minimally sufficient statistic for μ_C-μ_D with one that is insufficient for μ_C-μ_D. At (13.1.5) we then confound the statistic that is insufficient for μ_A-μ_B with a statistic that is vacuous for μ_A-μ_B and, by the same token, confound the statistic that is insufficient for μ_C-μ_D with a statistic that is vacuous for μ_C-μ_D. Here Definition 13.1.2 is appropriate.

Definition 13.1.2:

A statistic is *vacuous* for a parameter if and only if the distribution of that statistic is independent of that parameter.

For instance, the crux of the randomised test paradox discussed in Section 4.22 is that the auxiliary random number is vacuous for the parameter of interest. The reader should note, however, that the power of an array of hypothesis tests is maximised for a specified Type I error rate by adjoining the auxiliary random number, whereas a comparable gain is not at all achieved in the case of simultaneous statistical inference. On the contrary, whenever the ideas of such simultaneity are introduced, operating characteristics in the case of an array of hypothesis tests, or separating characteristics in the case of a suite of co-ordination tests, are adversely affected. The adversity can be horrendous. Referring to the denominator of t as defined at (13.1.2) as a standard error unit, we find for instance from the prepared tables of Harter (1960) for *Duncan's new multiple range test* that for a specified Type I error rate $\alpha = 0.01$, the required difference for significance between any two of ten treatment means arising as at (13.1.1) is three standard error units – and this might be considered too small, as Duncan's procedure has been criticised by amongst others Scheffé (1959, p.78) as incomprehensible.

13.2 Student's knowing subject

Suppose that any one of five alternative protein sources can be used to make up a feeding ration for dairy cows, and that in order to compare the five sources for their efficacy, we obtain the data in Table 13.2.1. A multiple range procedure originating in Student (1927)

Table 13.2.1: Imaginary data purportedly from a completely randomised design

Treatment (protein source)	A	B	C	D	E
Number of cows (replications)	5	5	5	5	5
Milk production per cow	21.1	21.5	22.0	20.2	27.1
Cost per cow in terms of milk	11.5	12.0	16.0	18.0	26.0
Treatment mean (profit)	9.6	9.5	6.0	2.2	1.1
Estimated standard error of a treatment mean: 1.2 on 20 degrees of freedom					

would then have us proceed as follows. We must choose the Type I error rate without any reference to the data. Let us choose 5%. We then require certain 'raising factors' that are obtainable from tables prepared by May (1952). In our case (five treatments, 20 df for error, and a 5% risk) the factors are 2.9, 3.6, 4.0 and 4.2, respectively. We must raise the value of the estimated standard error of a treatment mean by each of these factors in turn. (The factors already incorporate the square root of two, thus effectively raising the standard error of the *difference* between two treatment means.) Corresponding

values of multiple range statistics appropriate to our case are thus realised, as follows (in standard error units):

$R_2 = 1.2 \times 2.9$ = 3.5 units. $R_3 = 1.2 \times 3.6$ = 4.3 units. $R_4 = 1.2 \times 4.0$ = 4.8 units. $R_5 = 1.2 \times 4.2$ = 5.0 units.

The following display of ordered treatment means is then obtained by underlining every array of j consecutive means that vary by less than R_j, for j = 5, 4, 3 and 2, in that order, subject to the condition that, once an array of means is underlined, no sub-array found in that array is to be underlined (Miller 1981, p. 82):

A	B	C	D	E
9.6	9.5	6.0	2.2	1.1

Here for instance, ABC is underlined because $9.6 - 6.0 < R_3$, and DE is underlined because $2.2 - 1.1 < R_2$. The multiple range rules for inferential discourse then want us to reason that any differences between pairs of means that have been underlined are attributable to error, whereas (we are supposed to reason) any differences between pairs of means that have not been underlined are attributable, subject to negligible risk (5% in our case), to corresponding treatment differences. So in the present case we are *inter alia* supposed to reason that the apparent superiority of A over C might well be attributable to error, and so cannot be attributed to the superiority of A. But suppose the cost of B now increases from the previous equivalent of 12.0 units of milk per cow, to a new equivalent of 20.3 units of milk per cow. Then the previous display must be replaced by a new display, as follows:

A	C	D	B	E
9.6	6.0	2.2	1.2	1.1

The multiple range rules for inferential discourse would now have us reason that the difference between the measured mean performances of A and C is no longer attributable to error – because of the increased cost of B!

It would be futile to try to wriggle out of such implications by arguing that we were not supposed to alter the data. We have in fact not altered the data; the full data set simply tells us that a given protein source might be of a different cost, depending on the supplier. It is an inescapable principle of science that its reasoning must be able to survive any test of logical implication. So, if the protein sources labelled A, B and C, were soy bean meal, fish meal and lupine meal, respectively, the 'inferences' we are supposed to make then commit us, by their logical implication, to botanical reasoning according to which the efficacies of the two plant protein sources have come to differ as a result of declining numbers of pelagic fish in South African waters, owing to over-fishing by commercial trawlers, resulting in an increase in the price of fish meal. This is a perfect example of how the knowing subject of statistical inference looses track of substantive science.

It would also be futile to try to wriggle out of the foregoing by arguing that the multiple-range mathematics is correct. Of course it is correct! After all, the procedure used in the foregoing traces back to Student (1927), and its mathematics have been checked and re-

checked by countless other individuals. We must understand that the procedure is *mistaken*, which is not the same as being *wrong*. This distinction must be firmly grasped. For instance, if Koos brings a ladder to a party because the invitation said that drinks will be on the house, he has mistaken a figurative meaning for a literal meaning. Nevertheless, his reasoning is not wrong. On the contrary, according to his understanding, his reasoning is perfectly sound. The issue can hardly be overemphasised, as the whole sorry confusion of efforts to provide for statistical inference springs from a mathematical impatience with the reasoning of our customers in the substantive sciences, so that, instead, *our* reasoning can 'get on with the mathematics', which mathematics, despite being correct, then time and again turns out to proceed from a mistaken perception of the investigative method of substantive science.

It will be found that all *multiple-range* procedures suffer from the defect exemplified by the foregoing example, because all those procedures advocate rules for discourse where the significance, or insignificance, of the difference between any two means, is dependent on whether or not those two means straddle other means. Hence, we may dispense with any further procedures of that type, and in the rest of this chapter consider procedures of the *simultaneous-confidence-interval* type.

13.3 Example of a simultaneous-confidence-interval procedure

Let the set-up at (13.1.1) be for a yield trial with oat cultivars A, B and C, the data being

$$(\bar{x}_A, \bar{x}_B, \bar{x}_C, s^2) = (19, 14, 10, 4.41) \text{ from say n = 9 replications.} \qquad (13.3.1)$$

For reasons that will soon appear, we suppose that the investigator tries to address an extremely odd problem, as follows: the investigator envisages a point on the wall of the agronomy building by postulating that for some choice of origin and graphical scale its Cartesian co-ordinates are given in terms of the oat yield parameters as

$$[(\mu_A-\mu_B), (\mu_A+\mu_B-2\mu_C)\div\sqrt{3}] = [E(p), E(q)] \text{ say,} \qquad (13.3.2)$$

an estimator of this point being given by

$$[(\bar{X}_A-\bar{X}_B), (\bar{X}_A+\bar{X}_B-2\bar{X}_C)\div\sqrt{3}] = (p, q).$$

The (p, q) notation is intended to underscore the formal resemblance of the development that follows to a previous development of the travelling-gene problem in Section 6.7. The estimated standard error of either one of the members of (p, q) is given by

$$\sqrt{2(s^2\div n)} = 0.990 \text{ on } (3\text{-}1)(9\text{-}1) = 24 \text{ degrees of freedom.}$$

Since $(1\text{-}2\alpha)^2 = 0.99$ when $2\alpha = 0.005$, and since the 0.005 critical values for Student's $|t|$ on 24 degrees of freedom are given by -3.09 and +3.09, it follows that a 99% square confidence region for E(p) and E(q) *simultaneously,* takes on the form of the Cartesian product of the two 99.5% confidence intervals for E(p) and E(q) *separately*, as follows:

$$[p\text{-}3.09(0.990) < E(p) < p+3.09(0.990)]\times[q\text{-}3.09(0.990) < E(q) < q+3.09(0.990)].$$

For the data given at (13.3.1)

p = 19-14, i.e. 5, and q = [19+14-2(10)]÷$\sqrt{3}$], i.e. 7.5,

and since 3.09(0.990) = 3.1, the 99% confidence square for [E(p), E(q)] is given by

[5-3.1 < E(p) < 5+3.1]×[7.5-3.1 < E(q) < 7.5+3.1], as in Figure 13.3.1(a). (13.3.3)

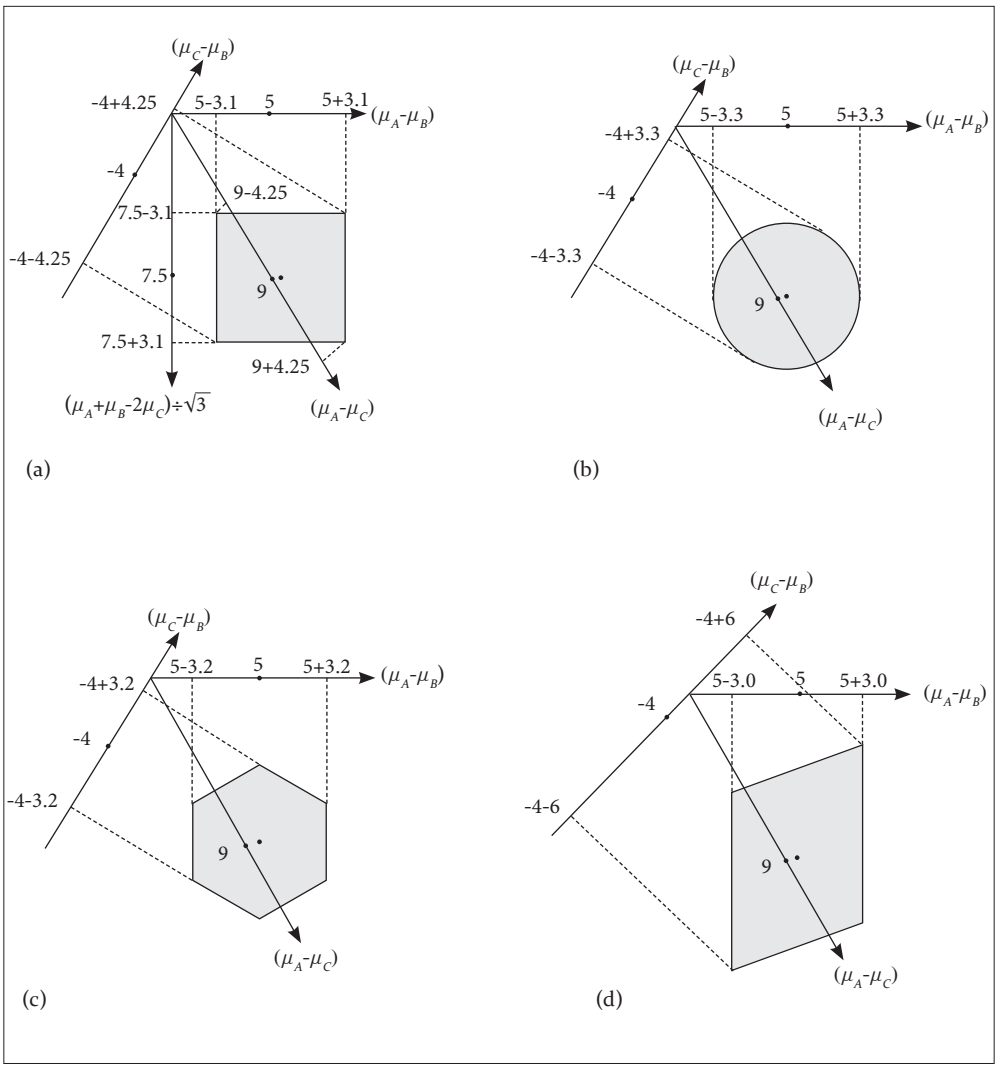


Figure 13.3.1: Four different 99% confidence regions for all possible parameter values of the form $\left[(\lambda_A\mu_A+\lambda_B\mu_B+\lambda_C\mu_C) \div \sqrt{\lambda_A^2+\lambda_B^2+\lambda_C^2}\right] \times \sqrt{2}$, simultaneously, where μ_A, μ_C, and μ_B are population means representing the yield totals of oat varieties *A*, *B* and *C*, respectively, and λ_A, λ_B, and λ_C are any real numbers that sum to zero

In the travelling-gene problem, the parameters E(p) and E(q) *jointly* conveyed a scientific meaning. In the present case, however, there is no such *joint* meaning – proof of this being as follows:

> In order to draw a version of Figure 13.3.1(a) onto the wall of the agronomy building we would, as already mentioned, require some choice of origin and of graphical scale. Must we for instance choose a graphical scale of 10 cm per kg of oats, or 20 cm per kg of oats, or 30 cm per kg of oats? And must the origin then be in the centre of the wall, or must it be lower down, at a suitable viewing height? These questions show that the *agronomical* point on the wall is completely ill defined, because whether or not the confidence square covers *any* given point on the wall will depend on our choice of the origin and scale. *Q.e.d.*

The issue is simply this: the confidence square is a well-defined behavioural concept, but the two parameters defined at (13.3.2) do not *jointly*, that is to say, *as a pair*, convey any corresponding agronomical concept. In other words, there does not exist an agronomical concept whose meaning is conveyed by, and only by, E(p) and E(q) *jointly*. So whatever was introduced to serve the idea of simultaneity in the present example was introduced to serve the idea of the knowing subject of statistics and not to serve any *agronomical* idea of simultaneity. The simultaneity envisaged for the knowing subject is therefore not anchored in agronomical science and must inexorably lead to the kind of silly gibberish we met in the previous section. Thus for instance if seed costs for variety A exceeds those for variety B by the equivalent of two additional units of yield, the 99 % confidence square at (13.3.3) would have us infer that:

> A is too costly compared to B, and the chance of this inference being wrong is a trifling 1%.

But, suppose we subsequently discover that the seed used as being for variety C, was in fact adulterated by seed of unknown origin, where we nevertheless observed that the oats labelled C was of the usual type, so that the error estimate remained appropriate. In that case the mean labelled C would be utterly worthless for the purposes of the trial and any analysis that retained that mean could not be defended as involving the correct error rate for the relevant agronomical conclusions. Thus forced to recalculate the 99% confidence region for E(p), we would obtain

> $[5\text{-}2.80(0.990) < E(p) < 5\text{+}2.80(0.990)]$, that is to say, $[5\text{-}2.8 < E(p) < 5\text{+}2.8]$,

and so be forced to infer that

> because the seed for C was adulterated, A is no longer too costly compared to B, and the chance of this inference being wrong is a trifling 1%.

Here the *simultaneous-confidence-region procedure* forces us to draw conclusions whose logical implications do not make substantive sense. It will be found that all simultaneous-confidence-region procedures suffer from this defect. Note also that despite differences in particular detail this is the same defect found in the previous section in respect of *multiple-range procedures*. In both kinds of procedure we are to draw conclusions that depend on irrelevant facts, and in both kinds of procedure that irrelevancy is dragged in

by rules for discourse whereby the statistic that is minimally sufficient for the parameter of interest is to be confounded with statistics that are vacuous for that parameter. Thus *both* of the two general types of multiple comparison procedure are defective in that logical implications of the conclusions they would have us draw amount to substantive gibberish. Also, in both kinds of procedure the defect is owing to beguiling probability calculus that has lost track of the investigative problem.

13.4 Vectors of experience

One of the sources of confusion leading to multiple comparison procedures is a failure to grasp that a vector of experiences (or conceptual experiences) is not a *further* experience (or *further* conceptual experience). Thus the point on the wall of the agronomy building is not an isolated example; on the contrary, it illustrates a general principle. Consider for instance the co-ordinates of the earth's magnetic pole. This might take the form of a pair of Cartesian co-ordinates, that is to say, a pair of conceptual experiences, in which case they might be thought to *simultaneously* comprise a *singular* experience, called for instance 'the position of the pole'. However, in order to grasp that such is not the case, we simply have to transform to polar co-ordinates, where the experiential pair, angle and distance, are clearly two distinct concepts. An *analyst* must always be exceedingly leery of dealing with an experiential vector, rather than with its separate components. (We note in passing that 'discriminant analysis' is a contradiction in terms arising from widespread misuse of the term 'analysis' in current statistical literature. 'Bayesian data analysis' is another such contradiction in terms.)

13.5 Two properties shared by all simultaneous-confidence-region procedures

The confidence square of the previous section provides confidence limits not only for

$$\mu_A-\mu_B \text{ and } (\mu_A+\mu_B-2\mu_C)\div\sqrt{3},$$

but in fact for all such contrasts, that is to say, for all possibilities of the form

$$\lambda_A\mu_A+\lambda_B\mu_B+\lambda_C\mu_C, \text{ where } \lambda_A+\lambda_B+\lambda_C = 0. \tag{13.5.1}$$

In order to exemplify this, we note that the transformation

$$(\cos\theta)p + (\sin\theta)q, \text{ and}$$
$$(\cos\theta)q - (\sin\theta)p,$$

rotates the original co-ordinates in Figure 13.3.1(a) clockwise by $\theta = 60°$ to co-ordinates given by

$$\left(+\frac{1}{2}\right)[(\mu_A-\mu_B)]+\left(+\frac{\sqrt{3}}{2}\right)[(\mu_A+\mu_B-2\mu_C)\div\sqrt{3}\,] = \mu_A-\mu_C, \text{ and} \tag{13.5.2}$$

$$\left(+\frac{1}{2}\right)[(\mu_A+\mu_B-2\mu_C)\div\sqrt{3}]+\left(+\frac{\sqrt{3}}{2}\right)[(\mu_A-\mu_B)] = (2\mu_B-\mu_A-\mu_C)\div\sqrt{3}$$

and rotates the original co-ordinates in Figure 13.3.1(a) anti-clockwise by θ = -60° to

$$\left(+\frac{1}{2}\right)[(\mu_A-\mu_B)]+\left(-\frac{\sqrt{3}}{2}\right)[(\mu_B+\mu_A-2\mu_C)\div\sqrt{3}] = \mu_C-\mu_B, \text{ and} \qquad (13.5.3)$$

$$\left(+\frac{1}{2}\right)[(\mu_A+\mu_B-2\mu_C)\div\sqrt{3}]+\left(-\frac{\sqrt{3}}{2}\right)[(\mu_A-\mu_B)] = (2\mu_A-\mu_B-\mu_C)\div\sqrt{3}.$$

The co-ordinates arising at (13.5.2) and (13.5.3) are of special interests. Perpendicular projections of the confidence square onto these co-ordinates show that our simultaneous confidence square not only provides confidence limits for the two contrasts we originally introduced as

$$\mu_A-\mu_B \text{ and } (\mu_A+\mu_B-2\mu_C)\div\sqrt{3},$$

but by implication also provides confidence limits for any further contrasts. In the case of

$$\mu_A-\mu_C \text{ and } \mu_C-\mu_B$$

the limits are calculated as follows: in Figure 13.3.1(a) the perpendicular giving the lower limit for $\mu_A-\mu_C$ coincides with the upper left corner of the square, at which corner we have from the result at (13.3.3) that

$$\mu_A-\mu_B = 5\text{-}3.1 \text{ and } (\mu_A+\mu_B-2\mu_C)\div\sqrt{3} = 7.5\text{-}3.1.$$

Solving for $\mu_A-\mu_C$ from these equations, the lower limit for $\mu_A-\mu_C$ is found to be

$$9\text{-}4.25 < \mu_A-\mu_C, \text{ where } x_A-x_C = 9.$$

In this way the following confidence limits are found to be implied by our square:

$$9\text{-}4.25 < \mu_A-\mu_C < 9+4.25 \text{ and } -4\text{-}4.25 < \mu_C-\mu_B < -4+4.25. \qquad (13.5.4)$$

This illustrates the following two properties of all simultaneous confidence regions.

Firstly, the width of the limits at (13.3.3) equals 2×3.1, whereas the width of the limits at (13.5.4) equals 2×4.5. Thus a simultaneous confidence region does not necessarily place different contrasts on an equal footing.

Secondly, any particular choice of values λ_A, λ_B, λ_C at (13.5.1), corresponds to a co-ordinate axis in Figure 13.3.1(a), where the perpendicular projection of our square region onto that axis provides a confidence interval for the value of

$$\lambda_A\mu_A + \lambda_B\mu_B + \lambda_C\mu_C$$

Thus infinitely many contrasts are being taken into account, of which the vast majority are substantively meaningless. For example, consider pi (π = 3.14 ...) and the Napierian constant (e = 2.71 ...), in order to form the contrast

$$[\pi\mu_A+e\mu_B-(\pi+e)\mu_C]\div\sqrt{\pi^2+e^2+\pi e},$$

for which our square would provide the confidence limits. This is surely a substantively meaningless contrast! Whatever could pi and the Napierian constant have to do with oat yields? It will be found that simultaneous-confidence-region procedures cannot avoid providing for such contrasts, because such procedures simply cannot logically be limited to account for certain contrasts only.

13.6 The extent of the confounding

In fairness to simultaneous-confidence-region procedures, the second property uncovered in the previous section should not be allowed to exaggerate the extent of the confounding of any significant data pattern with other data patterns, either contrary or vacuous. This consideration arises because, though the kind of region of present interest provides limits for any one of the infinitely many contrasts of the form defined at (13.5.1), it requires only as many *perpendicular* contrasts as there are dimensions to the Cartesian space involved, to define any region of interest. We speak of

> contrasts $C_1(\mu) = \kappa_A\mu_A+\kappa_B\mu_B+\kappa_C\mu_C$ and $C_2(\mu) = \lambda_A\mu_A+\lambda_B\mu_B+\lambda_C\mu_C$
> as *perpendicular* (or *orthogonal*) if $\kappa_A\lambda_A+\kappa_B\lambda_B+\kappa_C\lambda_C = 0$.

Thus, if for some or other reason one of the following two contrasts would be of interest, the other one would be represented by a perpendicular in two-dimensional Cartesian space:

$$C_1(\mu) = 1\mu_A+2\mu_B-3\mu_C \text{ and } C_2(\mu) = 5\mu_A-4\mu_B-1\mu_C.$$

So, if 'C_1 small' is a pattern of interest, simultaneous statistical inference might want to make allowance for selection by confounding that pattern with patterns of the type 'C_1 large', 'C_2 small' and 'C_2 large'. Any significant pattern is thereby confounded with one contrary pattern and two vacuous patterns. Again, consider the following region wherein the factors are a pair of one-sided lower-bounding regions

$$[4\text{-}3.1 < E(p) < \infty]\times[7.5\text{-}3.1 < E(q) < \infty].$$

Here any significant pattern is being confounded with zero contrary patterns and one vacuous pattern. And again, consider

$$[4\text{-}3.1 < E(p) < 4\text{-}3.1]\times[7.5\text{-}3.1 < E(q) < \infty].$$

Here 'E(p) extreme' is confounded with one contrary pattern and one vacuous pattern, and 'E(q) small' is confounded with zero contrary patterns and two vacuous patterns.

13.7 Scheffé's knowing subject

The normalised form of any contrast as defined at (13.5.1) is given by

$$(\lambda_A\mu_A+\lambda_B\mu_B+\lambda_C\mu_C)\div\sqrt{\lambda_A^2+\lambda_B^2+\lambda_C^2}\ ,$$

whose estimated standard error is given,

not by $\sqrt{2(s^2 \div n)}$, but by $\sqrt{s^2 \div n}$. (13.7.1)

For the present purposes, however, we express any contrasts of interest in 'normalised × $\sqrt{2}$ code', so the form on the left at (13.7.1) is always the correct one. Using this code let us then consider a (1-α) simultaneous confidence region of a kind introduced by Scheffé (1953), which in our oat-yield case is a circular region as depicted in Figure 13.3.1(b). The circle is centred for any contrast at the estimated value of that contrast, and has a radius given by

$$\sqrt{2(s^2 \div n)} \times \sqrt{(3\text{-}1)F_{(3\text{-}1),\, 3(n\text{-}1)}(\alpha)}\,, \tag{13.7.2}$$

where the $F(\bullet)$-like quantity is exceeded with probability α by Snedecor's F on 3-1 and 3(n-1) df (n = 9 in the present case). We previously noted the resemblance of the square region depicted in Figure 13.3.1(a) to the rectangular region developed for the travelling-gene problem of Example 6.7.1. But, though both regions convey *statistically* well-defined meanings, the travelling-gene region estimates a vector, [E(p), E(q)], which *as such* also conveys a *substantively* well-defined meaning, whereas the oat-yield region estimates a corresponding vector, but one that *as such* is substantively meaningless. Very much the same holds for the resemblance of the circular region depicted in Figure 13.3.1(b) to the circular region developed for the problem of the earth's magnetic pole in Example 6.7.2. Each circle can in its own right be *statistically* explicated in terms of repetitive attempts at covering a mathematically well-defined point with a specified long-run probability of success. However, the point called 'the position of the earth's magnetic pole' has a well-defined substantive meaning, whereas the point in the oat-yield example has no meaningful status in agronomical science, none at all. Certainly, agronomical science is unacquainted with any point on the wall of the agronomy building. In order to grasp the confounding implicit in the development of the circular oat-yields confidence region, let

$$C_1(\mu) = \kappa_A\mu_A + \kappa_B\mu_B + \kappa_C\mu_C \text{ and } C_2(\mu) = \lambda_A\mu_A + \lambda_B\mu_B + \lambda_C\mu_C$$

be any pair of normalised perpendicular contrasts. They are estimated in × $\sqrt{2}$ code by

$$C_1(x) = (\kappa_A\bar{x}_A + \kappa_B\bar{x}_B + \kappa_C\bar{x}_C)\sqrt{2} \text{ and } C_2(x) = (\lambda_A\bar{x}_A + \lambda_B\bar{x}_B + \lambda_C\bar{x}_C)\sqrt{2},$$

respectively, in which terms the circle depicted in Figure 13.3.1(b) is given by

$$[C_1(x)]^2 + [C_2(x)]^2 = \left[\sqrt{2(s^2 \div n)} \times \sqrt{(3\text{-}1)F_{(3\text{-}1),\, 3n\text{-}1)}(\alpha)}\right]^2. \tag{13.7.3}$$

If, for instance, interest is in μ_A- μ_B, we could estimate the two perpendiculars as

$$C_1(x) = (\bar{x}_A - \bar{x}_B) \text{ and } C_2(x) = (\bar{x}_A + \bar{x}_B - 2\bar{x}_C) \div \sqrt{3}.$$

The equation at (13.7.3) would then take the form

$$(\bar{x}_A - \bar{x}_B)^2 + [(\bar{x}_A + \bar{x}_B - 2\bar{x}_C)^2 \div 3] = \left[\sqrt{2(s^2 \div n)} \times \sqrt{(3\text{-}1)F_{(3\text{-}1)2(n\text{-}1)}(\alpha)}\right]^2.$$

This shows that data patterns of the type

'$\bar{x}_A - \bar{x}_B$ small' and '$\bar{x}_A - \bar{x}_B$ large',

are being confounded with each other, and with data patterns of the type

'$\bar{x}_A+\bar{x}_B-2\bar{x}_C$ small' and '$\bar{x}_A+\bar{x}_B-2\bar{x}_C$ large'.

Thus for any significant pattern of the type 'x_A-x_B extreme' the confounding is of the type 'one contrary pattern and two vacuous patterns'. In case of four treatments, the circle in Figure 13.3.1(b) is replaced by a sphere given by

$$[C_1(x)]^2+[C_2(x)]^2+[C_3(x)]^2=\left[\sqrt{2(s^2\div n)}\times\sqrt{(4\text{-}1)F_{(4-1)4(n-1)}(\alpha)}\right]^2, \qquad (13.7.4)$$

where the perpendiculars might for instance be

$C_1(x)=\bar{x}_A-\bar{x}_B$,
$C_2(x)=\bar{x}_A+\bar{x}_B-2\bar{x}_C$, and
$C_3(x)=\bar{x}_A+\bar{x}_B+\bar{x}_C-3\bar{x}_D$.

Thus for any significant pattern of the type '$\bar{x}_A-\bar{x}_B$ extreme', the confounding is of the type 'one contrary pattern and four vacuous patterns'. For 5, 6, 7, … treatments the foregoing algebra extends, even though the corresponding geometry (the circle or the sphere) falls away. Thus for m treatments, the confounding of any significant data pattern, for instance '$\bar{x}_A-\bar{x}_B$ extreme', is of the type 'one contrary pattern and $2(m\text{-}2)$ vacuous patterns'. For even quite small numbers of treatments the extent to which the instruments of investigation are blunted by this confounding is horrendous. Referring to the standard error of a difference between two treatment means as a standard error unit, the expressions given at (13.7.3) and at (13.7.4) show that for significance at the α level, an observed difference between two treatment means must exceed

$\sqrt{(m\text{-}1)F_{(m-1),\,m(n-1)}(\alpha)}$ such standard error units.

Thus for significance at the $\alpha = 0.05$ level, and with many degrees of freedom for error, an observed difference between two treatment means must exceed

3.1 rather than 1.96 standard error units in case of five treatments,
4.1 rather than 1.96 standard error units in case of ten treatments, and
4.9 rather than 1.96 standard error units in case of fifteen treatments. (13.7.5)

For finite numbers of degrees of freedom for error the situation is worse.

13.8 Tukey's knowing subject

At (13.5.4) we saw that a simultaneous-confidence-region procedure need not place its inferences on an equal footing; for any given specification of the confidence coefficient, confidence bounds for certain inferences might be chosen to be narrower than those for other inferences. Epistemologically the idea is extremely odd; we are expected to take seriously an epistemology where, for *given data,* the investigator can get to 'know' some things more precisely by being willing to get to 'know' other things less precisely. This is a remote consequence of epistemology that at the very outset was based on the idea of using Student's *t* statistic for one-sided, two-sided and more generally biased 'knowing'. The awful damage wreaked by Scheffé's knowing subject, as exemplified at (13.7.5), has

prompted procedures in which bounds for certain inferences of special interest are narrower than for inferences of lesser interest. A very popular outcome of these ideas is a simultaneous-confidence-region procedure attributable to Tukey (1953) and described by Scheffé (1959). It considers all possible normalised comparisons, and it chooses to draw unbiased confidence bounds for all possible comparisons between two means at a time, as narrowly as possible, at the cost of broader bounds for all other comparisons. In the case of our oat yields, the resulting region takes the form of the hexagon depicted in Figure 13.3.1(c). For a $(1 - \alpha)$ simultaneous-confidence region, the bounds for any of the three possible pair-wise comparisons are given by

$$\bar{x}_I - \bar{x}_J \pm q_{3,\,3(n-1)}(\alpha) \times \sqrt{s^2 \div n} \text{ for } (I, J) = (A, B), (A, C), (B, C), \qquad (13.8.1)$$

where the q(•)-like quantity is exceeded with probability α by the Studentised range statistic, and is obtainable for $\alpha = 0.05$ and $\alpha = 0.01$ from the tables of May (1952). For our case (m = 3 and n = 9) the requisite q(•)-like quantity for $\alpha = 0.01$, equals 4.54. So the bounds for a 99% region turn out to be

$\bar{x}_A - \bar{x}_B \pm 3.2$, $\bar{x}_A - \bar{x}_C \pm 3.2$, and $\bar{x}_C - \bar{x}_B \pm 3.2$, as depicted in Figure 13.3.1(c).

In the case of Scheffé's procedure, these bounds are

$\bar{x}_A - \bar{x}_B \pm 3.3$, $\bar{x}_A - \bar{x}_C \pm 3.3$, and $\bar{x}_C - \bar{x}_B \pm 3.3$, as depicted in Figure 13.3.1(b).

Note that in Figure 13.3.1(b) the limits for $\mu_A - \mu_C$ are to be found on the $\mu_A - \mu_C$ axis, and not on the circumference of the circle. With Scheffé's procedure the bounds for *any* normalised contrast in $\times \sqrt{2}$ code are obtained as the estimated value ± 3.3. With Tukey's procedure, however, the bounds for, for instance, the contrast

$$[(\bar{x}_A - \bar{x}_B - 2\bar{x}_C) \div \sqrt{6}] \times \sqrt{2} \qquad (13.8.2)$$

will be wider than ± 3.2. This contrast is perpendicular to $\bar{x}_A - \bar{x}_B$. Figure 13.3.1(c) therefore shows that its lower and upper confidence limits coincide with the lower and upper corners of the hexagon, respectively, at which corners

$\bar{x}_A - \bar{x}_B = 5$ and $\bar{x}_C - \bar{x}_B = -4 - 3.2$, and
$\bar{x}_A - \bar{x}_B = 5$ and $\bar{x}_C - \bar{x}_B = -4 + 3.2$, respectively.

Solving from these two sets of equations for the values of the contrast defined at (13.8.2), we find its confidence bounds for Tukey's procedure are given by

$[(\bar{x}_A + \bar{x}_B - 2\bar{x}_C) \div \sqrt{3}] \pm 3.7$.

For any number, m, of treatments the expression at (13.8.1) becomes

$$\bar{x}_I - \bar{x}_J \pm q_{(m-1),\,m(n-1)}(\alpha) \times \sqrt{s^2 \div n} \text{ for all } (I, J) \text{ such that } I \neq J. \qquad (13.8.3)$$

Recalling that here the q-like quantity incorporates $\sqrt{2}$ to provide for the standard error of the difference between two treatment means and, again referring to that standard error as a standard error unit, we find from the expression at (13.8.3) and from May's tables, that for significance at the $\alpha = 0.05$ level, and with many degrees of freedom for error, an observed difference between two treatment means must exceed

2.7 rather than 1.96 standard error units in case of five treatments,
3.2 rather than 1.96 standard error units in case of ten treatments, and
3.4 rather than 1.96 standard error units in case of fifteen treatments. (13.8.4)

For finite numbers of degrees of freedom for error the situation is worse. The facts at (13.8.4) are not quite as bad as those at (13.7.5), but remain bad, as we are dealing with normally distributed quantities that have known standard errors. We must ask: 'What has happened to the ubiquitous statistical rule that any discrepancy in excess of 2.58 standard error units is highly significant'? Clearly, any investigator who would employ Tukey's multiple comparison procedure stands to overlook glaringly obvious treatment effects. Small wonder then that Tukey had misgivings about publishing the procedure. Instead, by circulating the work in unpublished form, he cleverly gets credit for correct mathematics, as well as for suspecting an extra-mathematically flaw. But his suspicion never converts to understanding; he never comes to understand that the elaborate probability calculations that purport to account for the possible Type I errors of what he calls 'confirmatory data analysis' rely, for their tenability, on assumed zero probabilities of any Type II errors of what he calls 'exploratory data analysis'. He does not understand that *all* data analysis is exploratory, and that it is impossible to compute any 'probabilities of error' in respect of its findings; instead all that it is possible to compute, are measurements of the quality of fit of those findings.

13.9 Dunnet's knowing subject

A simultaneous-confidence-region procedure introduced by Dunnet (1955) considers all the possible normalised comparisons simultaneously and draws unbiased confidence bounds for all pair-wise comparisons with a control, as narrowly as possible and at the cost of broader confidence bounds for all the other comparisons. In the case of our oat-yield example the resulting region takes the form of a parallelogram, as depicted in Figure 13.3.1(d), treatment A being the control. For a (1-α) simultaneous-confidence region, Dunnet's bounds for the two possible pair-wise comparisons with the control are then given by

$$\bar{x}_A - \bar{x}_I \pm t'_{\text{2-one, 3(n-1)}}(\alpha) \times \text{for } I = B, C, \qquad (13.9.1)$$

where the t' (•)-like quantity is exceeded with probability α by Dunnet's many-one-*t* appropriate for the present case [two-one-t, on 3(9-1) df], and is obtainable from prepared tables (Dunnet 1955). For a 99% region in the present case these bounds turn out to be

$\bar{x}_A - \bar{x}_B \pm 3.07 \times 0.990$ and $\bar{x}_A - \bar{x}_C \pm 3.07 \times 0.990$, i.e.
$\bar{x}_A - \bar{x}_B \pm 3.0$ and $\bar{x}_A - \bar{x}_C \pm 3.0$, as depicted in Figure 13.3.1(d).

From Figure 13.3.1(d) we see that the lower and upper bounds simultaneously obtained for $\bar{x}_C$-$\bar{x}_B$ are given by the solutions of

$\bar{x}_A - \bar{x}_B = 5\text{-}3.0$ and $\bar{x}_A - \bar{x}_C = 9\text{+}3.0$, and of $\bar{x}_A$-$\bar{x}_B = 5\text{+}3.0$ and $\bar{x}_A - \bar{x}_C = 9\text{-}3.0$, i.e. by
$\bar{x}_C - \bar{x}_B - 6.0$, and $\bar{x}_C - \bar{x}_B + 6.0$, respectively.

Again, just as in previous sections we find that even for its intended results, the idea of simultaneity grievously blunts our instruments of investigation. From Dunnet's tables we find, for

instance, that for significance at the $\alpha = 0.05$ level, and with many degrees of freedom for error, an observed difference between the mean for control and that for any other treatment must exceed 2.23 (rather than just 1.96) standard error units in the case of five treatments.

13.10 One question many answers?

Figure 13.3.1 displays, for the comparison of the mean yields recorded for oat varieties A and B, four distinctly different answers. Moreover, the developments around this display have made it clear that there could be many more such answers arising from questions such as: Do you want to know 'one sidedly' or 'two sidedly'? If the latter, do you want to know 'biasedly'? For what error rates do you wish to settle? Are you of the 'conservative' school of thought that would unleash the Type II error rate in order to keep the Type I error rate under control, or are you of a less conservative school of thought? Must A be considered a control? And so on, and so forth. Further developments, for instance the Waller-Duncan Bayesian k ratio multiple comparison procedure, have also incorporated metaphysics of belief (Waller and Duncan 1969). We must ask: Is this endless variety of procedures being referred to by those of our statistical friends who exhort us to be proud of the fact that 'statistics is not so simple a science that it produces only one answer to a given question?' And must we not also ask: Is it not the case that any unambiguous question in substantive science can have only one *correct* answer? Or is that a snare and a delusion? Surely, as a first step toward answering these questions, we must strip away any manifest redundancy. *Otherwise what good is it to have a forceful theory of minimal sufficiency, if only to then ignore what it tells us?* So, note that when all redundancy is stripped away, we find for the class of models given at (13.1.1) that the information required for judging the tenability of any value assigned to μ_A-μ_B in respect of a given data set, and *all* of that information, is conveyed by the minimal sufficient statistic for μ_A-μ_B, that is to say, by

$$\{\overline{X}_A\text{-}\overline{X}_B,\ S^2\}.$$

Here Student's t is appropriate, owing to which 99% confidence limits for μ_A-μ_B on the oat-yield data given at (13.3.1) are obtained as

$$5\text{-}2.80(0.990) < \mu_A\text{-}\mu_B < 5\text{+}2.80(0.990), \text{ i.e. as}$$

$$5\text{-}2.77 < \mu_A\text{-}\mu_B < 5\text{+}2.77. \qquad (13.10.1)$$

Now recall that the 99% confidence limits given here can also be interpreted as a pair of (0.005, 0.995) and (0.995, 0.005) co-ordination limits, and note that the four pairs of corresponding confidence limits given in Figure 13.3.1 can be similarly interpreted. Then it appears at once from the width of the different pairs of limits that, because of the confounding of the minimal sufficient statistic with vacuous statistics by any of the multiple comparison procedures involved, any co-ordination tests derived from those procedures have inferior separating characteristics compared to the test derived from the confidence limits at (13.10.1) where any such confounding has been avoided. Now note further that confidence limits of the form given at (13.10.1), but based on 0.005 as *the physical norm of extremity*, are given by the 99.5% limits

$$5\text{-}3.09(0.990) < \mu_A\text{-}\mu_B < 5\text{+}3.09(0.990), \text{ i.e. by}$$

$$5\text{-}3.06 < \mu_A\text{-}\mu_B < 5\text{+}3.06. \qquad (13.10.2)$$

Comparison of the limits at (13.10.1) to those at (13.10.2) then shows that for the self-same physical norm of extremity (here 0.005 vs. 0.995, or 0.995 vs. 0.005) the pair of unbiased two-sided confidence limits are wider than the corresponding co-ordination limits based on that same norm. So, any co-ordination test we derive from two-sided confidence limits has inferior separating characteristics compared to those derived from the corresponding pair of one-sided confidence limits.

The reader should carefully note that the two sources of inferior separating characteristics uncovered in the foregoing arise from the same underlying source, that is to say, from the involvement of possible mistakes on the part of a knowing subject, as follows:

(1) We are supposed to envisage an array of different investigations (different repetitions) in which a knowing subject might mistakenly judge μ_A-μ_B to be larger than hypothesised in some of the repetitions and, by the same token, mistakenly judge μ_A-μ_B to be smaller than hypothesised in *other* repetitions.

(2) Having thus envisaged possible mistakes by the knowing subject in respect of the hypothesised values of μ_A-μ_B, we are then persuaded also to envisage possible mistakes by that knowing subject in respect of μ_A-μ_C and μ_C-μ_B.

Here the knowing subject has been represented as a 'deciding subject'. Opposed to that, Bayesian procedures represent the knowing subject as a 'believing subject', owing to which the foregoing arguments fall away, as they rely on the physical theories we refer to as those of operating characteristics and separating characteristics, respectively. As Bayesian reasoning is of a metaphysical nature, its critical evaluation requires a different approach. The distinction between co-ordination tests and Bayesian tests is in important ways on a par with the distinction between astronomy and astrology – and it is not at all the intention here to give offence; it is the intention to record important facts. It is a fact that astronomy is a physical discourse, as its concept of ultimate evidence is that and only that which the human body can be forced to perceive. As opposed to that, it is a fact that astrology is a metaphysical discourse, as its concept of ultimate evidence is that and only that which human belief can be persuaded to perceive. In section 13.13 we return to this point.

13.11 Mean separation procedures

Tukey (1949b) considers how we might assist an investigator who wishes to compare the mean yields of a number of varieties. He states:

> 'At a low and practical level, what do we want to do? We wish to separate the varieties into distinguishable groups as often as we can without too frequently separating varieties which should stay together.' (13.11.1)

This prompts the idea of using cluster analysis (Placket 1971; Jolliffe 1971). Scott and Knott (1974) develop a well-known method of cluster analysis for the purpose of making multiple comparisons. Gates and Bilbro (1978) illustrate its application to four different concrete examples. How must we react to this development? First and foremost, we must establish

how substantive science is being modelled by the development. So let us resort to our old standby: simulation. How then could we simulate the model? A Herculean but illuminating effort is as follows: cross two wheat varieties. Self the F_1. Cultivate 100 F_2 plants. Self the F_2. Cultivate one F_3 plant from each F_2 parent. Self the F_3. Cultivate one F_4 plant from each F_3 parent ... Continue until the F_{10}, at which point we have developed 100 different wheat varieties. Obtain 100 packets of seed from each of the 100 varieties. Draw a random 100 of the 10 000 packets. Discard the other 9 900 packets. Number the 100 random packets 1, 2, 3, ..., 100, and destroy all other means of identifying them. Replicate the 100 seed sources now identified only as numbers 1, 2. 3, ..., 100, using a randomised design, and determine the 100×n yields (n = the number of replicates). The outcome of such a simulation would produce exactly the kind of data Tukey and the cluster analysts have in mind. There could for instance be six clusters, respectively of size 11, size 78, size 1, size 3, size 1 and size 6. Or there might be a single cluster of size 100, or 100 clusters of size 1 each. It then appears that two things are wrong with the idea at (13.11.1). Firstly, there is not a shred of substantive evidence to have us believe that wheat varieties routinely belong to fewer clusters than there are varieties. On the contrary, such an assortment arises only under very special circumstances, such as in the presence of a few Mendelian factors with major effects, or in a grossly heterogeneous mixture of varieties. Secondly, cluster analysis is directed at certain characteristics that an entire *ensemble* of varieties might have, such as: 'This ensemble seems to comprise altogether three clusters of sizes 57, 18 and 25, respectively'. But the question, 'Do varieties A and B differ in yield?' is not effectively addressed by cluster analysis followed up by the question, 'Do A and B belong to the same cluster, or do they belong to different clusters?' That is so because in the set-up envisaged, any question about the comparative yields of varieties A and B is a question about the possible values of μ_A-μ_B, where such a question ought to be dealt with by means of, and only by means of,

$(\overline{X}_A\text{-}\overline{X}_B, S^2)$, which is *the minimal sufficient statistic* for μ_A-μ_B,

not by $(|\overline{X}_A\text{-}\overline{X}_B|, S^2)$ which is *insufficient* for μ_A-μ_B,

and *not* involving $\overline{X}_C$, $\overline{X}_D$, $\overline{X}_E$, ..., which are *vacuous* for μ_A-μ_B. (13.11.2)

Example 13.11.1

Shulkeum, as reported by Duncan (1955), performed a yield trial with seven varieties of barley in a randomised block design, and obtained the data reproduced in Table 13.11.1, where the 21 pair-wise differences amongst the seven means are also shown. We see at a glance that several of those differences exceed 10, i.e. exceed 2 × the standard error of the difference between those pairs of means. Much of the statistical literature would have us believe that these 21 tests involve redundancy because, so that literature tells us, amongst seven different means there are only 7-1 algebraically independent contrasts. The reason given is indeed a mathematical fact, but the conclusion drawn from it is false, because it tacitly assumes that beyond the numerical facts given in the table, the investigator is completely ignorant. The assumption is inadvertently promoted by calling the varieties 'A, B, C, …' whereas they are known to agronomy by names such as Brewer's Delight, Socks, Early Pearl, and so forth. To the agronomist these names convey a great deal of information – this one is exceedingly drought resistant, that one is prone to lodging, the next one has a particularly short growing season, and so forth. Owing to such substantive matter, each of the 21 tests performed in Table 13.11.1 is distinctly different from the other 20.

Table 13.11.1: Yields of seven varieties of barley from a randomised block design

Treatment (variety of barley)	A	B	C	D	E	F	G
Treatment mean	49.6	58.1	61.0	61.5	67.6	71.2	71.3
Minus the mean for treatment A	•	8.5	11.4	11.9	18.0	21.6	21.7
Minus the mean for treatment B		•	2.9	3.4	9.5	13.1	13.2
Minus the mean for treatment C			•	0.5	6.6	10.2	10.3
Minus the mean for treatment D				•	6.1	9.7	9.8
Minus the mean for treatment E					•	3.6	3.7
Minus the mean for treatment F						•	0.1
Estimated standard error of a treatment difference: 5.15 on 30 degrees of freedom							

The essence of this is reasoning to be found in the work of the 14th century philosopher William of Occam. In epistemology Occam is a data analyst. Consider the proposition 'Tom is a bald man'. It analyses the data named Tom into

(a datum of baldness)+(a datum of manliness)+(a residual datum).

The residual is a *co-ordinate datum*, where that goes to show that we are continually analysing (and adding to) the set of 'all data'; we are continually partitioning the set of all data in different ways. Occam's opening argument is then that of 'the displaced finger':

> See me move my finger. Where, in 'the analysis of Tom', does this new datum now belong?

It is not part of the 'datum of baldness'. And it is not part of the 'datum of manliness'. So we are compelled to make the 'displaced finger' become part of the 'co-ordinate datum' (the residual), where we then find (and this is the point of the argument) that we are able to analyse the set of 'all data' such that subsidiary sets of *relevant* data are co-ordinated by residual data. Toward clarifying that relevancy, Occam then advances his famous rule for discourse, reproduced here as Axiom 13.11.1.

Axiom 13.11.1: William of Occam's principle of paucity

> It is vain to try to do with more what can be done with fewer.

We have of course met this rule before:

> Never, ever introduce a constituent that is not needed.

Occam holds that different languages give rise to different *conventional signs* when just one *natural sign* (concept) might be involved. Thus for instance 'horse', 'equus', 'perd', … are different conventional signs that bring the selfsame natural sign into the human mind, that is to say, the natural sign *cannot be in any language*. So Occam says *inter alia*

that we must not try to do with more *conventional* signs 'what can be done with fewer' *natural* signs. But our interest in the principle of paucity must go beyond that, and here it will be helpful to note that the principle could also be called the *principle of minimal sufficiency*, which brings us back to the distinctions

minimal sufficient statistic, *insufficient statistic* and *vacuous statistic*.

When Occam says, 'It is vain to try to do with more' he rules out the *vacuous* statistic. When he says, 'what can be done' he calls for the *sufficient* statistic. And when he says 'with fewer' he calls for the latter to be *minimally* sufficient. That much is accomplished at (13.11.2). We must, however, refuse to agree to it that, inasmuch as Table 13.11.1 tells us that

'B-A' = 8.5 'C-A' = 11.4 'C-B' = 2.9, (13.11.3)

there is a 'redundancy' because purportedly

'C-A'-'C-B' = 'B-A'. (13.11.4)

We must refuse because at (13.11.3) and (13.11.4) the symbolism 'C-A', 'C-B' and 'B-A' is made to refer to the numerical data only, whereas a sound epistemology must insist that the symbolism 'A', 'B' and 'C' be made to convey any substantive information about the *varieties* involved. If interest in 'C-A' arises because amongst the seven, they are the only two varieties that produce high lysine seed, and interest in 'C-B' arises because amongst the seven, they are the only two varieties that have a growth season of less than 90 days, that does not at all imply that it would then be redundant for interest in 'B-A' to arise because amongst the seven, they are the only two varieties that do not lodge when grown under wet climatic conditions. The issue is not at all difficult. In fact, silly epistemology can almost always, if not always, be avoided in three easy steps, as follows:

Step 1: Analyse the substantive investigator's questions. What are they? Do they make sense? Try to replace them with analytic subsidiary questions. Determine which, if any, of those questions can be expressed as a question about the possible values of a scalar parameter. (Be leery of questions about a vector parameter that fails to convey just one singular concept, i.e. one single concept that cannot be analysed into several self-contained subsidiary concepts.)

Step 2: For each question in turn find the minimal sufficient statistic for addressing the question, *inter alia* avoiding any vacuous statistics. This is relatively easy in the event of questions about the possible values of a parameter.

Step 3: Address each question in turn, in terms of, and only in terms of, the minimal sufficient statistic for addressing that question. Try to avoid replacing the minimally sufficient statistic with an insufficient statistic.

This has been accomplished in Table 13.11.1 insofar as the table goes. We will now show that over and above the 21 contrasts in Table 13.11.1, numerous further contrasts might be considered, without substantive redundancy. Let us for that purpose pretend the barley data is South African, and let us imagine how a South African investigator might further analyse that data without redundancy, as follows: if varieties A and D are susceptible to damage

by an unseasonable cold snap as often experienced in the Golden Gate area, the suite of shortfall tests in Table 13.11.2 would indicate varieties G, F and E as suitable for that area.

Table 13.11.2: Shortfall tests for varieties not susceptible to unseasonable cold

Treatment (variety of barley)	B	C	E	F	G
Treatment mean	58.1	61.0	67.6	71.2	71.3
Shortfall	13.2	10.3	3.7	0.1	
Hypothesised left-most right co-ordinate	0.026	0.086	0.505	0.729	
Estimated standard error of a treatment mean: 3.64 on 30 degrees of freedom					

(Why not recommend just variety G? Because it is better not to put all our eggs in one basket.) If varieties A, E and G, produce inferior malt, the suite of shortfall tests in Table 13.11.3 would indicate variety F only as suitable for supplying breweries.

Table 13.11.3: Shortfall tests for those varieties that yield satisfactory malt

Treatment (variety of barley)	B	C	D	F
Treatment mean	58.1	61.0	61.5	71.2
Shortfall	13.1	10.2	9.7	
Hypothesised left-most right co-ordinate	0.021	0.070	0.084	
Estimated standard error of a treatment mean: 3.64 on 30 degrees of freedom				

And if only C, E and F, are drought tolerant, the suite of shortfall tests in Table 13.11.4 would indicate F and E as suitable for the dry-land farming practised to the west of Thaba 'Nchu.

Table 13.11.4: Shortfall tests for varieties of barley able to tolerate drought.

Treatment (variety of barley)	C	E	F
Treatment mean	61.0	67.6	71.2
Shortfall	10.2	3.6	
Hypothesised left-most right co-ordinate	0.050	0.364	
Estimated standard error of a treatment mean: 3.64 on 30 degrees of freedom			

Note that for a suite of shortfall tests, the minimal sufficient statistic obviously consists of all the non-negative shortfall contrasts and the error estimate. This is obvious, because the many-one contrast for any specified one of a suite of tests cannot be formed without knowing the yield of the specified one, and so being able to draw the appropriate order-statistical value from the yields of the many. Note also that Tables 13.11.J (J = 1, 2, 3, 4) have involved

$[7(7+1)\div 2]+(5-1)+(4-1)+(3-1) = 37$ different contrasts, and there could be more.

That the purely numerical evidence conveyed by these 37 contrasts is not derived from 37 algebraically independent variables is beside the point. The point is this: in each of the 37 cases, a question in substantive science leads to an item of

a *minimally sufficient* statistic of the form,
{the relevant contrast for the given question, S^2},
together with *further information of substantively relevant scientific nature*.

So, owing to the substantive part of any specified one amongst these 37 items, the full but minimally sufficient information required for that item, cannot be deduced from the full but minimally sufficient information required for the other 36 items. So Occam could not argue in any one of those 37 cases that we could 'do with less'.

We note in passing that certain literature wrongly attributes to Occam a principle known as Occam's razor, according to which, if there are more than one explanation for the self-same observation, we must prefer the simpler explanation. The razor is not due to Occam, and in any case it is nonsense, because if each of two explanations satisfies Occam's principle of paucity, the only way in which we can then discriminate between the two is by obtaining appropriately discriminative data.

13.12 On 'scientific coherence' as opposed to 'numerical coherence'

All multiple-comparison procedures are directed at producing conclusions of the type

'μ_A-$\mu_B < 0$', 'μ_A-$\mu_C > 0$', 'μ_A-μ_D might be 0' and so forth. (13.12.1)

Obviously, such a procedure would not be taken seriously if it failed to provide *numerical coherence*, meaning that if the procedure produces for given *numerical* data conclusions such as, for instance, those at (13.12.1), that procedure must not for the same *numerical* data produce other conclusions in conflict with those at (13.12.1), such as:

'μ_A-$\mu_B = 0$', 'μ_A-$\mu_C < 0$', 'μ_A-$\mu_D < 0$' and so forth. (13.12.2)

However, a data set in the scientific sense, cannot be merely numerical, and owing to that we have been able to develop proof that multiple-comparison procedures cannot provide *scientific coherence*, meaning that any conclusions that a given procedure might arrive at with respect to a given data set, do not contradict other conclusions that the procedure might *imply* in respect of the same data set. In Section 13.2 we saw that if a judgement

of 'significance' or 'non-significance' in respect of the difference between two treatment means is by a given procedure made to depend on whether or not the two means straddle other means, that procedure can very simply be shown to be scientifically incoherent. We also remarked that all multiple-range procedures are incoherent on that basis. In Section 13.3 we saw more generally that inasmuch as all multiple-comparison procedures would have the significance of any observed difference between two means depend on the mere presence or absence of other means, the procedures are scientifically incoherent. It is simply nonsense to have an investigator reporting:

> 'Because ten other peach cultivars were present, the yields of cultivars A and B did not differ significantly, but because only three of the ten are yellow cultivars, the yield of A and B, viewed as yellow cultivars, differed significantly.'

So it must be firmly grasped that numerical coherence does not ensure scientific coherence. The crux of the matter is simply an instance of the principle of paucity:

> Never, ever draw conclusions involving a vacuous statistic.

Mead and Curnow (1983 p. 41) note a common violation of this rule in summaries such as:

> 'A was shown to be superior to B but not superior to C; there was no difference between B and C.'

In the matter of A being, or not being, superior to B, the performance of C is vacuous. There cannot be any objection to adding that the performance of C was intermediate between that of A and of B without its being 'statistically separable' from either of them.

13.13 Lindley's 'knowing subject'

The literature on simultaneous statistical inference is arguably the largest sub-literature of statistics. Certainly it is so extensive that it would be foolish here to try to describe all the numerous recipes proposed for such inferences. Our coverage so far is, however, sufficient for the purpose of coming to grips with the general idea, which is simply this: an investigator conducts an experiment involving several treatments and proceeds to draw conclusions from this, that and the other comparison between treatment means. The *idée fixe* would then have it that for any one conclusion there is a probability of its being erroneous, and the more conclusions drawn, the larger the probability that an error is involved. So we are told to be concerned about the probability of one or more errors amongst several simultaneous conclusions. Here the first of a number of flaws in the reasoning surfaces because if, on the one hand, the conclusions are entirely unrelated, *scientific* simultaneity does no apply, and if on the other hand, the conclusions contribute to a coherent scientific opinion, statistical calculations do not provide for such coherence, as they provide only for *statistical* simultaneity. In order to grasp this, we have but to ask: Do the procedures of Student, Scheffé, Dunnet, etc. ensure that simultaneous conclusions drawn by such procedures, make simultaneous *scientific* sense? We can be compelled to answer that they obviously do not, because if several tests are performed according to the statistical rules for tests performed *separately*, all multiple-comparison procedures hold that the probability of one or more erroneous conclusions

increases with the number of tests performed. However, science holds that the more a coherent scientific view is tested, the more such a probability of error, if at all meaningful, must *decrease*. The next obvious flaw in the procedures is that they would in effect have us replace any observed treatment difference by shrinking it to a smaller difference. (The term 'shrinking', in this sense, is actually used in present-day statistical literature.) With for instance Tukey's procedure, an observed treatment difference, when expressed in its actual standard error units, must, for significance at the 1% level, exceed

2.58 if the number of means is m = 2,
2.91 if the number of means is m = 3,
3.11 if the number of means is m = 4,
3.25 if the number of means is m = 5, ...

A way of implementing this would be to fix the required excess at 2.58 or more, and then to shrink the observed differences as follows for m > 2:

If m = 3, shrink each observed difference by expressing it in × (2.58÷2.91) code.
If m = 4, shrink each observed difference by expressing it in × (2.58÷3.11) code.
If m = 5, shrink each observed difference by expressing it in × (2.58÷3.25) code.
...

This consequence of simultaneous statistical inference would surely strike visitors from another planet as extremely odd, as it results in the findings of substantive investigators being treated on a par with fishermen's tales about the size of the one that got away. It would seem to such visitors that when a substantive investigator earnestly tries to assure a statistician about the integrity of a measurement of the difference between say responses A and B, the statistician's stock reaction is: 'Come come my dear fellow, it could not have been that much. We'll just have to shrink it down to a believable size.' Shrinking is not limited to frequentists; there are Bayesian forms of shrinking as well. For such data as we have here envisaged, Lindley (1971) explains this as follows (the italics are Lindley's, the notation is ours, and he simplifies the variance notation by considering samples of size n = 1, so that we must interpret his variance notation as 'variance ÷ n'):

'Let $\bar{x}_I$ and μ_I be respectively the sample and population means for the I^{th} sample. Then ... $\bar{x}_I$ is *not* a sensible estimate of μ_I: technically, it is *inadmissible*. A more satisfactory estimate is of the form

$$\left[\frac{\bar{x}_I}{\sigma^2} + \frac{\bar{x}_\bullet}{\tau^2}\right] \div \left[\frac{1}{\sigma^2} + \frac{1}{\tau^2}\right], \qquad (13.13.1)$$

where τ^2 is a number that need not (here) concern us … , and $\bar{x}_\bullet$ is the overall mean. The form of this estimate is a weighted average of the sample mean and the overall mean, and the effect of this weighting is to pull all the "estimates" $\bar{x}_I$ toward the central value $\bar{x}_\bullet$: in particular, the extreme values of $\bar{x}_I$ get the most shift. So already we have a partial answer to the multiple comparison conundrum.

But we can go further. Suppose we are interested in estimating the difference $\mu_I - \mu_J$ $(I \neq J)$. It is reasonably estimated by the corresponding difference of the above estimates, i.e. by

$$(\bar{x}_I - \bar{x}_J)\left[\frac{\tau^2}{\sigma^2 + \tau^2}\right],$$

which, although it does not involve the overall mean, is less than the "natural" estimate $(\bar{x}_I - \bar{x}_J)$, so that the differences have been diminished in magnitude in accordance with common sense.'

He notes that the foregoing 'only deals with point estimation', then uses it as a platform to develop, as a pivot for Bayesian 'interval estimation', the quantity

$$\frac{(\bar{x}_I - \bar{x}_J)}{\sigma\sqrt{2}}\left[\frac{\tau^2}{\sigma^2 + \tau^2}\right]^{1/2}, \text{ to be treated as if an } N(0, 1^2) \text{ variable.} \qquad (13.13.2)$$

He points out the 'shrinking' when he points out that the quantity at (13.13.2) is

'… less than the value we are used to considering, namely $\frac{(\bar{x}_I - \bar{x}_J)}{\sigma\sqrt{2}}$ so that a larger observed difference is needed to obtain a given "significance".' (13.13.3)

He concludes from the foregoing that:

'It seems to me that this recognition that one cannot do the obvious in the *estimation* problem is of great help in considering the *multiple comparison* situation.' (13.13.4)

Clearly, Lindley is using the expressions '*not* a sensible estimate', 'is *inadmissible*', and 'in accordance with common sense' with provocative intent. In fact, he tacitly admits to that when he subsequently uses the expressions 'the value we are used to considering' at (13.13.3) and 'one cannot do the obvious' at (13.13.4). The reader might well be mystified by these provocations. However, they are simply derived (which is not to say explained) as follows:

Let $\bar{X}_I = \mu_I + \varepsilon_I$, where ε_I denotes an error random variable with $E(\varepsilon_I) = 0$, so μ_I represents the true potential of treatment *I*. The error variance is the variance of ε_I, and (true to Bayesian form) the parameter μ_I is treated, not as a fixed constant, but as a random variable. So we have two random variables with

$$\text{Variance}(\mu_I) = \tau^2, \text{ Variance}(\varepsilon_I) = \sigma^2, \text{ and Covariance}(\mu_I, \varepsilon_I) = 0.$$

Lindley then wants us to estimate μ_I by means of the linear regression formula:

$$\mu_\bullet + \beta(\bar{X}_I - \mu_\bullet), \qquad (13.13.5)$$

where $\mu_\bullet$ is the population mean of the μ_I ($I = 1, 2, 3, \ldots$) and β is the coefficient of the regression of the μ_I on the $\bar{X}_I$. Then, reasoning as if n = 1, as Lindley does,

$$\beta = \text{Covariance}(\mu_I, \overline{X}_I) \div \text{Variance}(\overline{X}_I)$$

$$= \text{Covariance}(\mu_I, \mu_I + \varepsilon_I) \div \text{Variance}(\mu_I + \varepsilon_I)$$

$$= (\tau^2 + 0) \div [\tau^2 + 2(0) + \sigma^2]. \qquad (13.13.6)$$

We note further that

$$E(\overline{X}_I) = E(\mu_I + \varepsilon_I), \text{ which equals } E(\mu_I) + 0 = \mu_\bullet.$$

So the overall mean of the treatment means, $\overline{X}_\bullet$, is an estimate of $\mu_\bullet$. Inserting this estimate and the form at (13.13.6) into the form at (13.13.5), we obtain

$$\overline{X}_\bullet + [\tau^2 \div (\tau^2 + \sigma^2)](\overline{X}_I - \overline{X}_\bullet) = (\tau^2 \overline{X}_I + \sigma^2 \overline{X}_\bullet) \div (\tau^2 + \sigma^2).$$

Multiplying both the numerator and the denominator of the expression on the right by $1 \div (\tau^2 \sigma^2)$, we obtain the form given at (13.13.1).

Now how to *explain* this?

Let us begin with something that clearly *is* 'in accordance with common sense'. It can be shown that optimally an AI centre for dairy cattle should each year mate a certain proportion of the pool of cows involved with a certain number new sires for progeny testing. For each sire one obtains a number of contemporary comparisons of the form

> $\overline{x}_I - \overline{x}_J$, where $\overline{x}_I$ denotes the mean production of that sire's daughters in a given herd, and $\overline{x}_J$ denotes the mean production of their contemporaries in the same herd.

The variances of such comparisons will differ, being proportional to

> $(n_I + n_J) \div n_I n_J$, where n_I denotes the number of daughters in the given herd, and n_J denotes the number of contemporaries in that herd. (13.13.7)

Each year a new cohort of young bulls is tested, the contemporary comparisons are used to identify a given number of them as the superior members of that cohort, and seed from the superior members is harvested and frozen for use during the coming year. Any seed from previous cohorts is discarded, and the bulls, superior or otherwise, are slaughtered. This practice enables maximum selection pressure, without undue inbreeding, as the gene pool is a closed one. The point here is that there is no interest in any one particular bull *as such*. No particular one of them would be identified as 'The Great iThemba Bulelani who has occupied the Number 1 Bull Pen for the past nine years'. On the contrary, each cohort is viewed as a random sample from the potential sons of the best cows. Thus, the breeding values of the various bulls can be represented as a random sample, just as Lindley would have us formally represent the μ values in his scheme as a random sample. A weighted mean performance for each bull is obtained by combining his contemporary comparisons in accordance with weights based on the form at (13.13.7). Thus any bull's performance can be represented as $\mu + \varepsilon$, where μ denotes his breeding value, and ε denotes the statistical error of his weighted mean performance. As in Lindley's scheme, μ and ε are uncorrelated, and as in Lindley's scheme, μ is a random variable with constant variance. In the case of ε, however, the weighted mean performances of the different bulls vary in precision, as the numbers of daughters and contemporaries vary from bull to bull. And precisely for that reason regression estimators of the type derived in our

explanation of Lindley's reasoning are, in case of the AI problem 'in accordance with common sense'. The crux of the matter is now simply this:

> If we cannot represent the μ values as a random sample, both of the foregoing two developments collapse.

In the case of the AI centre, both the genetics and the physical acts of the breeding procedure ensure that such a representation, *as a physical model*, is in accordance with common sense. In the case of Lindley's reasoning, however, it is beyond any reasonable contest that, *as a physical model*, the μ values cannot be represented otherwise than as a set of unknown constants. So Lindley has to resort to *a metaphysical model* in which the μ values *are* constants, but 'a knowing subject's *beliefs*' about them are to be modelled by treating them as a random sample of values. The resulting development inexorably ends up finding the perennial albatross of Bayesian inference hanging from its neck when it has to deal with the ultimate question: How are 'the knowing subject's beliefs' to be mapped onto corresponding physical facts, that is to say, onto facts that can be forced upon the human body?

We can now usefully return to the remarks on Bayesian procedures made toward the end of Section 13.10. There we began to consider how such procedures might be scientifically evaluated. There are just three possibilities or approaches (see below), of which the first, though entirely justified, is currently rebuffed by Bayesians and the second and third are the moieties of a proof by dilemma that Bayesian inference can be sensibly viewed only as a technology.

Approach 1: It is entirely justified, but currently quite futile, to point out to Bayesians that their system of verities is not compatible with the evidential rules for discourse of the physical sciences. As noted in Section 13.10, the concept of the 'operating characteristics of a decision-making procedure' and the concept of the 'separating characteristics of a data-analytical procedure', which are exemplified by the 'error rates of an array of hypothesis tests' and 'the sensitivities of a suite of co-ordination tests', respectively, are physically meaningful concepts because their meanings can, by simulation, be forced upon the human body. However, current Bayesian reasoning brushes such notions aside by claiming that it is not the experiences of human body, but the probabilistic beliefs then developed in the human psyche that are the ultimate arbiters of scientific truth. Hence, for the purposes of such arbitration, they hold that the statistician's task is to develop the 'psychological (that is to say, metaphysical) "probabilities of truth" of those beliefs'.

Approach 2: A second method for the evaluation of Bayesian reasoning arises because a Bayesian procedure necessarily proceeds from an array that constitutes the members of a class of models, as required for the introduction of Bayes's formula. Thus, in the context of any investigative procedure, we must call for explication of those commencement tests that necessarily had to justify the introduction of the class characteristic involved. And, as the only discernable methods for such testing rely on frequency physics, any attempt at adjoining such an explication to Bayesian metaphysics generates incoherence.

Approach 3: A third method proceeds from the second method by noting that the incoherence of attempting to adjoin the frequency physics of commencement testing to the metaphysics of Bayesian inference can be avoided by, and only by, introducing the class

characteristic as an *assumption*, rather than an *inference*. However, that can be justified as, and only as, requisite for decision-making under risk, because the statistician must then be able to say 'I *have to* decide on this; so I *have to assume where necessary*, and I have to do so as best I can. If you can convince me that other assumptions would be better, I will adopt them.' The point here is that reasonable assumptions, when necessary, are perfectly acceptable in the context of the use of knowledge, but we must never, ever in the context of the pursuit of knowledge pretend that 'to assume' is 'to know'.

The reasoning in Approaches 2 and 3 provide the respective horns of proof by dilemma that Bayesian inference cannot coherently contribute to investigative discourse, that is to say, to the discourse of the pursuit of knowledge. So, such inference is compelled to try to justify itself in the context of, and only of, the discourse of the use of knowledge. After all, this should hardly come as a surprise, as the stock argument for introducing a Bayesian prior is to plead that such metaphysics 'conveys "knowledge", albeit "vague", additional to the sample information, and we must try to *use* that additional knowledge'. Thus, returning to Definitions 1.2.1 and 1.2.2, we see that Bayesian inference cannot survive otherwise than in the discourse of the *use* of knowledge. (We note in passing that Definition 1.35.3 need not concern us here, as it can involve Bayes's theorem only as a formula in frequency physics.) Hence, having confined Bayesian inference to the discourse of the *use* of knowledge, the term 'Bayesian inference' must be discarded in favour of the alternative designation 'Bayesian decision-making under risk', and consequently we can now develop a greatly strengthened form of Approach 1. This is so because it compels Bayesians to recognise that the *use* of knowledge is directed at *physical* rewards. If for instance the required reward is 'bigger and better potatoes' we cannot possibly take that to mean that our customers will be happy to *believe* that they are busy eating 'bigger and better potatoes' when in fact those potatoes are the same old potatoes as before. The only discernable way in which a Bayesian posterior could have a physical meaning, is for the Bayesian prior to have a physical meaning, and the only discernable way for the prior to have such a physical meaning, is for it to be capable of defence as a frequency distribution – either theoretical, or estimated, or conjectured. Or does the reader know about any other appropriate physical meaning?

It is not the purpose of the foregoing argument to assert that the knowledge arrived at by statistical reasoning cannot fall short of other sources of information. Let us suppose for instance that a zoologist has discovered a new species of rodents and, closely examining just ten of the animals, finds two are female and eight are male. The (0.05, 0.95) and (0.95, 0.05) co-ordinate bounds for binomial sampling would then indicate a population proportion of females somewhere between 0.04 and 0.69. However, the zoologist's knowledge of rodent biology would indicate population proportions much closer to 0.50. Therefore, what our argument *does* assert is only that rodent biology is incapable of making any acquaintance with the results of Bayesian attempts at combining the two sources of information.

It is also not the purpose of the foregoing argument to assert that all personal probabilities are incapable of providing useful information, either in the pursuit of knowledge or in the use of knowledge. What our argument *does* assert is that when Bayesian inference tries to persuade us that its metaphysical inferences represent scientific knowledge, it relies on the ill-defined notion of statistical inference to inadvertently reverse the roles of the

concomitant variable (personal probability) and the target variable (physical reality). This type of confusion is not all that unusual; there exists a statistical literature that obstinately persists with the claim that the linear calibration problem is best solved by using standard regression formulae so to speak 'the wrong way round' (Randall 1985). There also exists a horticultural literature that pursues the 'ideal architecture' for this, that or the other species of fruit tree, without any discernable objective other than that 'ideal architecture' itself.

We note that the argument here developed from Approaches 1, 2 and 3, is just a version of a basic argument developed in Section 4.8. In Chapter 14 we will meet another version of the same basic argument. In fact, with the possible exception of likelihood inference, all the various received theories of statistical inference fall victim to it. Likelihood inference escapes it only by reinterpretation as odds-ratio testing, whereby it relinquishes any claim to be a theory of inference, as any theory of inference must necessarily have a knowing subject; otherwise, who infers?

13.14 Cox's 'knowing subject'

D.R. Cox has addressed the notion of multiple comparisons (more correctly referred to as 'simultaneous statistical inference') on very general grounds by way of what he refers to as 'making allowance for selection' (e.g. Cox 1977, p. 51). He reasons in the context of significance tests rather than of hypothesis tests. However, his reasoning fails to properly separate two quite different possibilities, of which the first is explained as follows: an investigator of a solitary real-world data set wishes to test a hypothesised model against a particular alternative, and finds that different tests of significance might be used for that purpose. So, using various different tests in turn, he obtains the significance levels $p_{obs}(1)$, $p_{obs}(2)$, $p_{obs}(3)$ ... . The investigator selects the smallest of those observed levels, and wrongly interprets that as the appropriate significance level, that is to say, as the probability of mistakenly concluding that the evidence thus obtained is just decisive against the hypothesised model. Cox proposes (correctly so) that allowance must be made for that level having been selected from several levels, which allowance is made by interpreting the selected level as the smallest order-statistical value in the set

$$\{p_{obs}(1), p_{obs}(2), p_{obs}(3), ...\}, \tag{13.14.1}$$

and then taking the distribution of the corresponding order statistic as the test distribution. In order to see that this reasoning is correct, we could resort to simulation, where it will be found that the smallest significance level might be $p_{obs}(7)$ in a first repetition, $p_{obs}(2)$ in a second repetition, $p_{obs}(9)$ in a third repetition, ..., thus showing that as far as the choice of test statistic goes, Cox's reasoning is correct. We subsequently show that he employs the statistic wrongly, but for the moment that need not concern us. Let us consider a concrete example.

Example 13.14.1

In sheep one would expect body mass and number of crimps per centimetre of the wool to be uncorrelated. This was confirmed by within-flock correlation coefficients obtained from 39 different flocks of sheep. The 39 null-hypothetical right statistical co-ordinates were uniformly distributed over the unit interval, except for one co-ordinate that seemed

remarkably small. Taking the statistical rounding to be zero, the j^{th} smallest co-ordinate, $V_{(j)}$, is transformed as follows to Snedecor's *F* (Wilkinson 1933; Ling 1992):

$$\{[1-V_{(j)}]\div V_{(j)}\}\div\{2(39+1-j)\div 2j\} = F \text{ on } 2(39+1) \text{ and } 2j \text{ df. for } j = 1, 2, 3 \ldots, 39.$$

This is improved by employing mid-co-ordinates (Stone 1969); these may be taken to be

$$\text{either of the form } U_{(j)} + 0.5\,\varepsilon, \text{ or of the form } 0.5\,\varepsilon + V_{(j)}.$$

In the present case the ε values were ≈ 0. So

$$V_{(1)} = 0.0001,\ V_{(2)} = 0.0758,\ V_{(3)} = 0.0994,\ V_{(4)} = 0.1029, \ldots$$

could be considered mid-co-ordinates and transformed to corresponding *F* values whose hypothesised right co-ordinates thus turned out to be

$$0.004,\ 0.806,\ 0.758,\ 0.580 \ldots, \text{ respectively.} \qquad (13.14.2)$$

The first of these co-ordinates points at a possible outlier. So, as a test for zero correlation of body mass and crimps per centimetre of the wool, it seems appropriate to test the co-ordinates of the median V (of $V_{(20)} = 0.5621$), whose co-ordinates under the hypothesised model, (0.63, ε, 0.37), are not indicative of non-zero correlation. Cox (1977, p. 51) views such problems from a significance tester's point of view; so he would have us imagine the behaviour of a knowing subject who examines *k* significance levels, selects the smallest level and (ignoring the other $k - 1$ levels) regards the selected level as just decisive against the hypothesised model. In order to then make allowance for selection, Cox would have us calculate the significance level as the Type I error rate of such behaviour. He makes q_{obs} denote the smallest original level, noting that in cases such as ours the required error rate is given by

$$p_{obs} = 1-(1-q_{obs})^k, \text{ which in our case} = 1-(1-0.0001)^{39} = 0.004. \qquad (13.14.3)$$

It would require immensely precise calculations to display the numerical distinction between the result at (13.14.3) and the initial term of the array at (13.14.2). But that is not the point of this example. The point is this: the reasoning that has lead us to the array at (13.14.2) did not involve a knowing subject; it simply established what test statistic was being used, as in fact we could have established by a simulation as described. So, contrary to Cox's reasoning, the behavioural device invoking his knowing subject is not needed for problems of the present kind. In order to judge his device we must consider problems in which its results could not otherwise be obtained. That brings us to the second of the two possibilities covered by Cox's reasoning.

The second possibility can be explained as follows: an investigator of a solitary real-world data set, wishing to test whether or not the value of a certain parameter might be zero, finds that a recommended test statistic tends to be positive when the parameter is positive, and tends to be negative when the parameter is negative. Thus, finding that the value observed is (say) negative, the investigator computes the significance level as the probability under the hypothesised parameter value zero, that the test statistic would take a value as small, or smaller, than observed, and then interprets this level as one of which a small value points at a negative parameter value. Denote this significance level as $p_{obs}^{(-)}$.

Cox argues that, had the observed value been a positive one, the investigator would then have computed the significance level as the probability under the hypothesised parameter value zero, that the value of the test statistic would be as large, or larger, than observed, and would have interpreted that level as one in which a small value points at a positive parameter value. Denote this significance level as $p_{obs}(+)$. Cox argues that in either case the investigator would in effect be performing two different tests corresponding to $p_{obs}(-)$ and $p_{obs}(+)$, respectively, and would be selecting the smaller of the two levels as the appropriate significance level. So he argues that allowance for that selection be made by interpreting the test statistic actually used as the smallest order-statistical value in the set

$$\{p_{obs}(-), p_{obs}(+)\}, \tag{13.14.2}$$

and then taking the distribution of the corresponding order statistic as the test distribution. In order to show that 'allowance for section' from the set at (13.14.1) and 'allowance for selection' from the set at (13.14.2) are two very different kettles of fish, we might again resort to simulation, in which case a glaringly obvious distinction will appear, as follows: for the simulation previously required, repetitive drawing from a hypothesised model and from *just one* alternative will suffice. But for the simulation presently required, the hypothesised model and *more than just one* alternative must be considered, because the hypothesised parameter value cannot, for just one solitary real-world data set, at one and the same time be both too large and too small. The next example shows how that leads to a fundamental disagreement between a co-ordination test and a significance test.

Example 13.14.2

Let the mean difference in observed yields of two bean cultivars be +19, whose estimated standard error is 8.6 on 10 degrees of freedom, and let Student's *t* be appropriate as a test statistic for judging the quality of fit of the population model whose mean is zero. A co-ordination test places the mental correlate of the observed value of *t* at (0.948, ε, 0.052) in Student's test distribution. A reasonable person would surely find this a rather awkward fit, pointing instead at certain models where the population mean deviates from zero in favour of the cultivar with the higher observed yield. Opposed to that, Cox would have us imagine how his knowing subject might have observed a difference of -19, and thus have concluded in favour of the other cultivar. So we must calculate that such behaviour would lead to wrong conclusions in as much as 10.4% of cases, and we must use that to argue that the co-ordination test has misled us *in this particular case*. Moreover, we must argue that a co-ordination test in the imaginary case would have been equally misleading. A co-ordination tester will reject this argument, arguing instead as follows:

> The *observed* co-ordination (0.948, ε, 0.052) favours the first of the two cultivars. The possibility of *another* case with observed co-ordination (0.052, ε, 0.948), is utterly irrelevant, as that is not the data in hand, and the imaginary behaviour of a knowing subject is not part and parcel of how the data in hand came about. So, any valid explanatory model for the data in hand, cannot partake of that imaginary behaviour.

Cox's knowing subject reflects a failure to distinguish between Definitions 1.2.1 and 1.2.2, owing to which, as the foregoing example demonstrates, Cox confounds indicative data patterns with contrary data patterns, and this then leads him to further confounding with vacuous data patterns, as was demonstrated in Example 1.31.1.

It might well be thought that there is no difference between Cox's knowing subject and Fisher's knowing subject, and that might well be so. However, in Chapter 6 we noted that Fisher introduces two-tailed tests without explanation. Moreover, Fisher wears different hats in different writings. In his book on inference he even alternates between different hats in the same book. That makes it difficult for us to know whether or not we interpret him correctly. So it seems appropriate to deal with Cox's epistemology in its own right.

Note in conclusion that Examples 13.14.1 and 13.14.2 have once again underscored that a co-ordination test is not 'a significance test with frills'.

13.15 A stupid question

It is demonstrable that the literature on multiple-comparison procedures perceives the prime question motivating those procedures as:

> 'How does one compare means arising from a set of unstructured treatments following analysis of variance?'

The question is incredibly stupid. There is no point in mincing words; the stupidity is manifested by failure to understand that such comparison must follow on the question:

> 'What was the motivation for choosing those particular treatments? What question(s) did the substantive investigator wish to address?'

When for instance Nelder (1971) says of multiple-comparison procedures that 'their principle use appears to be to lend an air of respectability to otherwise uninteresting sets of data', we must respond by asking: 'How can a data set be "uninteresting", otherwise than not being linked to any specific question that it was intended to address, or that it might subsequently have been perceived to bear upon?' Possibly Nelder means just that when he explains: 'By uninteresting I mean data to which no prior structure attaches and which do not themselves clearly show posterior structure.' We note, however, that a data set does not ask questions; it is *we* who must ask questions. And, in cases where a data set has 'prior structure', as in the case of a factorial structure, it is *we* who prepared the way for certain questions to follow. And, when a data set shows 'posterior structure' it is *we* who perceive that structure and it is *we* who then ask: 'Does it reflect such-and-such?' So, it would be counterproductive not to observe that the multiple comparison muddle has originated in failure to grasp that most of its various procedures begin with vague ideas of the kind expressed at (13.11.1) – ideas that have not been motivated by the purposes of the substantive investigator. Shulkeum's data of Example 13.11.1 is a case in point. More than half a century has elapsed since Duncan (1955) placed Shulkeum's *numerical* data in the literature on multiple-comparison procedures, and since then those numerical data have been used in enumerable papers to exemplify numerical recipes for the implementation of this, that and the other such procedure, without asking why in the first place Shulkeum might have produced those data.

> In the case of a multiple-range procedure: What substantive questions by Shulkeum would make it significant for a pair of means to straddle or not to straddle another mean?

In the case of a simultaneous-confidence-interval procedure: What substantive questions by Shulkeum would make it necessary to consider comparisons *jointly*?

In the case of a cluster analysis: Could Shulkeum have thought the number of true means would be less than the number of observed means? If so, what could have prompted such an idea?

And in each of the foregoing: What question could Shulkeum conceivably have asked about a subset of the seven barley varieties, such that the answer to the question would depend on *the mere existence* of recorded mean yields from the rest of those varieties?

With due respect to Nelder we must nevertheless point out that it is not the data sets that are 'uninteresting'; it is the results of silly procedures that are uninteresting. It is impossible for a substantively sensible question and appropriate data to be found uninteresting by our profession.

13.16 The multiple-comparison 'lore' on factorials

The ill-defined purposes of simultaneous statistical inferences have resulted in profound confusion about when, how and why such inferences are called for. An example of that confusion is an almost universally accepted 'lore' (there is no other way to describe it) according to which such inferences are not applicable to factorial experiments, despite the analysis of such experiments being clearly directed at drawing related conclusions. A typical 2×2 factorial experiment, for instance, usually results in a conclusion conveyed by just three related contrasts in one of two possible ways, as follows:

Possibility 1: The interaction is significant, so the simple effects of one of the factors at the two different levels of the other factor are examined.

Possibility 2: The interaction is not significant, so the main effects of the two different factors are examined.

In both cases the results of the analysis are conveyed by exactly three different contrasts jointly. So it would seem obvious that simultaneous statistical inference would specify an overall error rate, α, and apply a subsidiary error rate, γ, to each of the three contrasts separately, where γ is the solution of

$$1-(1-\gamma)^3 = \alpha.$$

However, the literature on simultaneous statistical inference overwhelmingly holds that that is not appropriate, positing, as the arcane reason for that, that 'the three contrasts are orthogonal', thus generating a meme nested within the wider meme named 'simultaneous statistical inference', which in turn is a meme nested within the yet wider meme named 'statistical inference', which in turn is a meme nested within the yet wider meme named 'probability inference'.

13.17 'More conservative' procedures and 'less conservative' procedures

Facts such as those listed at (13.7.5) and (13.8.4) are referred to in multiple-comparison jargon, not by saying that the former procedure is 'the more destructive' of the two, and that the latter procedure is 'the less destructive' of the two, but by using 'conservative' as a euphemism for destructive. Nevertheless there has been a demand for less conservative procedures. The demand has resulted in a stream of increasingly obscure procedures, so much so that Monte Carlo studies have been conducted to try to fathom what the methods actually do. Chew (1976) for instance describes four Monte Carlo studies (by Boardman and Moffit, 1971; Carmer and Swanson, 1973; Thomas, 1974; and Einot and Gabriel, 1975), but he seems to have been unable to make much use of them. It cannot but be considered very odd that a community of mathematically talent individuals have devised many procedures, persuaded thousands of investigators in substantive science to use those procedures and, having done that, have then resorted to simulation studies in the hope of finding out what they have wrought by way of the procedures.

13.18 A fundamental error

Let us return to a point made in Section 13.13, namely that if several tests are performed according to the statistical rules for tests performed *separately*, all multiple-comparison procedures hold that the probability of erroneous conclusion *increases* with the number of tests performed. But science holds that the more a coherent scientific view is tested, the more such probability of error, if at all meaningful, will *decrease*. For a simple example, consider Theodosius Dobzhansky's book, *Genetics and the origin of species* (1951). He tests his thesis on literally hundreds of independent empirical data sets, many of which are of a statistical nature. So, simultaneous statistical inference must hold that if α_1, α_2, α_3, ... denote the probabilities of error for those of his tests that are of a statistical nature, the probability that his book contains an error is given by

$$1-[(1-\alpha_1)(1-\alpha_2)(1-\alpha_3) \ldots], \text{ which must be very large, owing to the many tests.}$$

And it is believed that his book *does* contain an error in that he cites Ford (1937, 1945) on the latter's evidence for evolution of industrial melanism as protective colouring in moths, evidence which has subsequently been questioned (Hooper, 2002). Yet that has virtually no impact on whether or not we are prepared to accept Dobzhansky's thesis. The lack of impact arises because of the overall coherence of Dobzhansky's reasoning. Such reasoning is capable of surviving this, that or the other factual error not of crucial importance. The reader should note that one can find hundreds of such books. A remarkable example of such a book, and one dealing with the findings of many wide-ranging scientific investigations, but requiring no specialised knowledge for its understanding, is Randy Shilts' book *And the band played on*.

The point here is this: simultaneous statistical inference rests on the fundamental error of thinking that the scientific investigator is completely dependent on the numerical data under consideration, that is to say, is otherwise utterly ignorant, and so can be persuaded to adopt the findings of a set of numerical rules, without having the ability to judge the scientific sense of those findings when viewed in the wider perspective of substantive science.

13.19 A PROFESSION IN DENIAL

The introduction of multiple-comparison procedures has thrown – and continues to throw – the statistical profession into profound disarray. Different schools of thought are deeply divided as to what the methods are supposed to achieve, and even as to what, if anything, they do achieve. A review by O'Neill and Whetherill (1971) has a bibliography of 254 references; yet, despite such a wealth of information, they declare that 'in spite of the importance of the subject, and the amount written on it ... there is still much confusion as to what the basic problems are, what the various procedures achieve, and what criteria and properties should be studied.'

Bryan-Jones and Finney (1983) commend the use of 'Duncan's new test' for pair-wise comparisons 'if an experiment concerns a totally unstructured set of treatments'. But they furiously attack the use of such methods if treatments have a 'logical structure'. They evidently subscribe to the widely maintained lore that simultaneous statistical inference does not apply in cases where treatments are structured, as for instance in the case of a factorial treatment array. However, Hartley (1955) would evidently recognise (quite correctly) that the lore in question is without reasoned defence, as he demonstrates very simply how simultaneous statistical inference can be applied in a factorial case. We can easily supply forceful examples of our own, as indicated in Section 13.16.

Steel and Torrie (1980) devote an entire chapter to a variety of the procedures and then systematically apply them in the rest of their book. The same is true of Ott (1992). In both books various procedures are presented as part and parcel of correct statistical practice. However, both books lack clarity of explanation as to how, in any given case, one chooses amongst the various procedures. Ott, for instance, tries to explain as follows, referring to the procedure we described in Section 13.2 as the SNK procedure (i.e. the Student-Newman-Keuls procedure):

> 'Which procedure should you use? We generally prefer the SNK procedure for efficacy (effectiveness) comparisons. But our reasons for this choice have a great deal to do with our work setting and the regulations surrounding our decision.'

However, any of the procedures described by Ott can be used for efficacy comparisons. So all we learn from this 'explanation' is that a regulatory authority has prescribed the SNK procedure. The reader may well surmise that the prescription did not spring from any proper understanding. More likely than not, it reflects the more persuasive contributions to a learned confusion by certain members of a committee. Ott tries to explain further, as follows:

> ' ... the decision regarding which procedure to use, and when to use it, is up to the individual. For a given problem, determine whether your decisions regarding differences should, in general, be more (or less) conservative. Then choose a procedure that exhibits the desired characteristic.' (13.19.1)

But how 'conservative' *ought* we to be? All that we learn from this further 'explanation' is that it is 'up to the individual'. In other words, we are dealing with a question to which there is no *correct* answer. A similar 'explication' is found in Miller (1981) in a book wide-

ly considered the definitive account on simultaneous statistical inference. He opens the concluding section of his introductory chapter by stating with disarming frankness:

> 'Time has now run out. There is nowhere left for the author to go but to discuss what constitutes a family.' [He means 'family of simultaneous tests'.]
>
> 'This is the hardest part of the entire book because it is where statistics takes leave of mathematics and must be guided by subjective judgement.'

This is followed by three to four pages of attempted explications that are as clear as mud, and then, finally abandoning the attempt, he concludes that (his italics):

> '*There are no hard-and-fast rules for where the family lines should be drawn, and the statistician must rely on his own judgement for the problem at hand.*' (13.19.2)

What kind of 'science' is that, where for a given question, Tom has this answer, Dick has that answer, and Harry has another answer? The 'recommendations' at (13.19.1) and (13.19.2) are a cop-out; they ask *us* to answer the question *they* are supposed to answer.

Snedecor (1956) devotes a section of his book to the procedures, and then ignores them in the rest of the book. His message to substantive investigators is clearly: 'You will come across the use of these recipes in the literature of your subject. Therefore you should know about them. We do not use them.' This stance is maintained in subsequent editions of Snedecor's famous book.

Mead (1990) points out that: 'For many experimenters, and even for editors of journals, [multiple comparison procedures] have become automatic in the less desirable sense of being used as a substitute for thought'. He strongly advises against the use of any of the procedures, 'unless, ... you decide that the method is exactly appropriate'. He gives several examples where such a method is not at all appropriate, but does not give as much as a single example for which they *would* be exactly appropriate, and concludes that 'multiple comparison methods be avoided'. His message to substantive investigators is clearly: 'I am entirely unable to think of any example for which such a procedure could be commended. In fact I strongly suspect there aren't any such examples. But to be on the safe side, I won't go so far as to declare the procedures utterly useless.'

Nelder (1971) states: 'In my view, multiple comparison methods have no place at all in the interpretation of data.'

How can a profession of highly intelligent individuals continually be so divided on a question that arises in their subject, and that has been widely debated for the better part of a century? There can be only one answer: they share and they cling to a deeply entrenched source of confusion. We have clearly seen, by way of reasoning whose import can be forced upon the human body, the introduction of that source at (6.4.1) and (6.4.2). There we saw how R.A. Fisher, in developing his theory of significance tests, violated a fundamental principle of science:

> Never, ever introduce a constituent that is not needed.

The knowing subject of statistical inference, who must infer the inferences, and therefore might, with probability such-and-such, 'infer mistakenly', is not needed. By introducing

that subject Fisher fell into the profoundly confusing notion that statistics requires its very own form of inference; that notion kept confusing him for ever after, and the rest of us, like so many lemmings, followed him into that confusion. In order to abolish the dreadful notion of simultaneous statistical inference, indeed, in order to abolish all the other silly theories of statistical inference, we must learn not to confound significant data patterns with contrary data patterns, and we must learn not to confound informative data patterns with vacuous data patterns. So, as a first step in that direction, we must learn to just say no to Student's two-tailed *t*-test.

Chapter 14

FIDUCIAL INFERENCE

Metaphysical probabilities *sans* metaphysical priors

14.1 Introduction

R.A. Fisher held that 'Bayesian inference', in the sense we considered in Chapter 12, invalidly addresses a valid problem in statistical inference and therefore proposed to develop, under the name 'fiducial inference', the valid solution to that problem. His development is unclear to such an extent that many, even most, statisticians dismiss it as beyond comprehension. However, Fisher is arguably the giant among the founders of mathematical statistics. So we must try to come to grips with the idea of fiducial inference, and as scientists rather than gullible disciples, we must critically judge that idea as best we can.

14.2 The basic idea

Fisher (1973; first published in 1956) quotes Keynes (1921) with evident approval:

> 'If logic investigates the general principles of valid thought, the study of arguments, to which it is rational to attach *some* weight, is as much part of it as the study of those which are demonstrative.' (original italics) (14.3.1)

The italicised word points at probability inference, likelihood inference, or some such notion. That would no longer be the case should we replace the phrases 'to which it is rational to attach *some* weight' and 'are demonstrative', with the phrases 'to which it is possible to attach relevant physical evidence' and 'are logically demonstrative', respectively. However, Fisher throughout his long career was clearly convinced that the valid form of probability inference somehow existed to be discovered; and he was convinced that he was on the brink of that discovery. He took every attempt at such inference seriously, but favoured the idea that his 'fiducial' method would, in some or other clarified form, supply the breakthrough.

He explains (1973, p. 54) that 'fiducial probability', like 'Bayesian probability' in the sense of our Chapter 12, is intended to express a justifiable degree of metaphysical 'belief' about the unknown value of a parametric constant of interest, the difference between the two kinds of probability being, he explains, that whereas

> '... the argument of Bayes requires a distribution *a priori* involving probability statements of the same logical form as those finally obtained *a posteriori*, the application of the fiducial argument can only be made in the absence of such information *a priori*.'

Let a datum, x, be modelled as a realisation of an $N(\mu, 1^2)$ random variable, X, where $-\infty < \mu < +\infty$. The fiducial argument would somehow have our 'beliefs' about the possible values of μ take the form of an $N(x, 1^2)$ distribution of those values. Thus, for instance, a 0.95 fiducial interval for μ is somehow (by fiducial argument) obtained as follows, where $\Pr f(\bullet)$ denotes fiducial probability:

$\Pr(|X-\mu| \geq 1.96) = 0.05$. So, by the fiducial argument,
$\Pr f(|\mu-x| \geq 1.96) = 0.05$, i.e. $\Pr f(x-1.96 < \mu < x+1.96) = 1-0.05$.

In this case, the fiducial interval, $x-1.96 < \mu < x+1.96$, is indistinguishable from the corresponding 0.95 confidence interval. Since this is often – though not invariably –the case, it is important to note from the outset that Fisher emphatically denies that a fiducial probability requires, for its justification, to be capable of any frequency interpretation. This fact comes to the fore for instance in the fiducial treatment of the Behrens-Fisher problem, where it then turns out that a $1-\alpha$ fiducial interval for the difference in the means of two normal populations with unknown and possibly different variances, does not cover the true difference with frequency $1-\alpha$ in repetitive sampling of those two populations (Neyman, 1941). Unfortunately, the examples of such disagreement between frequency physics and fiducial metaphysics tend to be so involved that it will be best for us to avoid them. But that need not concern us, as long as we firmly grasp that fiducial probabilities, as such, are not interpretable as frequencies; they are to be interpreted as degrees of belief only. Any fiducial inference is directed at elimination testing, and Fisher requires us to proceed from a minimally sufficient statistic for the parameter of interest. Furthermore, in the event of an ancillary partitioning we are required to condition on the ancillary. This of course requires clarification, and will strengthen, rather than weaken, Fisher's position if we agree to the clarification developed in Chapter 9.

14.3 A CONCRETE EXAMPLE

Fisher (1973, pp. 54-63) proposes to exemplify the fiducial mode of reasoning by applying it to a random sample of exponential waiting times. So, in order to provide a concrete example, let us apply his reasoning to the waiting times for the first $n = 20$ eruptions of Vesuvius after that of AD 1631, given in years and in order of occurrence in Table 1.15.3. Denoting the expected waiting time as θ^{-1} $(0 < \theta < \infty)$, the maximum likelihood estimator of θ is given by Fisher as

$$T = \frac{n}{X}, \text{ where X is the sum of the waiting times.}$$

Fisher notes that T is minimally sufficient for θ, no ancillary statistics are involved, and X is 'continuous over all positive values, uniformly for all values of θ' – these being requirements (so he asserts) for the fiducial argument to be applicable. Since $2\theta X$ is a χ^2 variable on $2n$ degrees of freedom, it follows that

if P is the frequency with which χ^2 on 2n degrees of freedom exceeds $\chi^2_{2n}(P)$,

then the statement

$$\theta > \frac{T}{2n}\ \chi^2_{2n}(P)$$

'is verified with the frequency P, for all values of P chosen' (to quote Fisher). (14.3.1)

He points out: 'The reasoning developed so far has been entirely deductive.' He then proposes to develop 'the fiducial argument' whereby (so he claims) the status of θ in the form at (14.3.1) is altered from that of a constant to that of a random variable. Hence, for T = t he obtains

$$\Pr f[\theta > \frac{T}{2n}\chi^2_{2n}(P)] = P \qquad (14.3.2)$$

For instance, as the sum of the waiting times for the first n = 20 eruptions of Vesuvius after that of AD 1631, is 313 years, and as χ^2 on 40 degrees of freedom exceeds 55.76 with frequency $P = 0.05$, a 0.95 fiducial interval for θ is given by

$$0 < \theta < \frac{1}{2(313)}(55.76).$$

Thus a 0.95 fiducial interval for the expected waiting time is given by

$$\infty > \theta^{-1} > \frac{2(313)}{55.76}, \text{ i.e. by}$$

$$11.2 \text{ years} < \theta^{-1} < \infty. \qquad (14.3.3)$$

But look at what this leads to! The 20 individual waiting times given in Table 1.15.3 enabled us in Example 1.15.2 to perform a commencement test of the quality of fit of the exponential class characteristic. Using the Cramer-Von Mises test, we found that the test datum is CvM = 0.150 whose mental correlate co-ordinates at (0.87, 0.13*) in the test distribution. Now suppose, for argument's sake, that despite knowing that the 1st eruption preceded the 20th eruption by 313 years, for some reason or another only every second one of the 20 individual waiting times are available for the Cramer-Von Mises test. Then the test datum turns out to be CvM = 0.055 whose mental correlate co-ordinates at (0.77, 0.23*) in the test distribution. In both cases, however, we know that there were n = 20 waiting times totalling to 313 years. So in both cases we arrive at precisely the same fiducial intervals as exemplified at (14.3.3). Thus the fiducial argument is not coherently adjoined to its commencement. This is obviously so for *any* fiducial argument. This must be firmly grasped: it is surely the intention that fiducial probabilities such as 0.99, 0.95 and 0.90 must in some or other sense express different levels of 'absolute credibility', but they cannot possibly be doing so inasmuch as they simply do not in any reasonable sense depend on the credibility of the commencement tests that lead to them. It is futile to try to wriggle out of this difficulty by pleading that all the other theories of statistical inference suffer from the same defect. We reply: Indeed they do! So what does that tell us about them?

14.4 The shortcomings of fiducial inference

The shortcomings that have so far been noted, and some further shortcomings, may be summarised as follows:

(1) According to Fisher himself, and as made explicit by the Behrens-Fisher test, fiducial reasoning is metaphysical. That is to say, it cannot coherently be adjoined to the reasoning of the physical sciences.

(2) Fiducial reasoning is not only metaphysical, but is peculiarly so. For instance, Lindley (1958) has shown that it cannot in general be coherently adjoined to Bayesian metaphysics.

(3) Apart from the foregoing two kinds of exogenous incoherence, we have seen above that fiducial reasoning also precipitates endogenous incoherence, because the reasoning it would have us use for elimination tests cannot coherently be adjoined to the reasoning required for the corresponding commencement tests.

(4) Another source of endogenous incoherence is that fiducial reasoning (and this is according to Fisher himself) is applicable to a continuous test statistic only. Thus a discourse that must *also* rely on inferences involving a discrete test statistic, cannot be coherent.

(5) The fiducial rules for discourse are unclear. Buehler and Feddersen (1963) for instance, show that Fisher's favourite example of a fiducial interval (i.e. one based on Student's t for the population mean with normal sampling, as in Fisher 1973, p. 193) fails to satisfy a requirement that Fisher himself laid down as necessary.

(6) Another obscurity in the fiducial rules for discourse is noted by Kendall and Stuart (1961, p. 136) when they point out: 'As to what should be done to construct an interval for a single parameter θ where a single sufficient statistic does not exist, writers on fiducial theory are for the most part silent.' By 'a single' they mean 'a scalar'; consider, for instance, the continuous form of the example at (1.12.1).

(7) As will be found, the rules for calculating fiducial probabilities are, by Fisher's own account, inexorably subject to the notion of simultaneous statistical inference.

All this is bad enough. Yet there is worse to come.

(8) Fisher (1951, p. 60), referring to the example of exponential waiting times, states

> 'if ... there had been 500 accurately measured time intervals, calculations based on the distribution of χ^2 for 1 000 degrees of freedom would show that the probability was 25% of the true value lying below .96957 of the estimate, and 25% of lying above 1.02988, times the same quantity. These values then bracket a central region ... within which the true value will lie with a probability of just one half.' (14.4.1)

Still referring to the example of 1 000 exponential waiting times, he states (on p. 61) that:

> 'When, as in the example chosen, the data are simple and the meaning of the calculations completely clear, other relevant probability statements may be made with equal confidence and exactitude. For example, the probability is 5% each way of the true value lying outside the limiting ratios .92732 and 1.07439, and it is only 1%

> of it lying below .89819 and another 1% of lying above 1.10622, so that the odds are 49 to 1 that it should lie within these last limits. The fiducial distribution in this way comprises a complete set of probability statements appropriate to any chosen level of probability, or to any chosen limits. In such cases the precision of the estimate has been completely specified.' (14.4.2)

This is circular reasoning: consider Investigators 1 and 2, having the same grounds to expect, but not to know for sure, that the waiting times for the first 20 eruptions of Vesuvius after that of AD 1631, might satisfactorily be modelled as a random sample of exponential waiting times indexed by θ for some θ in $0 < \theta < \infty$, where Investigator 1 is given all 20 waiting times, and Investigator 2 is given only that the 20th eruption after that of AD 1631 took place 313 years later. Then a commencement test using the Cramer-Von Mises test tells Investigator 1 that the exponential class characteristic, as thus tested, fits the data satisfactorily, whereas Investigator 2 cannot perform any such commencement test. However, following Fisher, each investigator can calculate that the maximum likelihood estimate is 15.65 years, and each investigator would very much like to have the precision of the estimate 'completely specified' as Fisher promises at (14.4.2). So, consider how each investigator might follow his instructions for obtaining limits for θ^{-1} such that the odds are 19 to 1 that it should lie within those limits. The recipe previously given at (14.3.2) is then more conveniently expressed as

$$\Pr f\left[\theta^{-1} < \frac{2n}{\chi^2_{2n}(P)}\right],$$

where $\chi^2_{2n}(P)$ for $2n = 2(20)$, equals 59.34 when P equals 0.025, and 24.43 when P equals 0.925.

So the requisite 0.95 fiducial interval is by *each one of the two investigators* found to be

$$10.55 \text{ years} < \theta^{-1} < 25.62 \text{ years}.$$

This clearly implies that in respect of a fiducial argument, any commencement test is completely irrelevant, and the reader should carefully note why this is so: for a Cramer-Von Mises test on the 20 waiting times CvM = 0.150 whose mental correlate co-ordinates at (0.87, 0.13*) in the test distribution, from which we are surely justified to conclude:

> The class characteristic, *as tested* by the Cramer-Von Mises test, *fits the given data* well.

Here, the phrases, 'as tested' and (more importantly) 'fits the given data', are warning us that we have not obtained *proof* that the class characteristic is that of exponential sampling. Moreover, this applies not only to the Cramer-Von Mises test; it would also apply to *any* commencement test.

> So the fiducial argument must *assume* that the class characteristic *is* exponential,
> *so that* the probability statement at (14.3.1) can be made,
> *so that* the fiducial argument will produce the probability statement at (14.3.2),
> *so that* probability statements such as the one at (14.3.3) can be made, and
> *that* is circular reasoning.

It might be thought that we can escape this circularity by interpreting fiducial tests as tests of fit. However, that is precluded by the explanation at (14.4.2). It is not possible to justify the pooling of tail areas when testing fit. In any case, it also is not possible to justify a metaphysical number as a measure of physical fit.

14.5 Probability inference

Definition 4.20.1 attracted our attention to a general notion of probability inference, which transcends that of statistical inference. We remarked on it that if the idea of such inference cannot be viable in the latter case, it cannot be viable in the more general case. So, inasmuch as the vicious circle uncovered in Sections 4.8, 12.7, 13.12 and 14.4 obviously extends to any form of statistical inference, for the simple reason that statistical inference cannot apply to the outcome of any commencement test, and inasmuch as such tests are inescapably needed to provide the platform from which any form of statistical inference must try to proceed, we have provided proof beyond any reasonable contest that the notion of statistical inference is not viable – it is a mere will-o'-the-wisp. It follows that the more general notion of probability inference is not viable. As a matter of fact, the wider literature on probability inference indicates that it is indeed a will-o'-the-wisp. Thus for instance, two attempts at a general theory of probability inference by intellectual giants such as John Maynard Keynes and Rudolf Carnap, respectively, evidently turned out to be fruitless (Keynes 1921; Carnap 1950). Nowhere in the vast body of present-day scientific knowledge do we find that their theories made any contribution whatsoever to that knowledge.

Chapter 15

EPILOGUE

Challenging the statistical profession

This book originated in an attempt to resolve the dreadful confusion surrounding simultaneous statistical inference. The reader might, or might not, agree with the outcome. Either way, however, one cannot simply laugh the matter off. Our obligation to our customers in the substantive sciences, and to consulting statisticians at the interface, disallows that. All of us, as professionals, are responsible for it. If the manner of the present development seems overly aggressive, even arrogant, I would plead for some understanding, on three grounds. Firstly, after a long time of trying to engage fellow statisticians on this matter, it became clear that an immense prejudice would have to be overcome. Notions about probabilities of drawing this, that, or the other wrong conclusion, have become so deeply entrenched that asking present-day statisticians to dispense with these ideas is akin to trying to row a boat up the Victoria falls. Secondly, whilst it is true that science cannot advance by way of the method of authority, it is a fact of life that much of what passes for 'science', is simply inspired by the views of this, that, or another authoritative person, and so the need to attack such authority seemed unavoidable. Thirdly, and most importantly, it is demonstrable that our notions on statistical inference and the inevitable correlative notions on simultaneous statistical inference do dreadful damage to substantive science; for proof, one need only to scan the journals of such science, where one finds that substantive investigators have been persuaded to consider various probabilities of dubious relevance, instead of considering what appropriate analyses of their data would show. So clearly, for all three reasons, an attempt at subtlety would have courted failure to communicate. In order to compel the whole of the statistical community to deal with this problem, it needs to be attacked with a broad sword. And we must draw in the rank and file, because debates on the ideas of statistical inference have for too long been a province of the ivory tower. It is time to drag these nonsensical ideas into the market place, that is to say, into where they can be tested at the interface between statistics and the substantive sciences. For that, a suitable point of departure would for instance be the problem formulated at (4.20.4) in Example 4.20.1, in respect of which we can test each of the various received theories of statistical inference in turn, as to how that problem is to be solved, such that we can claim to have done so:

> without answers that mistake the pursuit of knowledge for the use of knowledge,
> without answers that confound data patterns with contrary data patterns,
> without answers that confound relevant data patterns with irrelevant data patterns,
> without answers that confound physics with metaphysics,
> without answers that rely on circular reasoning,
> without answers that partake of a retrospectively invented knowing subject.

References

Agresti, A. 2002. *Categorical data analysis*. Second Edition. Hoboken, NJ: Wiley.

Anscombe, F.J. 1963. Tests of goodness of fit. *J. Roy. Stat. Soc.,* Series B, 25:81-94.

Arbuthnott, J. 1710. An argument for divine providence taken from the constant regularity of the births of both sexes. *Phil. Trans. Roy. Soc.,* 27:186-90.

Bain, L.J. and Engelhardt, M. 1989. *Introduction to probability and mathematical statistics*. Boston: PWS-kent.

Barnard, G.A. and Sprott, D.A. 1970. A note on Basu's examples of anomalous ancillary statistics. *Waterloo Symposium on Foundations of Statistical Inference*. Toronto: Holt, Rhinehart & Winston.

Bartlett, M.S. 1937. Some examples of statistical methods of research in agriculture and applied biology. *Suppl. J. Roy. Stat. Soc.,* 4:137-183.

Basu, D. 1964. Recovery of ancillary information. *Sankhyā* 26, Series A:3-16.

Basu, D. 1975. Statistical information and likelihood (with discussion). *Sankhyā* 37, Series A, Pt1:1-71, and *corrigendum* p. 456.

Bayes, T. 1763. An essay toward solving a problem in the doctrine of chances. *Phil. Trans. Roy. Soc*., 53:370-418.

Berkson, J. 1938. Some difficulties of interpretation encountered in the application of the chi-square test. *J. Amer. Stat. Assoc.,* 33:526-536.

Berkson, J. 1942. Tests of significance considered as evidence. *J. Amer. Stat. Assoc.,* 37: 325-335.

Birnbaum, A. 1962. On the foundations of statistical inference (with discussion). *J. Amer. Stat. Assoc.,* 57: 269-326.

Bliss, C.I. 1967. *Statistics in biology*. New York: McGraw-Hill.

Boardman, T.J. and Moffit, D.R. 1971. Graphical Monte Carlo Type I error rates for multiple comparison procedures. *Biometrics,* 27:738-744.

Boussy, I.A. and Kidwell, M.G. 1987. The P-M hybrid dysgenesis cline in eastern Australian *Drosophila melanogaster*. Discrete P, Q, and M regions are nearly contiguous. *Genetics*, 115:737-745.

Box, G.E.P. 1953. Non-normality and tests on variances. *Biometrika*, 40:318-335.

Brownlee, K.A. 1949. *Industrial experimentation.* Fourth Edition. London: Her Majesty's Stationary Office.

Bryan-Jones, J. and Finney, D.J. 1983. On an error in 'Instructions to authors'. *HortScience,* 18(3):279-281.

Buehler, R.J. and Feddersen, A.P. 1963. Note on a conditional property of Student's *t*. *Ann. Math. Stat.*, 34:1098-1100.

Campbell, F.L. 1926. Speed of toxic action of arsenic in the silkworm. *J. Gen. Physiol.,* 9:433-443.

Carmer, S.G. and Swanson, M.R. 1973. Evaluation of ten pairwise comparison procedures by Monte Carlo methods. *J. Amer. Stat. Assoc.*, 68:66-74.

Carnap, R. 1950. *Logical foundations of probability*. Second Edition. Chicago: University of Chicago Press.

Chew, V. 1976. Comparing treatment means: a compendium. *Hortscience,* 11:348-357.

Cochran, W.G. and Cox, G.M. 1957. *Experimental design*. Second Edition. New York: Wiley.

Cox, D.R. 1970. *The analysis of binary data*. London: Methuen.

Cox, D.R. 1971. The choice between alternative ancillary statistics. *J. Roy. Stat. Soc.,* Series B, 33:251–255.

Cox, D.R. 1977. The role of significance tests. *Scand. J. Stat.,* 4:49-70.

Cox, D.R. 1984. Contribution to discussion of paper by Yates. *J. Roy. Stat. Soc.*, Series A, 147:426-463.

Cox, D.R. and Hinkley, D.V. 1974. *Theoretical statistics*. London: Chapman and Hall.

D'Agostino, R.B. 1970. Transformation to normality of the null distribution of g_1. *Biometrika,* 57:679-681.

D'Agostino, R.B. and Tietjen, G.L. 1971. Simulation probability points of b_2 for small samples. *Biometrika,* 58:669-672.

Daniel, C. and Wood, F.S. 1971. *Fitting equations to data*. New York: Wiley.

Dawkins, R. 2003. *A devil's chaplain*. London: Weidenfield and Nicolson.

Dixon, W.J. 1950. Analysis of extreme values. *Ann. Math. Stat.,* 21:488-506.

Dixon, W.J. 1951. Ratios involving extreme values. *Ann. Math. Stat.,* 22:68-78.

Dobzhansky, Th. 1951. *Genetics and the origin of species*. Third Edition. New York: Columbia University Press.

Duncan, D.B. 1955. Multiple range and multiple *F* tests. *Biometrics,* 11:1-42.

Dunnett, C.W. 1955. A multiple comparison procedure for comparing several treatments with a control. *J. Amer. Stat. Assoc.,* 50:1096-1121.

Edwards, A.W.F. 1972. *Likelihood*. Cambridge: Cambridge University Press.

Edwards, J.H. 1961. Seasonal incidence of congenital disease in Birmingham. *Ann. Hum. Gen.,* 25:89-93.

Einot, I. and Gabriel, K.R. 1975. A study of the powers of several methods of multiple comparisons. *J. Amer. Stat. Assoc.,* 70:574-583.

Feller, W. 1970. *An introduction to probability theory and its applications. Volume I.* Third Edition. New York: Wiley.

Finney, D.J. 1946. Standard errors of yields adjusted for regression on an independent measurement. *Biom. Bull.,* 2:53-55.

Finney, D.J. 1948. The Fisher-Yates test of significance in 2×2 contingency tables. *Biometrika,* 35:145-156.

Fisher, R.A. 1926. The arrangement of field experiments. *J. Ministry Agric.,* XXXIII:503-513.

Fisher, R.A. 1934. Randomisation, and an old enigma of card play. *The Mathematical Gazette,* 18:294-297.

Fisher, R.A. 1935. The logic of inductive inference. *J. Roy. Stat. Soc.,* Series A, 98:39-54.

Fisher, R.A. 1936. Uncertain inference. *Proc. Amer. Acad. Arts and Sciences,* 71(4):245-258.

Fisher, R.A. 1943. Note on Dr. Berkson's criticism of tests of significance. *J. Amer. Stat. Assoc.,* 38:103-104.

Fisher, R.A. 1950. The significance of deviations from expectation in a Poisson series. *Biometrics,* 6:17-24.

Fisher, R.A. 1953. Dispersion on a sphere. *Proc. Roy. Soc.,* Series A, 217: 295-305.

Fisher, R.A. 1966. *The design of experiments*. Eighth Edition. Edinburgh: Oliver and Boyd.

Fisher, R.A. 1970. *Statistical methods for research workers*. Fourteenth Edition. Edinburgh: Oliver and Boyd.

Fisher, R.A. 1973. *Statistical methods and scientific inference*. Third Edition. London: Oliver and Boyd.

Ford, E.B. 1937. Problems of heridity in the Lepidoptera. *Biological Reviews*, 12:461-503.

Ford, E.B. 1945. Polimorphism. *Biological Reviews*, 20:73-88.

Fraenkel, G. 1927. Phototropotaxis bei Meerestieren. *Naturwissenschaften,* 15:117-122.

Fraser, D.A.S. 1968. *The structure of inference*. New York: Wiley.

Freund, J.E. and Perles, B.M. 1993. Observations on the definition of P-values. *Teaching Statistics,* 15:8-9.

Gates, C.E. and Bilbro, J.D. 1978. Illustration of a cluster analysis method for mean separation. *Agron. J.,* 70:462-465.

Gibbons, J.D. 1986. P-value. In S. Kotz and N.L. Johnson (eds), *Encyclopedia of statistical sciences*. *Volume* 7. New York: Wiley, pp. 366-368.

Gibbons, J.D. and Pratt, J.W. 1975. P-values: interpretation and methodology. *Amer. Stat.*, 20:20-25.

Gibbons, J.D., Olkin, I. and Sobel, M. 1977. *Selecting and ordering populations: a new statistical methodology.* New York: Wiley.

Goulden, C.H. 1939. *Methods of statistical analysis*. New York: Wiley.

Gupta, S.S. 1965. On some multiple decision (selection and ranking) rules. *Technometrics,* 7:225-245.

Gupta, S.S. and Panchapakesan, S. 1979. *Multiple decision procedures: theory and methodology of selecting and ranking populations.* New York: Wiley.

Hald, A. 1952. *Statistical theory with engineering applications*. New York: Wiley.

Harter, H.L. 1960. Critical values for Duncan's new multiple-range test. *Biometrics,* 16: 671-685.

Harter, H.L. 1961. Expected values of normal order statistics. *Biometrika,* 48:151-165.

Hartley, H.O. 1955. Some recent developments in analysis of variance. *Comm. Pure & Appl. Math.,* 8:47-72.

Hill, I.D. 1984. Contribution to discussion of paper by Yates. *J. Roy. Stat. Soc.*, Series A, 147:426-463.

Hill, I.D. and Pike, M.C. 1965. Algorithm 4: TWOBYTWO. *Computer Bulletin,* 9:56-63.

Hinkelmann, K. and Kempthorne, O. 2007. *The design and analysis of experiments*. Second Edition. New York: Wiley.

Hogg, R.V. and Craig, A.T. 1970. *Introduction to mathematical statistics.* New York: Macmillan.

Hooper, J. 2002. *Of moths and men.* London: Fourth Estate.

Irwin, J.O. 1935. Tests of significance for differences between percentages based on small numbers. *Metron,* 12:83-94.

Jeffreys, H. 1961. *Theory of probability*. Third Edition. Oxford: Clarendon Press.

Jolliffe, I.T. 1971. Contribution to discussion of paper by R.T. O'Neill and B.G. Wetherill. *J. Roy. Stat. Soc.*, Series B, 33:218-250.

Kalbfleisch, J.G. 1979. *Probability and statistical inference.* Volume 2. Second Edition. New York: Springer-Verlag.

Kempthorne, O. and Folks, L. 1971. *Probability, statistics, and data analysis.* Ames, Iowa: Iowa State University Press.

Kendall, M.G. and Stuart, A. 1961. *The advanced theory of statistics*. *Volume* 2. Third Edition. London: Charles Griffin.

Keynes, J.M. 1921. *A treatise on probabilities*. New York: Harper and Row.

Kolmogoroff, A. 1956. *Foundations of the theory of probability.* Second English Edition. New York: Chelsea.

Krumbein, W.C. 1939. Preferred orientation of pebbles in sedimentary deposits. *J. Geology,* 47:673-706.

Laplace, Pierre Simon de. 1814. *A philosophical essay on probabilities*. New York: Dover.

Leakey, R. and Lewin, R. 1993. *Origins reconsidered.* London: Abacus.

Lehmann, E.L. 1959. *Testing statistical hypotheses*. New York: Wiley.

Lehmann, E.L. 1986. *Testing statistical hypotheses*. Second Edition. New York: Wiley.

Lehmann, E.L. and Scheffé, H. 1950. Completeness, similar regions, and unbiased estimation. Part I. *Sankhyā* 10:305-340.

Lehmann, E.L. and Scheffé, H. 1955. Completeness, similar regions, and unbiased estimation. Part II. *Sankhyā* 15:219-236.

Lindley, D.V. 1958. Fiducial distributions and Bayes' theorem. *J. Roy. Stat. Soc.,* Series B, 20: 102-107.

Lindley, D.V. 1965. *Introduction to probability and statistics from a Bayesian viewpoint*. Cambridge: Cambridge University Press.

Lindley, D.V. 1971. Contribution to discussion of paper by R.T. O'Neill and B.G. Wetherill. *J. Roy. Stat. Soc.*, Series B, 33:218-250.

Ling, R.F. 1992. Just say no to binomial (and other discrete distributions) tables. *Amer. Stat.,* 46:53-54.

Lowenfeld, J. 1955. *An experiment relating the concepts of repression, subception, and perceptual defence.* Unpublished PhD dissertation, Pennsylvania State University.

Maclean, G.L. 1985. *Roberts' Birds of Southern Africa.* Cape Town: Trustees of the John Voelker Bird Book Fund.

Maritz, J.S. 1989. *Empirical Bayes' methods*. Second Edition. London: Chapman and Hall.

May, J.M. 1952. Extended and corrected tables of the upper percentage points of the 'Studentized' range. *Biometrika,* 39:192-193.

Maynard Smith, J. 1972. *On evolution.* Edinburgh: Edinburgh University Press.

Mead, R. and Curnow, R.N. 1983. *Statistical methods in agriculture and experimental biology*. London: Chapman and Hall.

Mead, R. 1990. *The design of experiments: statistical principles for practical applications*. Cambridge: Cambridge University Press.

Meulepas, E. 1998. A two-tailed P-value for Fisher's exact test. *Biom. J.,* 40:3-10.

Miller, R.G. Jr. 1981. *Simultaneous statistical inference.* Second Edition. New York: Springer.

Milliken, G.A. and Graybill, F.A. 1970. Extensions of the general linear hypothesis model. *J. Amer. Stat. Assoc.,* 65:797-807.

Milliken, G.A. and Graybill, F.A. 1971. Tests for interaction in the two-way model with missing data. *Biometrics,* 27:1079-1083.

Mood A.M., Graybill, F.A. and Boes, D.C. 1974. *Introduction to the theory of statistics.* Third Edition. New York: McGraw-Hill.

Nelder, J.A., 1971. Contribution to discussion of paper by R.T. O'Neill and B.G. Wetherill. *J. Roy. Stat. Soc.*, Series B, 33: 218-250.

Neyman, J. 1941. Fiducial argument and the theory of confidence intervals. *Biometrika* 32:120-150.

Neyman, J. 1952. *Lectures and conferences on mathematical statistics and probability*. Washington D.C. Graduate School, U.S. Dept. of Agriculture.

Neyman, J. and Pearson, E.S. 1933. On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. Roy. Soc.,* (A)231:289-337.

O'Neill, R.T. and Wetherill, B.G. 1971. The present state of multiple comparison methods (with discussion). *J. Roy. Stat. Soc.,* Series B, 33:218-250.

Ott, R.L. 1992. *An introduction to statistical methods and data analysis*. Fourth Edition. Belmont, CA: Duxbury Press.

Pearson, K. 1900. On a criterion that a given set of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen in random sampling. *Phil. Mag.,* 50:157-76.

Plackett, R.L. 1971. Contribution to discussion of paper by R.T. O'Neill and B.G. Wetherill. *J. Roy. Stat. Soc.*, Series B, 33:218-250.

Popper, K.R. 1979. *Objective knowledge; an evolutionary approach*. Revised Edition. Oxford: Oxford University Press.

Pratt, J.W. 1962. Contribution to discussion of paper by A. Birnbaum. *J. Amer. Stat. Assoc.*, 57: 269-326.

Raleigh. 1880. On the result of a large number of vibrations of the same pitch and arbitrary phase. *Phil. Mag.,* 10:73-78.

Randall, J.H. 1985. Controlled linear calibration. Unpublished Ph.D. dissertation, University of Stellenbosch, Stellenbosch, Republic of South Africa.

Reichenbach, H. 1949. *The theory of probability*. Berkley, CA: University of California Press.

Rosenthall, E.B. 1965. *Understanding the new maths*. London: Souvernir Press.

Sadie, A. 1996. The robustness of a subset selection procedure in the case of a non-orthogonal analysis of variance. Unpublished M.Sc. dissertation, University of Stellenbosch, Stellenbosch, Republic of South Africa.

SAS. 1992. SAS Institute Inc., SAS® Technical Report P-229, SAS/STAT®, Software, Changes and Enhancements, release 6.07. Cary, NC: SAS Institute Inc.

Saunders, A.R. and Rayner, A.A. 1951. *Statistical methods with special reference to field experiments*. Third Edition. Pretoria: Government Printer.

Savage, L.J. 1954. *The foundations of statistics*. New York: Wiley.

Savage, L.J. 1962. *The foundations of statistical inference*. London: Methuen.

Scheffé, H. 1953. A method for judging all contrasts in the analysis of variance. *Biometrika,* 40:87-104.

Scheffé, H. 1959. *The analysis of variance*. New York: Wiley.

Scott, A.J. and Knott, M. 1974. A cluster analysis method for grouping means in the analysis of variance. *Biometrika,* 40:87-104.

Shapiro, S.S. and Wilk, M.B. 1965. An analysis of variance test for normality (complete samples). *Biometrika,* 52:591-611.

Shilts, R. 1987. *And the band played on*. New York: St. Martin's Griffin.

Snedecor, G.W. 1956. *Statistical methods*. Fifth Edition. Ames, IA: Iowa State University Press.

Snedecor, G.W. and Cochran, W.G. 1989. *Statistical methods*. Eighth Edition. Ames, IA: Iowa State University Press.

Steel, R.G.D. and Torrie, J.H. 1980. *Principles and procedures of statistics*. Second Edition. New York: McGraw-Hill.

Stone, M. 1969. The role of significance testing; some data with a message. *Biometrika,* 56:485-493.

Student, 1908. The probable error of the mean. *Biometrika,* 6:1-25.

Student, 1927. Errors of routine analysis. *Biometrika,* 19:151-164.

Tallis, G.M. 1988. Goodness of fit. In S. Kottz and N.L. Johnson (eds), *Encyclopedia of statistical science,* Volume 3. New York: Wiley, pp. 451-461.

Thomas, D.A.H. 1974. Error rates in multiple comparisons among means – results of a simulation exercise. *J. Roy. Stat. Soc.,* Series C, 23:284-294.

Tocher, K.D. 1950. Extension of the Neyman-Pearson theory of tests to discontinuous variates. *Biometrika,* 37:139-144.

Tukey, J.W. 1949a. One degree of freedom for non-additivity. *Biometrics,* 5:232-242.

Tukey, J.W. 1949b. Comparing individual means in the analysis of variance. *Biometrics,* 5:99-114.

Tukey, J.W. 1953. *The problem of multiple comparisons.* Unpublished notes. Princeton University, Princeton, New Jersey.

Van Aarde, I.M.R. 1994. Pivotal quantities for the substitution values of various entries, each in place of its best contender, given the standard analysis of variance model. *Biom. J.,* 36:673-687.

Wackerly, D.D., Mendenhall, W. and Scheaffer, R.L. 1996. *Mathematical statistics with applications*. Fifth Edition. Belmont, CA: Wordsworth.

Waller, R.A. and Duncan, D.B. 1969. A Bayes rule for the symmetric multiple comparisons problem. *J. Amer. Stat. Assoc.,* 64:1484-1503.

Welch, B.L. 1939. On confidence limits and sufficiency, with particular reference to parameters of location. *Ann. Math. Stat.,* 27:58-69.

Wells, M.J. 1961. Weight discrimination by Octopus. *J. Exp. Biol.,* 38:127-133.

Wilson, J.W. 1975. *Vegetable gardening*. Menlo Park, CA: Lane.

Wilkinson, B. 1933. A statistical consideration in psychological research. *Psychol. Bull.,* 48:156-158.

Yates, F. 1984. Tests of significance for 2×2 contingency tables (with discussion). *J. Roy. Stat. Soc.,* Series A, 147:426-463.

Young, A. 1771. *A course of experimental agriculture, Volume I.* Dublin: Exshaw.

Young, J.Z. 1965. The organization of a memory system. *Proc. Roy. Soc.* B, 163:285-320.

In this book four different schools of inference, namely, frequentist, Bayesian, likelihood and fiducial are introduced and then with appropriate examples and discussions, the gaps and holes in the traditional sense of inferences are exposed. The author goes back deep into the literature, exploring views and fallacies introduced by outstanding scientists and statisticians. The author proposes a coordination test of any model posted for a given (observed) data set. A coordination test is a calculable ordered number triplet, whose three members are statistical coordinates. Various appropriate examples are given in each chapter where the traditional analysis and inference are applied, and then the statistical coordinates are calculated for each example. After each example the shortfalls, gaps and holes in the traditional methods are discussed.

Prof Christien Thiart

University of Cape Town

This is a bold book. It asks statisticians to revise deeply entrenched ways of thinking about not only the practice of their craft, but also about the philosophical basis of their subject. It does so by an exhaustive analysis, using a host of examples to demonstrate specific points.

Prof Jannie Hofmeyr

Stellenbosch University

SUN PRESS

www.africansunmedia.co.za
www.sun-e-shop.co.za