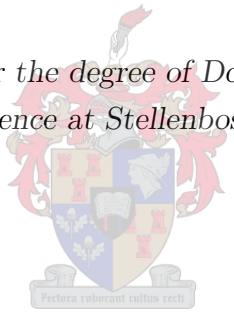
From stable priors to maximum Bayesian evidence via a generalised rule of succession

by

Michiel Burger de Kock

Dissertation presented for the degree of Doctor of Philosophy in the Faculty of Science at Stellenbosch University



Merensky Building, Merriman Ave.
Stellenbosch University
Private Bag X1, Matieland, 7602, South Africa.

Promoter: Prof. H.C. Eggers

April 2014

Declaration

By submitting this dissertation electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: April 2014 .....

Abstract

From stable priors to maximum Bayesian evidence via a generalised rule of succession

M.B. de Kock

Merensky Building, Merriman Ave.
Stellenbosch University
Private Bag X1, Matieland, 7602, South Africa.

Dissertation: PhD (Theoretical Physics)

April 2014

We investigate the procedure of assigning probabilities to logical statements. The simplest case is that of equilibrium statistical mechanics and its fundamental assumption of equally likely states. Rederiving the formulation led us to question the assumption of logical independence inherent to the construction and specifically its inability to update probability when data becomes available. Consequently we replace the assumption of logical independence with De Finetti's concept of exchangeability. To use the corresponding representation theorems of De Finetti requires us to assign prior distributions for some general parameter spaces. We propose the use of stability properties to identify suitable prior distributions. The combination of exchangeable likelihoods and corresponding prior distributions results in more general evidence distribution assignments. These new evidence assignments generalise the Shannon entropy to other entropy measures. The goal of these entropy formulations is to provide a general framework for constructing models.

Uittreksel

Van stabiele a priori-verdelings tot maksimum Bayes-getuienis via 'n veralgemeende Laplace-opvolgwet

M.B. de Kock

Merenskygebou, Merrimanlaan
Stellenbosch Universiteit
Privaatsak X1, Matieland, 7602, Suid-Afrika.

Proefskrif: PhD (Teoretiese Fisika)

April 2014

Ons ondersoek the prosedure om waarskynlikhede aan logiese stellings toe te ken. Die eenvoudigste geval is die van ewewig-statistiese meganika en die ooreenkomstige fundamentele aanname van ewekansige toestande. Herafleiding van die standaard formulering lei ons tot die bevraagtekening van die aanname van logiese onafhanklikheid en spesifiek die onmoontlikheid van opdatering van waarskynlikheid wanneer data beskikbaar raak. Gevolglik vervang ons die aanname van logiese onafhanklikheid met De Finetti se aanname van omruilbaarheid. Om die ooreenkomstige voorstelling stellings te gebruik moet ons a priori verdelings konstrueer vir 'n paar algemene parameter-ruimtes. Ons stel voor dat stabiliteits-eienskappe gebruik moet word om geskikte a priori distribusies te identifiseer. Die kombinase van omruilbare aanneemlikheids funksies en die ooreenkomstige a priori verdelings lei ons tot nuwe toekennings van getuienis-verdelings. Hierdie nuwe getuienes-verdelings is n veralgemening van Shannon se entropie na ander entropie-maatstawwe. Die doel van hierdie entropie formalismes is om 'n raamwerk vir modelkonstruksie te verskaf.

Acknowledgements

This dissertation is the result of many hours of work and was a large part of my life for some time. Many people contributed to this large (never-ending) project.

First I would like to thank my supervisor, Prof. Hans Eggers, for the discussions, ideas, mentoring and especially for the enthusiasm Hans shows for understanding and getting things right. He also deserves credit for sending a reluctant student to many conferences and editing my funny sentences. The Friday Seminar group was also source of many entertaining and stimulating discussions, thanks to Dr. Hannes Kriel, Dr. Carl Rohwer and Stefan Astl.

I would also like to thank my parents Lizette and Andre de Kock for their unfailing love and support for this endeavour. We were equally ignorant about the length and breadth of this project when we started out but happily we all made it to the end.

And finally special thanks to the NRF and NITheP for providing financial support for this dissertation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

As most students in physics, my education regarding statistics was based on the view that probability is the number of data points with a particular outcome b, divided by the total number. While this demonstrably works in many cases, the difficulties for a small number of measurements and/or many possible outcomes (such as highly differential measurement quantities) eventually cannot be ignored. In my masters thesis de Kock (2009), such difficulties occurred: we tried to construct a systematic expansion of nongaussian probability distributions, and following textbook literature we tried to apply the Gram-Charlier and Edgeworth expansions using different reference functions. While these do in fact work under some circumstances, as e.g. shown in de Kock *et al.* (2011*b*) and de Kock *et al.* (2011*c*), the fundamentally asymptotic character of such expansions led us to question their foundations and in turn led us to Bayesian statistics.

This dissertation does not answer the original questions raised by such expansions. Rather, the journey into the Bayesian world has proven to be a long and unsettling reorganisation of many concepts and relations which are commonly taken for granted. Many ideas commonly accepted, such as the above idea of probability as data ratio, were found to be correct in some limit but only in that limit. Others needed to be placed in different relations or reinterpreted. In time and with continued digging, the original goal turned out to be less interesting than the reasons why such simple concepts fail and how the strict thinking and careful accounting required by Bayesian reasoning resolves those failures.

We will not therefore concern ourselves mainly with correlations as originally intended, nor with high energy physics, nor even with classical statistical mechanics, even though it may look that way. The dissertation is about two things: firstly, the discovery that a simple assumption underlying many calculations in physics, statistics and probably many other fields may only be a special case, and that the general case is as yet not fully understood, that "Independence" is a special case of "Exchangeability". Secondly, the thrust and contents of this dissertation are about bringing the ideas and mathematics of Bayesian analysis to bear on this change in foundations.

The rigour and consistency of Bayes has resulted in many new insights. Especially satisfying was that parameter estimation and choosing a model are two sides of the same

coin and use the same procedure. It is also gratifying to discover that statistical mechanics and data analysis are also part of that same framework. In many cases this does not contradict the orthodox methods but explains (at least to us) what they are really doing. Also we could clarify the logical reasoning used in statistics.

Our immediate goal in this dissertation is to separate issues relating to physical objects and the laws governing them from that which is statistics, understood as inductive logic. This makes it easier to generalise results. More importantly, it gives us a different perspective on what is important and what is trivial.

One of the unexpected fruits of this endeavour is the Bayes Factor which shows which model is preferred by the data. This is an exciting development for us since it might provide a natural way of truncating series expansions. Among other things, we will also extend the principle of maximum entropy in various directions but will not consider it as a principle per se. It does not have the status of the sum and product rule which are far more fundamental; in fact, we consider Maximum Entropy only a limited type of model, namely a multinomial distribution in disguise. Correlations will appear sporadically in different places in the dissertation because they do remain one of our long-term goals. For the moment, we are concerned with sorting out the fundamentals and doing so as well as possible.

Chapter 2

Preliminaries

In this chapter, we provide some mathematical background information as a basis for later use. This is only a reasonable overview of the topics; for a more comprehensive review see Johnson *et al.* (2005) for the binomial and factorial section, Whittaker and Watson (1927) for asymptotic series, Cover and Thomas (2012) for convexity, Khinchin (1949) for generating functions and Bleistein and Handelsman (1986) for saddlepoint approximations. We shall try to use the following consistent notation:

Entry	**Symbol** S	**Total** T	**Ratio** (S/T)
Logical Proposition	$\mathcal{A}, \mathcal{B}, \ldots, \mathcal{Z}$		
Real number	α, β, γ		
Positive integers	a, r, n	A, R, N	ρ, ρ'
Dual space	t, μ, λ		
Sum Index	j, k, l, ℓ		
Real Part	$\Re[\ldots]$		
Small number or error	ϵ		
Probability of ...	$p(...)$		
Bin index	b, c, d	B	

2.1 Factorials and binomials

The Euler definition of a gamma function is

$$\Gamma[\alpha] = \int_0^\infty t^{\alpha-1} e^{-t}\, \mathrm{d}t, \qquad \Re[\alpha] > 0 \tag{2.1.1}$$

and if we integrate by parts we find a recursion relation for the gamma function,

$$\Gamma[\alpha + 1] = \alpha \Gamma[\alpha], \tag{2.1.2}$$

which allows us to extend the definition of the gamma function to negative and noninteger α. If α is a positive integer n the gamma function is a factorial

$$(n-1)! = \Gamma[n], \tag{2.1.3}$$

where

$$n! = n(n-1)(n-2)\cdots 1 = \prod_{j=0}^{n-1}(n-j). \tag{2.1.4}$$

The factorial is very useful because it is the number of different orderings of n elements. Saying that we have an ordered sequence or unordered sequence is different from considering the elements of a set distinguishable or undistinguishable. An ordered or unordered sequence is a matter of notation and our intent of keeping the information in the ordering. There are n different choices for the first element, $(n-1)$ choices for the second element and so forth giving us the total number of permutations of n elements. We shall also need some asymptotic expansions of the gamma and factorial functions, the first being Stirling's expansion,

$$\Gamma[\alpha+1] \sim \sqrt{2\pi}\,\alpha^{\alpha+1/2}e^{-\alpha}\exp\left(\frac{1}{12\alpha} - \frac{1}{360\alpha^3} + \cdots\right). \tag{2.1.5}$$

While the second which has no name but has a similar accuracy,

$$\frac{\Gamma[\alpha+j]}{\Gamma[\alpha+k]} \sim \alpha^{(j-k)}\left(1 + \frac{(j-k)(j+k-1)}{2\alpha} + \cdots\right). \tag{2.1.6}$$

It is useful to count how many times an element α occurs in a sequence $\mathcal{A}$. So, defining an indicator function as,

$$\delta(\alpha, \mathcal{A}) = \begin{cases} 1 & \text{if } \alpha \in \mathcal{A} \\ 0 & \text{if } \alpha \notin \mathcal{A}, \end{cases} \tag{2.1.7}$$

which is used to define an *occupation number* that counts the number of times, that $\alpha \in \mathcal{A}$ in N occurrences

$$n_{\mathcal{A}} = \sum_{j=1}^{N} \delta(\alpha_j, \mathcal{A}). \tag{2.1.8}$$

We next ask how many ways there are of choosing k ordered elements out of a n sized set. For example, from the set $\{a, b, c, d\}$ we can choose two elements in twelve ways,

$$\begin{array}{llllll} \{a,b\}, & \{a,c\}, & \{a,d\}, & \{b,c\}, & \{b,d\}, & \{c,d\} \\ \{b,a\}, & \{c,a\}, & \{d,a\}, & \{c,b\}, & \{d,b\}, & \{d,c\}. \end{array} \tag{2.1.9}$$

Assuming ordering is important, there are n choices for the first element, $n-1$ choices for the second element and $n-k+1$ choices for the kth element. This gives us a *falling factorial*

$$n^{\underline{k}} = n(n-1)\cdots(n-k+1) = \prod_{j=0}^{k-1}(n-j) = \frac{n!}{(n-k)!} \tag{2.1.10}$$

where we follow the notation from Graham *et al.* (1994). If we want the number of ways of choosing k unordered elements out of n we divide with the number of permutations $k!$ to find the *binomial coefficient*

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n^{\underline{k}}}{k!} = \binom{n}{n-k}. \tag{2.1.11}$$

In our example we get the six elements $\{a,b\}\ldots\{c,d\}$, $\binom{4}{2} = 6$ and we read the binomial coefficient as 4 choose 2 is equal to 6. When we expand the product $(\alpha+\beta)^n$ every term is the product of n factors, each either an α or β. The number of terms with k factors of α and $n-k$ factors of β is the binomial coefficient; yielding the binomial theorem

$$(\alpha+\beta)^n = \sum_{k=0}^{\infty} \binom{n}{k} \alpha^k \beta^{n-k}, \tag{2.1.12}$$

for a positive integer power n and real numbers α and β. Notice that it is a finite sum since $\binom{n}{k} = 0$ if $k < 0$ or $k > n$. The equivalent of the binomial theorem for multiple variables, is called the *multinomial theorem*

$$\left(\sum_{b=1}^{B} \alpha_b\right)^N = \sum_{U[\boldsymbol{n}]}^{N} N! \prod_{b=1}^{B} \frac{a_b^{n_b}}{n_b!}, \tag{2.1.13}$$

where $\boldsymbol{n} = \{n_1, \ldots, n_B\}$ and

> `Universal set` $U[\boldsymbol{n}]$`:`
> `The set of all non-negative integers` $n_1, n_2, \ldots n_B$ `that sum to` N`.`

A Taylor expansion of (2.1.12) with $\beta = 1$

$$(1+\alpha)^\gamma = \sum_{k=0}^{\infty} \frac{\gamma^{\underline{k}}}{k!} \alpha^k, \qquad -1 < \alpha < 1, \tag{2.1.14}$$

may be used to generalise the binomial coefficient to arbitrary real γ,

$$\binom{\gamma}{k} = \frac{\gamma^{\underline{k}}}{k!}. \tag{2.1.15}$$

The Taylor expansion remains valid for negative values of γ implying that negative binomial coefficients and also negative falling factorials can be defined. A negative falling factorial we will write in terms of a *rising factorial*

$$n^{\overline{k}} = n(n+1)\cdots(n+k-1) = \prod_{j=0}^{k-1}(n+j) = \frac{\Gamma[n+k]}{\Gamma[n]}, \tag{2.1.16}$$

by using

$$(-n)^{\underline{k}} = (-n)(-n-1)\ldots(-n-k+1) = (-1)^k n^{\overline{k}}. \tag{2.1.17}$$

A rising factorial $n^{\overline{k}}$ is interpreted combinatorically in terms of partitions into sets, and specifically as the number of ways to insert k ordered elements into n ordered sets. For

example, let there be k=3 elements a, b and c which are to be inserted into n=2 ordered sets, which in the table below are separated by the vertical line |. In the first step of constructing all such possibilities, a is inserted into either Set 1 or Set 2. In the second step, b is inserted into each of the three possible positions with regard to existing elements a and |. Following the third step, we have $2^{\overline{3}} = 2(2{+}1)(2{+}2) = 24$ possible orderings:

Step 1			$a\|$	$\|a$		
Step 2	$ba\|$	$ab\|$	$a\|b$	$b\|a$	$\|ba$	$\|ab$
Step 3	$cba\|$	$cab\|$	$ca\|b$	$cb\|a$	$c\|ba$	$c\|ab$
	$bca\|$	$acb\|$	$ac\|b$	$bc\|a$	$\|cba$	$\|cab$
	$bac\|$	$abc\|$	$a\|cb$	$b\|ca$	$\|bca$	$\|acb$
	$ba\|c$	$ab\|c$	$a\|bc$	$b\|ac$	$\|bac$	$\|abc$

Based on this, the *negative binomial coefficient*

$$\binom{-n}{k} = (-1)^k \frac{n^{\overline{k}}}{k!} = (-1)^k \binom{n+k-1}{k} \tag{2.1.18}$$

is interpreted as $(-1)^k$ times the number of ways k *unordered* elements can be inserted into n ordered sets. In terms of the above example, making a=b=c yields the $2^{\overline{3}}/3! = 4$ distinct enumerations $aaa|$, $aa|a$, $a|aa$ and $|aaa$ or, in terms of occupation numbers, $\{3,0\}$, $\{2,1\}$, $\{1,2\}$ and $\{0,3\}$.

Using the negative binomial coefficient the binomial theorem can be extended to the negative integers,

$$(\alpha+\beta)^{-n} = \sum_{k=0}^{\infty} \frac{(-1)^k n^{\overline{k}}}{k!} \alpha^k \beta^{-n-k} \tag{2.1.19}$$

and the negative real numbers,

$$(1+\alpha)^{-\gamma} = \sum_{k=0}^{\infty} \frac{(-1)^k \gamma^{\overline{k}}}{k!} \alpha^k \qquad -1 < \alpha < 1. \tag{2.1.20}$$

2.2 Asymptotic series

As we shall be using the above asymptotic expansions of Stirling, a definition of an *asymptotic series* would be useful. We use the definition of Poincaré (1896): A divergent series,

$$\begin{aligned} f[z] &\sim \lim_{n\to\infty} S_n \\ S_n &= A_0 + \frac{A_1}{z} + \frac{A_2}{z^2} + \cdots + \frac{A_n}{z^n}, \end{aligned} \tag{2.2.1}$$

in which the sum of the first $(n{+}1)$ terms is $S_n[z]$, is said to be an *asymptotic expansion of a function* $f[z]$ for a given range of values of $\arg z$, if the remainder $R_n[z] = z^n\,(f[z] - S_n[z])$

satisfies the condition

$$\lim_{|z|\to\infty} R_n[z] = 0 \qquad (n \text{ fixed}), \tag{2.2.2}$$

even though

$$\lim_{n\to\infty} |R_n[z]| = \infty \qquad (z \text{ fixed}). \tag{2.2.3}$$

When this is the case, we can for a given small ϵ make $|z|$ large enough to let

$$|z^n\,[f[z] - S_n[z]]| < \epsilon. \tag{2.2.4}$$

We denote the fact that the series is the asymptotic expansion of $f[z]$ by writing

$$f[z] \sim \sum_{n=0}^{\infty} A_n z^{-n}. \tag{2.2.5}$$

To illustrate the concept we will consider the function $f[x] = \int_x^\infty (e^{x-t}/t)\,\mathrm{d}t$, where x is real and positive and the path of integration is the real axis, Whittaker and Watson (1927). By repeated integration by parts, we obtain

$$f[x] = \frac{1}{x} - \frac{1}{x^2} + \frac{2!}{x^3} - \cdots + \frac{(-1)^{n-1}(n-1)!}{x^n} + (-1)^n n! \int_x^\infty \frac{e^{x-t}}{t^{n+1}}\,\mathrm{d}t. \tag{2.2.6}$$

Naturally we consider the expansion,

$$S_n[x] = \frac{1}{x} - \frac{1}{x^2} + \frac{2!}{x^3} - \cdots + \frac{(-1)^n n!}{x^{n+1}} \tag{2.2.7}$$

in connection with the function $f(x)$. The ratio test $\left|\frac{S_{n+1}-S_n}{S_n-S_{n-1}}\right| = \frac{n}{x}$ shows that in the limit as n becomes large, n/x also goes to ∞. The partial sum $S_n(x)$ therefore diverges for all values of x in the limit $n \to \infty$. Nevertheless, the series can be used to calculate $f(x)$ for large x in the following way. Take any fixed value of n and calculate the value of S_n. We have

$$f[x] - S_n[x] = (-1)^{n+1}(n+1)! \int_x^\infty \frac{e^{x-t}}{t^{n+2}}\,\mathrm{d}t, \tag{2.2.8}$$

and therefore, since $e^{x-t} \le 1$,

$$|f[x] - S_n[x]| = (n+1)! \int_x^\infty \frac{e^{x-t}}{t^{n+2}}\,\mathrm{d}t < (n+1)! \int_x^\infty \frac{1}{t^{n+2}}\,\mathrm{d}t = \frac{n!}{x^{n+1}}. \tag{2.2.9}$$

For values of x which are sufficiently large, the right hand side of this equation is very small. Thus, if we take $x \ge 2n$, we have

$$|f[x] - S_n[x]| < \frac{1}{2^{n+1}n^2}, \tag{2.2.10}$$

which can be made as small as we like by appropriate choice of n. *Hence, the value of the function $f[x]$ can be calculated with great accuracy for large values of x by choosing a suitable number n in the expression $S_n[x]$.*

2.3 Convex functions

Here we will introduce some simple properties of quantities we will discuss later.

A function $f[x]$ is said to be *convex* over an interval (a, b) if for every $x_1, x_2 \in (a, b)$ and $0 \leq \lambda \leq 1$,

$$f[\lambda x_1 + (1-\lambda)x_2] \leq \lambda f[x_1] + (1-\lambda) f[x_2]. \tag{2.3.1}$$

A function f is said to be *strictly convex* if equality holds only if $\lambda = 0$ or $\lambda = 1$.

A function f is *concave* if $-f$ is convex.

A function is convex if it always lies below any chord connecting two points x_1 and x_2. A function is concave if it always lies above any chord. Examples of convex functions: $x^2, |x|, e^x$ and $x \log x$ for $x \geq 0$. While $\log x$ and $\sqrt{x}$ are concave for $x \geq 0$. The following useful Lemma applies:

If the function f has a second derivative which is non-negative (positive) everywhere, then the function is convex (strictly convex).

We use the Taylor expansion of the function around x_0 to prove this statement,

$$f[x] = f[x_0] + f'[x_0](x - x_0) + \frac{f''[x^*]}{2}(x - x_0)^2, \tag{2.3.2}$$

where x^* lies between x_0 and x. By convexity, $f''(x_0) \geq 0$, and thus the last term is always non-negative for all x. Let $x_0 = \lambda x_1 + (1-\lambda)x_2$ and take $x = x_1$ to obtain

$$f[x_1] \geq f[x_0] + f'[x_0]\left[(1-\lambda)(x_1 - x_2)\right]. \tag{2.3.3}$$

Similarly, take $x = x_2$ to obtain,

$$f[x_2] \geq f[x_0] + f'[x_0]\left[\lambda(x_2 - x_1)\right]. \tag{2.3.4}$$

Multiplying (2.3.3) with λ and (2.3.4) with $(1-\lambda)$ and adding them together, we obtain (2.3.1). The proof for strict convexity proceeds along similar lines.

The next inequality is one of the most widely used and useful results:

Jensen's inequality: If f is a convex function and $p[x]$ is non-negative for all x, then

$$\int f[x]p[x]\,\mathrm{d}x \geq f\left[\int xp[x]\,\mathrm{d}x\right] \quad \text{or} \quad \sum_x f[x]p[x] \geq f\left[\sum_x xp[x]\right], \tag{2.3.5}$$

in the continuous and discrete cases respectively. Moreover, if f is strictly convex, then equality in (2.3.5) implies that $p[x]$ is a constant.

We will prove this assertion for discrete distributions by induction. For a discrete function with two-point support,

$$\rho[x] = \rho_1\delta(x - x_1) + \rho_2\delta(x - x_2) \tag{2.3.6}$$

the inequality becomes,

$$\rho_1 f[x_1] + \rho_2 f[x_2] \geq f[\rho_1 x_1 + \rho_2 x_2], \tag{2.3.7}$$

which follows from the definition of a convex function. Suppose the assertion is true for functions with $k-1$ points in its support. Then writing $\rho'_j = \frac{\rho_j}{(1-\rho_k)}$ for $j = 1, 2, \ldots, k-1$, we have

$$\begin{aligned}\sum_{j=1}^{k} \rho_j f[x_j] &= \rho_k f[x_k] + (1-\rho_k)\sum_{j=1}^{k-1} \rho'_j f[x_j] \\ &\geq \rho_k f[x_k] + (1-\rho_k) f\left[\sum_{j=1}^{k-1} \rho'_j x_j\right] \\ &\geq f\left[\rho_k x_k + (1-\rho_k)\sum_{j=1}^{k-1} \rho'_j x_j\right],\end{aligned} \tag{2.3.8}$$

where the first inequality follows from the induction hypotheses and the second from the definition of convexity. So that finally

$$\sum_{j=1}^{k} \rho_j f[x_j] \geq f\left[\sum_{j=1}^{k} \rho_j x_j\right]. \tag{2.3.9}$$

This can be extended to continuous functions by taking the appropriate limits.

2.4 Moment generating function

There are many different generating functions, with the probability generating function (p.g.f.) most appropriate for discrete outcomes. Nevertheless, we shall concentrate on the moment generating function (m.g.f.) because it ties up with the saddlepoint approximation which we shall discuss subsequently. Let the outcomes of a particular system be $\mathcal{A} = \{0, 1, 2, \ldots, B\}$. Let $\rho_0, \rho_1, \ldots, \rho_B$ be a set of non-negative real numbers constrained by $\sum_{b\geq 0} \rho_b = 1$, and hence $\sum_{b\geq 0} \rho_b e^{-\lambda b}$ converges for $\lambda > 0$, even if we let $B \to \infty$. The moment generating function of this set defined by

$$\Phi[\lambda] = \sum_{b\geq 0} \rho_b e^{-\lambda b}, \tag{2.4.1}$$

where $\Phi[\lambda]$ is the abbreviation of $\Phi[\lambda, b|\boldsymbol{\rho}]$ and has the following useful properties:

(a) *For $\lambda > 0$ the generating function $\Phi[\lambda]$ is a positive and monotonically decreasing function starting from $\Phi[0] = 1$.* Note that we define $0^0 = 1$.

(b) *For $\lambda > 0$, $\Phi[\lambda]$ has derivatives of all orders.*

$$\Phi^{(n)}[\lambda] = \sum_{b\geq 0}(-b)^n \rho_b e^{-b\lambda} \quad n = 1, 2, \ldots, \tag{2.4.2}$$

which converge uniformly because the sequence is bounded and can be compared to a geometric sequence.

(c) *The second logarithmic derivative is non-negative for $\lambda > 0$,*

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\lambda}\log\Phi[\lambda] &= \frac{\Phi'[\lambda]}{\Phi[\lambda]} = -\frac{\sum b\rho_b e^{-\lambda b}}{\sum e^{-\lambda b}} && \leq \quad 0 \\ \frac{\mathrm{d}^2}{\mathrm{d}\lambda^2}\log\Phi[\lambda] &= \frac{\Phi[\lambda]\Phi''[\lambda] - \Phi'[\lambda]^2}{\Phi[\lambda]^2} \\ &= \frac{1}{\Phi[\lambda]}\sum_{b\geq 0}\left(b - \frac{\Phi'[\lambda]}{\Phi[\lambda]}\right)^2 \rho_b e^{-\lambda b} && \geq \quad 0. \end{aligned} \tag{2.4.3}$$

Consequently our generating function is a convex function for $\lambda > 0$.

(d) *The m.g.f. of the convolution of two sequences is the product of the two individual m.g.f.'s.* Consider a convolution of two sequences $\{\alpha_j\}_{j=0}^{J}$ and $\{\beta_k\}_{k=0}^{K}$. Extend the sequences to arbitrary ranges by adding zeros, i.e. $\alpha_j = 0\ \forall j < 0$ and $j > J$ etc. Then the product of m.g.f.'s is

$$\begin{aligned} \Phi_\alpha[\lambda]\Phi_\beta[\lambda] &= \sum_{k=0}^{K}\sum_{j=0}^{J}\alpha_j\beta_k e^{-\lambda(j+k)} = \sum_{k=0}^{K}\sum_{\ell=k}^{J+k}\alpha_{\ell-k}\beta_k e^{-\lambda\ell} = \sum_{k=0}^{J+K}\sum_{\ell=k}^{J+K}\alpha_{\ell-k}\beta_k e^{-\lambda\ell} \\ &= \sum_{\ell=0}^{J+K}\sum_{k=0}^{\ell}\alpha_{\ell-k}\beta_k e^{-\lambda\ell} = \sum_{\ell=0}^{J+K}\gamma_\ell e^{-\lambda\ell} = \Phi_\gamma[\lambda] \end{aligned} \tag{2.4.4}$$

where $\ell = j + k$ and we used the identity

$$\sum_{k=0}^{J+K}\sum_{\ell=k}^{J+K}[\ldots] = \sum_{0\leq k\leq\ell\leq J+K}[\ldots] = \sum_{\ell=0}^{J+K}\sum_{k=0}^{\ell}[\ldots]. \tag{2.4.5}$$

This is the important property of the generating function, where the convolution of sequences is equivalent to the product of their generating functions. This property also generalises to multiple convolutions which is easy to see if we add another convolution to our sequences,

$$\sum_{j=0}^{\ell}\rho_{\ell-j}\gamma_j = \sum_{j=0}^{\ell}\rho_{\ell-j}\sum_{k=0}^{j}\alpha_{j-k}\beta_k \tag{2.4.6}$$

and the corresponding generating functions are then clearly

$$\Phi_\gamma[\lambda]\Phi_\eta[\lambda] = \Phi_\gamma[\lambda]\Phi_\alpha[\lambda]\Phi_\beta[\lambda]. \tag{2.4.7}$$

When convolving the same sequence n times, we hence have $\Phi^n[\lambda]$ as the moment generating function of the convolution, which we are usually interested in for large n.

(e) *The inversion formula*

$$\rho_b = \frac{1}{2\pi i} \int_{-i\pi}^{i\pi} \Phi\left[\lambda\right] e^{b\lambda} \, \mathrm{d}\lambda. \tag{2.4.8}$$

is easy to prove: Multiply both sides of (2.4.1) with $\exp\left[j\lambda\right]$, where j is an arbitrary integer, and integrate the expression obtained with respect to λ from $-i\pi$ to $i\pi$,

$$\int_{-i\pi}^{i\pi} \Phi\left[\lambda\right] e^{j\lambda} \, \mathrm{d}t = \sum_{b\geq 0} \rho_b \int_{-i\pi}^{i\pi} e^{(j-b)\lambda} \, \mathrm{d}t, \tag{2.4.9}$$

The series on the right hand side being uniformly convergent for $\lambda > 0$ can be integrated term by term and since the right hand side is either equal to $2\pi i$ if $j = b$ or zero if $j \neq b$ the inversion formula follows.

(f) We can also make the substitution $\lambda = -\log z$ and simply take derivatives in z at $z = 0$ to find the inverse,

$$\rho_b = b! \, \frac{\mathrm{d}^b}{\mathrm{d}z^b} \Phi\left[-\log z\right]\Big|_{z=0}. \tag{2.4.10}$$

It may seem pointless to replace a simple operation like derivatives with something more complicated like complex integration but is easier to approximate the complex integration than the derivatives.

2.5 Moments and cumulants

Consider the convolution of two continuous distributions:

$$h(z) = \int f(x)g(y)\delta(z - x - y)dydx. \tag{2.5.1}$$

For continuous distributions it is usually easier to use the Fourier transform than the moment generating function,

$$\Phi[t, z|h(z)] = \int_{-\infty}^{\infty} h(z)e^{itz}dz, \tag{2.5.2}$$

which we abbreviate as $\Phi_h[z]$. Substituting in the convolution we have

$$\begin{aligned} \Phi_h[t] &= \int f(x)g(y)e^{itx+ity}dydx \\ &= \Phi_f[t]\Phi_g[t] \end{aligned} \tag{2.5.3}$$

which is the usual property of the generating functions that the product of the Fourier transforms is the Fourier transform of the convolutions of the distributions. This allows

us to define quantities of moments that are invariant under convolution which is a very useful property to have. Remembering the definition of the moments,

$$\mu_j = \left\langle x^j \right\rangle = \int z^j h[z] dz, \tag{2.5.4}$$

they can easily be found from the Fourier transform by taking the derivatives at zero,

$$\left.\frac{\mathrm{d}^j}{\mathrm{d}t^j}\Phi_h[t]\right|_{t=0} = \left.\frac{\mathrm{d}^j}{\mathrm{d}t^j}\int_{-\infty}^{\infty} h[z]e^{itz}dz\right|_{t=0} = \left.\frac{\mathrm{d}^j}{\mathrm{d}t^j}\int_{-\infty}^{\infty} h[z]\left[\sum_k \frac{(itz)^k}{k!}\right]dx\right|_{t=0} \tag{2.5.5}$$

$$= \left.\frac{\mathrm{d}^j}{\mathrm{d}t^j}\sum_k \frac{(it)^k \left\langle z^k \right\rangle}{k!}\right|_{\lambda=0} = i^k \left\langle z^k \right\rangle. \tag{2.5.6}$$

Taking the logarithm of the generating function changes the product structure of the convolutions into an additive structure which is amenable to a Taylor expansion around zero, and defines the cumulants κ_j,

$$\begin{aligned}\Phi_h[t] &= 1 + \mu_1 i\lambda + \mu_2\frac{(i\lambda)^2}{2!} + \mu_3\frac{(i\lambda)^3}{3!} + \mu_4\frac{(i\lambda)^4}{4!} + \dots \\ &= \exp\left[\kappa_1 i\lambda + \kappa_2\frac{(i\lambda)^2}{2!} + \kappa_3\frac{(i\lambda)^3}{3!} + \kappa_4\frac{(i\lambda)^4}{4!} + \dots\right].\end{aligned} \tag{2.5.7}$$

On expanding the exponential, we find the relations

$$\begin{aligned}\kappa_1 &= \mu_1 & \kappa_2 &= \mu_2 - \mu_1^2 \\ \kappa_3 &= \mu_3 - 3\mu_1\mu_2 + 2\mu_1^3 & \kappa_4 &= \mu_4 - 4\mu_1\mu_3 - 3\mu_2^2 + 12\mu_1^2\mu_2 - 6\mu_1^4\end{aligned} \tag{2.5.8}$$

or

$$\begin{aligned}\mu_1 &= \kappa_1 & \mu_2 &= \kappa_2 + \kappa_1^2 \\ \mu_3 &= \kappa_3 + 3\kappa_1\kappa_2 + 2\kappa_1^3 & \mu_4 &= \kappa_4 + 4\kappa_1\kappa_3 + 3\kappa_2^2 + 12\kappa_1^2\kappa_2 + 6\kappa_1^4\end{aligned} \tag{2.5.9}$$

and so forth. The first two cumulants are called the mean and the variance, while the third and fourth are related to the skewness ($\frac{\kappa_3}{\kappa_2^{3/2}}$) and kurtosis ($\frac{\kappa_4}{\kappa_2^2}$) respectively. Skewness is a measure of the asymmetry of the distribution and kurtosis indicates the strength of the decay in the tails of the distribution.

2.6 Saddlepoint approximations

Following Bleistein and Handelsman (1986) we introduce the saddlepoint approximation as we shall repeatedly need to invert the convolutions of discrete distributions in their generating function form. From the inversion formula (2.4.8) if follows that the mean of n convolutions of the same distribution is given by,

$$\rho_b = \frac{1}{2\pi i}\int_{-i\pi}^{i\pi} \Phi^n[\lambda]e^{b\lambda}\,\mathrm{d}\lambda. \tag{2.6.1}$$

Taking $x = b/n$ as the variable (which becomes continuous for large n) we assume $n \gg b$

$$\rho\left[x\right] = \frac{1}{2\pi i}\int_{-i\pi}^{i\pi} \Phi^n\left[\lambda\right] e^{nx\lambda}\,\mathrm{d}\lambda. \tag{2.6.2}$$

Visualise the behaviour of the integral as we move from zero to infinity on the real positive axis. The first factor of the integrand, $\Phi^n\left[\lambda\right]$, starts at one and decreases monotonically. The second factor, $e^{n\lambda x}$, also starts at one when $\lambda = 0$ but increases monotonically to infinity. In addition, the relative increase of the second factor,

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\log e^{\lambda nx} = nx, \tag{2.6.3}$$

is constant and the relative decrease of the first factor,

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\log \Phi^n\left[\lambda\right] = -n\frac{\sum_b b\rho_b e^{-b\lambda}}{\sum_b \rho_b e^{-b\lambda}} \tag{2.6.4}$$

decreases itself monotonically. Indeed we know from (2.4.3) that it is a convex function on the real positive axis. Under these circumstances integrand $\Phi^n[\lambda]e^{nx\lambda}$ will have only one minimum and no other maxima or minima on the positive real axis. This minimum will be very steep because both exponents, n and nx, will be very large numbers for any constant x. At this point on the real positive axis (which we will call the saddlepoint) the first derivative of the integrand will disappear and the second derivative will be positive and very large. We first determine the saddlepoint, where the first derivative of the integrand vanishes. It is convenient to use the logarithm of the exponential generating function here,

$$K\left[\lambda\right] = \log\Phi[\lambda] + \lambda b. \tag{2.6.5}$$

Then λ^* is defined by the equation,

$$K^{(1)}\left[\lambda^*\right] = \frac{\mathrm{d}}{\mathrm{d}\lambda}\log\Phi\left[\lambda\right]\Big|_{\lambda=\lambda^*} + b = 0 \tag{2.6.6}$$

and we also know that $K^{(2)}\left[\lambda^*\right] > 0$ from (2.4.3) and that there is only one solution to this equation. Define the path of integration for (2.6.2) as a straight line segment through λ^* parallel to the imaginary axis terminating at $\lambda^* \pm i\pi$. Since $K[\lambda]$ has a minimum at λ^* for real λ, the modulus of the integrand must have a maximum at λ^* on the chosen path[1]. Consequently, for the particular path chosen only values near the neighbourhood

[1]It can be shown that for any such terminating straight line segment parallel to the imaginary axis the integrand attains its maximum modulus only where the line crosses the real axis. For on the line $\lambda = \mu + i\nu$,

$$\left|e^{K[\lambda]}\right| = e^{\mu b}\left|\sum_j \rho_j e^{-(\mu+i\nu)j}\right| \leq e^{K[\mu]}. \tag{2.6.7}$$

Assume equality holds for some $\nu \neq 0$, so that $\sum \rho_b e^{-(\mu+i\nu)b} = \Phi[\lambda]e^{i\alpha}$ where α is some real constant, which gives

$$\sum \rho_b e^{-\mu b}\left[1 - \cos(\nu b - \alpha)\right] = 0. \tag{2.6.8}$$

Implying $\cos(\nu b - \alpha) = 1$ for all integral b and some α, but $v = 0$ is the only possible solution in $(-\pi, \pi)$ and consequently equality can only hold on the real axis

of λ^* need be considered when n is large. Intuitively all the terms in (2.4.1) will reinforce each other on the real axis but as we add an imaginary component the terms will start to 'rotate' at different speeds according to the values of b and thus the integrand will in general be considerably less as we move away from the real axis.

Making the substitution $t = \sqrt{n}(\lambda - \lambda^*)$ and Taylor expanding around $t = 0$ changes the integral (2.6.2) into

$$\rho\left[x\right] = \frac{e^{nK[\lambda^*]}}{2\pi i\sqrt{n}} \int_{\sqrt{n}(-i\pi-\lambda^*)}^{\sqrt{n}(i\pi-\lambda^*)} \exp\left[K^{(2)}[\lambda^*]\frac{t^2}{2} + \sum_{j=3}^{\infty} \frac{K^{(j)}[\lambda^*]}{\sqrt{n}^{j-2}}\frac{t^j}{j!}\right] \mathrm{d}t. \tag{2.6.9}$$

Expanding the exponential of higher order terms yields,

$$\begin{aligned}\rho\left[x\right] &= \frac{e^{nK[\lambda^*]}}{2\pi i\sqrt{n}} \int_{\sqrt{n}(-i\pi-\lambda^*)}^{\sqrt{n}(i\pi-\lambda^*)} \exp\left[\frac{K^{(2)}[\lambda^*]}{2}t^2\right] \prod_{j\geq 3}\left[1 + \sum_{k=1}^{\infty}\frac{1}{k!}\left(\frac{K^{(j)}[\lambda^*]t^j}{\sqrt{n}^{j-2}j!}\right)^k\right] \mathrm{d}t \\ &= \frac{e^{nK[\lambda^*]}}{2\pi i\sqrt{n}} \int_{\sqrt{n}(-i\pi-\lambda^*)}^{\sqrt{n}(i\pi-\lambda^*)} \exp\left[\frac{K^{(2)}[\lambda^*]}{2}t^2\right]\left[1 + \sum_{j\geq 3}^{\infty} A_j t^j\right] \mathrm{d}t,\end{aligned} \tag{2.6.10}$$

where A_j is a set of coefficients that are functions of the logarithmic derivatives of the exponential generating function at the saddlepoint and $A_j \sim n^{1-j/2}$. Assuming n is large we take the contour from minus infinity to infinity (the corrections are sub-exponential, see Bender and Orszag (1999)) which we can then displace to the origin using the Cauchy theorem (we know that there are no singularities on the real positive axis). Using the standard integral true for $a > 0$ and $j \geq 0$,

$$\frac{1}{2\pi i}\int_{-i\infty}^{i\infty} e^{at^2}t^j \,\mathrm{d}t = \begin{cases} \frac{(-1)^{j/2}}{2\pi}\left(\frac{j-1}{2}\right)!a^{-\frac{j+1}{2}} & \text{for even } j \\ 0 & \text{for odd } j, \end{cases} \tag{2.6.11}$$

with $a = \frac{K^{(2)}[\lambda^*]}{2}$.

Finally we have the saddlepoint approximation to the mean of n convolutions of a distribution,

$$\rho\left[x\right] = \frac{e^{nK[\lambda^*]}}{\sqrt{2\pi n K^{(2)}[\lambda^*]}}\left[1 + \sum_{j=2}^{\infty}\frac{A_{2j}}{(K^{(2)}[\lambda^*])^j}\frac{(-1)^j(2j-1)!}{2^j(j-1)!}\right]. \tag{2.6.12}$$

The advantage that the saddlepoint has over a conventional Gaussian approximation is that the saddlepoint ensures that all the odd coefficients do not contribute; thus the saddlepoint approximation with no corrections is order $O\left(\frac{1}{n}\right)$ while a Gaussian approximation gives $O\left(\frac{1}{\sqrt{n}}\right)$, see Daniels (1954) and Stuart and Ord (1994). The first coefficient in the expansion is,

$$A_4 = \frac{K^{(4)}[\lambda^*]}{24n} \tag{2.6.13}$$

and adding it would give us $O\left(\frac{1}{n^2}\right)$ accuracy.

It is not obvious from the above derivation that the saddlepoint approximation is a proper asymptotic expansion. To prove the assertion requires the use of the method of steepest descent: an account is given in Jeffreys and Jeffreys (1950) and uses the lemma of Watson (1948).

Chapter 3

Basics of Bayesian probability

> They say that Understanding ought to work by the rules of right reason. These rules are, or ought to be, contained in Logic; but the actual science of logic is conversant at present only with things either certain, impossible, or entirely doubtful, none of which (fortunately) we have to reason on. Therefore the true logic for this world is the calculus of Probabilities, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind.
>
> J. Clerk Maxwell

3.1 Introduction

The fundamental problem of science, and a fundamental problem of everyday life, is that of learning from experience. Knowledge obtained in this way is in part a description of what we have observed (past experiences) and in part a prediction of future experiences. The prediction part is called induction or generalisation and is the part that we are interested in. This is not to say that the description of unrelated things are not important as they might lead to predictions in the future. Logic in its usual sense does not contain any of these inferences. Formal logic only admits three attitudes true, false and ignorance. Therefore as pointed out by Maxwell, we have to use probabilities as a degree of belief if we want to formalise the thinking of a scientist. The discussion in this chapter is based on Jeffreys (1961), Jaynes (2003) and Lindley (2006).

As we can draw an infinite number of functions through a finite number of fixed points, generalisations or inductions are by definition subjective. In the same way probabilities are predictions and thus theoretical quantities. So when Boltzmann (1974) says "the task of theory consists in constructing a picture of the external world that exists purely internally" and De Finetti (1974) says "probabilities do not exist", they are in fact saying the same thing. On the one hand we have the ontological world of experimental observation and on the other hand we have an epistemological world which consists of theory and probabilities. The ontological world is called objective and the epistemological world is called subjective.

This is of course a trivial observation but it does seem to cause confusion. For example it is impossible to measure a probability. If two observers measure a distance they can agree on the length, but if two observers assign a probability to an event they do not have to agree at all because they can have different background information. Take a black box and place either a red or blue ball in it. Now place an observer inside the box and an observer outside the box. The observer inside is absolutely sure which ball will be drawn and the observer outside does not know, thus they assign different probabilities to the same event and they are both correct. Also if the inside observer now tells the outside observer what he knows the outside observer changes his prediction and assigns different probabilities, and thus the changing of probabilities does not necessarily imply any physical change.

Each observer has his own information time line with the things he has learned in his information past and things that he predicts in his information future. This information time line does not correspond to physical time and can even work in the opposite direction. For example, learning something new from a fossil changes the prediction we make of something that has occurred physically a few million years in the past.

To use probabilities as an element of logic or a degree of belief in a *logical proposition* places a certain burden on us. We have to be reasonable men. First we have to define the logical propositions we are discussing. We will call propositions *uncertain* if we do not know whether they are false or true. A typical example would be,

$$\mathcal{A} \equiv \texttt{It will rain tomorrow.} \tag{3.1.1}$$

Then we must describe the *knowledge* that we will use to assign a probability. This may require an onerous process of identifying and specifying, relevant facts or issues such as

- Is tomorrow from midnight tonight to midnight tomorrow or from morning till evening?
- How much water constitutes rain, a bit of dew or at least one millimetre in a cup placed where exactly?
- What do we assume is known about the physical process of rain?
- Do historical observations of rain influence our probability assignment?
- Is there enough consistency in nature to permit such a prediction at all?

The set of answers and questions describes our knowledge base,

$$\mathcal{K} \equiv \texttt{The knowledge we consider relevant to the problem.} \tag{3.1.2}$$

Being reasonable men also implies that we must reason consistently: if two observers are provided with the same knowledge base they should assign the same probabilities. All this goes to show that every probability is a conditional probability,

$$\rho(\mathcal{A}|\mathcal{K}), \tag{3.1.3}$$

where $\mathcal{A}, \mathcal{B}, \ldots$ will be called Boolean propositions because we shall assume they obey Boolean algebra (see Appendix B). Everything to the right of the vertical line is assumed given and known and everything to the left is uncertain. As a matter of principle all probabilities should have something to the right of the line. Before we can translate these logical operations into mathematical rules, we have to translate our *degrees of belief*[1] onto a mathematical scale and the traditional choice is to denote true with one and false with zero and the rest with a number between zero and one depending on their uncertainty. This leads to the

> Convexity Rule: For any proposition $\mathcal{A}$ and knowledge base $\mathcal{K}$, the observer assigns the probability $p(\mathcal{A}|\mathcal{K})$ for $\mathcal{A}$ given $\mathcal{K}$ as a rational number between zero and one. The observer only assigns $p(\mathcal{A}|\mathcal{K}) = 1$ if $\mathcal{K}$ logically implies $\mathcal{A}$ and only assigns $p(\mathcal{A}|\mathcal{K}) = 0$ if $\mathcal{A}$ is logically false on $\mathcal{K}$.

3.2 Sum rule and Product rule

We can write the four basic logical constructs algebraically or logically as

$$\begin{aligned} \mathcal{A}|\mathcal{K} &\leftrightarrow \mathcal{A} \text{ GIVEN } \mathcal{K} \\ \mathcal{A}, \mathcal{B} &\leftrightarrow \mathcal{A} \text{ AND } \mathcal{B} \\ \mathcal{A} + \mathcal{B} &\leftrightarrow \mathcal{A} \text{ OR } \mathcal{B} \\ \overline{\mathcal{A}} &\leftrightarrow \text{NOT } \mathcal{A}. \end{aligned} \tag{3.2.1}$$

It seems there should be an infinite number of mappings of these rules but consistency forces all the different rules to be equivalent. The proof of this, known as Cox's theorem, and is set out in Appendix B. We could have chosen a different convention but we will show later that different conventions will have exactly the same content i.e. that there is a one to one mapping between the different conventions and we would gain nothing by doing things differently.

The logical OR applied to proposition $\mathcal{A}$ and its contradiction $\overline{\mathcal{A}}$ should give certainty as we know from deductive logic, meaning that

$$p(\mathcal{A}|\mathcal{K}) + p(\overline{\mathcal{A}}|\mathcal{K}) = 1 \qquad (\textbf{Sum rule}). \tag{3.2.2}$$

From Cox's theorem we know that all other choices are identical to this choice. Next consider the logical AND operation: we have to think on how to combine logical propositions. From Cox (1946), let proposition $\mathcal{A}$ denote *an athlete can run from one place to another* and proposition $\mathcal{B}$ *he can run back again without stopping.* The knowledge base $\mathcal{K}$ is what the observer knows about running from here to the distant place and what he knows about the physical condition of the athlete. $p(\mathcal{B}|\mathcal{A}, \mathcal{K})$ is the probability that the athlete will

[1] We use the term *degree of belief* strictly in the operational sense defined here. Matters of metaphysics and ontology do not form part of our investigation in this dissertation.

return assuming that he reached the distant place and $p(\mathcal{A}|\mathcal{K})$ is the probability that he will reach the distant place. Common sense says the product of the two is the probability $p(\mathcal{AB}|\mathcal{K})$ that he completes the race. Cox's theorem shows that all other choices are equivalent and that generally

$$p(\mathcal{A},\mathcal{B}|\mathcal{K}) = p(\mathcal{A}|\mathcal{B},\mathcal{K})p(\mathcal{B}|\mathcal{K}) = p(\mathcal{B}|\mathcal{A},\mathcal{K})p(\mathcal{A}|\mathcal{K}) \qquad (\textbf{Product rule}). \tag{3.2.3}$$

3.2.1 Bayes' Rule and its associates

We define **logical independence** as

$$p(\mathcal{A}|\mathcal{B},\mathcal{K}) = p(\mathcal{A}|\mathcal{K}). \tag{3.2.4}$$

The uncertainty in the proposition $\mathcal{A}$ remains unaffected by observing or learning proposition $\mathcal{B}$. Through the product rule, this is equivalent to the **joint probability** factorising,

$$p(\mathcal{A},\mathcal{B}|\mathcal{K}) = p(\mathcal{A}|\mathcal{K})p(\mathcal{B}|\mathcal{K}). \tag{3.2.5}$$

We will call probabilities that factorise a set of *chances*. It is also essential to realise that logical independence is again not a property of an event or an object. Examine a bent coin for example: If we are told that a bent coin has a 60% chance to land heads and a 40% chance of landing tails, this information makes one throw of the coin a logically independent event against all the other throws of the coin. But if we are told the coin is bent but not towards which side it favours the throws of the coin are not logically independent, but physically nothing has changed between the two scenarios. There is a fundamental connection between ignorance and independence.

To proceed we need to refine the sum rule further: we seek a formula for the logical sum $\mathcal{A}+\mathcal{B}$. Using De Morgan's (B.2.5) law on $\mathcal{A}+\mathcal{B}$ and repeatedly applying our rules (see also Jaynes (2003)) gives

$$\begin{aligned} p(\mathcal{A}+\mathcal{B}|\mathcal{K}) &= 1 - p(\overline{\mathcal{A}},\overline{\mathcal{B}}|\mathcal{K}) = 1 - p(\overline{\mathcal{A}}|\mathcal{K})p(\overline{\mathcal{B}}|\overline{\mathcal{A}},\mathcal{K}) \\ &= 1 - p(\overline{\mathcal{A}}|\mathcal{K})\left[1 - p(\mathcal{B}|\overline{\mathcal{A}},\mathcal{K})\right] = p(\mathcal{A}|\mathcal{K}) + p(\overline{\mathcal{A}},\mathcal{B}|\mathcal{K}) \\ &= p(\mathcal{A}|\mathcal{K}) + p(\mathcal{B}|\mathcal{K})p(\overline{\mathcal{A}}|\mathcal{B},\mathcal{K}) \\ &= p(\mathcal{A}|\mathcal{K}) + p(\mathcal{B}|\mathcal{K})\left[1 - p(\mathcal{A}|\mathcal{B},\mathcal{K})\right] \end{aligned} \tag{3.2.6}$$

and we end up with

$$p(\mathcal{A}+\mathcal{B}|\mathcal{K}) = p(\mathcal{A}|\mathcal{K}) + p(\mathcal{B}|\mathcal{K}) - p(\mathcal{A},\mathcal{B}|\mathcal{K}). \tag{3.2.7}$$

The sum rule (3.2.2) is a special case $\mathcal{B} = \overline{\mathcal{A}}$ of this **Extended Sum Rule**. In view of this rule it would be convenient to define a set of exclusive and exhaustive propositions. Propositions $\mathcal{A}$ and $\mathcal{B}$ are *exclusive* when $p(\mathcal{A},\mathcal{B}|\mathcal{K}) = 0$ i.e. if one of them is true the other is false. Propositions $\mathcal{A}$ and $\mathcal{B}$ are *exhaustive* when one of them must be true

$p(\mathcal{A}+\mathcal{B}|\mathcal{K})=1$. After constructing an exclusive and exhaustive set of propositions $\mathcal{F}_b$, we can partition any proposition $\mathcal{A}$ into this set,

$$\begin{aligned} &p(\mathcal{F}_1+\cdots+\mathcal{F}_B|\mathcal{K})=1 \\ &p(\mathcal{F}_a,\mathcal{F}_b|\mathcal{K})=0 \quad \forall\, a\neq b \\ &p(\mathcal{A}|\mathcal{K})=p(\mathcal{A}\mathcal{F}_1+\cdots+\mathcal{A}\mathcal{F}_B|\mathcal{K})=\sum_b p(\mathcal{A},\mathcal{F}_b|\mathcal{K}), \end{aligned} \tag{3.2.8}$$

which is much simpler than the B-fold extended sum rule. For an arbitrary parameter θ we would for example take the propositions $\mathcal{F}$ as intervals on the support of the parameter so that it forms an exhaustive set and non-overlapping intervals and thus an exclusive set as well. Then on making each interval smaller while increasing the number of intervals, the sum rule becomes,

$$p(\mathcal{D}|\mathcal{K})=\int p(\mathcal{D},\boldsymbol{\theta}|\mathcal{K})d\boldsymbol{\theta}. \tag{3.2.9}$$

Summing or integrating out a parameter is called **marginalisation** and gives the **evidence** for that parameter.

If we divide the evidence (assuming $p(\mathcal{D}|\mathcal{K})\neq 0$) into our joint probability distribution $p(\mathcal{D},\boldsymbol{\theta}|\mathcal{K})$ we derive **Bayes Rule**,

$$\boxed{p(\boldsymbol{\theta}|\mathcal{D},\mathcal{K})=\frac{p(\boldsymbol{\theta},\mathcal{D}|\mathcal{K})}{p(\mathcal{D}|\mathcal{K})}=\frac{p(\mathcal{D}|\boldsymbol{\theta},\mathcal{K})p(\boldsymbol{\theta}|\mathcal{K})}{p(\mathcal{D}|\mathcal{K})}} \tag{3.2.10}$$

or

$$\textbf{posterior}=\frac{\textbf{likelihood}\times\textbf{prior}}{\textbf{evidence}}, \tag{3.2.11}$$

where we also write $L[\boldsymbol{\theta}]$ for the likelihood if we consider the data fixed and $\pi[\boldsymbol{\theta}]$ for the prior. Bayes Rule is used when we want to invert the conditioning of the probabilities. Of course in science we need this in all parameter estimation problems. We usually have a model that predicts the data given a set of parameters, but what we need is the probability for the parameters given the data that we have observed.

3.3 Example: straight line fitting

Let us illustrate these rules by applying them to a perennial problem: drawing the optimal straight line through a scatter plot of data points, see Zellner (1971) and Jaynes (1990). To make the example interesting we assume error in both x and y variables:

$$y=y^*+\epsilon_y \qquad x=x^*+\epsilon_x \qquad y^*=\beta x^*+\alpha \tag{3.3.1}$$

where α is the unknown intercept of our straight line, β the slope, and x^* and y^* are the 'true' values. The first logical proposition we make is what we observe (namely the data),

$$\mathcal{D}\equiv\{(x_1,y_1),(x_2,y_2),\ldots,(x_J,y_J)\} \tag{3.3.2}$$

and as part of our knowledge base we will assume that the errors are distributed like a Gaussian distribution,

$$p(x|x^*,\sigma_x,\mathcal{K}) = \frac{1}{\sqrt{2\pi}\sigma_x}\exp\left[-\frac{(x-x^*)^2}{2\sigma_x^2}\right]. \tag{3.3.3}$$

While we will derive from first principles later, usually nobody would object to this assumption. The probability $p(x|x^*,\sigma_x,\mathcal{K})$ is called a forward probability: our observations are explained in terms of a set of parameters that are assumed known; in this case x^* and σ_x. The total list of parameters is of course,

$$\boldsymbol{\theta} = \{\alpha,\beta,x^*,y^*,\sigma_x,\sigma_y\}. \tag{3.3.4}$$

Next we have to assign a probability to the whole of the data set, $p(\mathcal{D}|\boldsymbol{\theta},\mathcal{K})$. To do this we will need to make some assumptions: First, that there is no time evolution in our data set and hence that there is no preferred ordering of the data, implying **exchangeability** i.e. we assign the same probability for all permutations of the data set,

$$p(\{x_1,y_1\},\{x_2,y_2\},\ldots,\{x_J,y_J\}|\mathcal{K}) = p(\{x_{\pi_1},y_{\pi_1}\},\{x_{\pi_2},y_{\pi_2}\},\ldots,\{x_{\pi_J},y_{\pi_J}\}|\mathcal{K}), \tag{3.3.5}$$

where $\{\pi_j\}$ is some reordering of the data. In terms of our physical/information time picture we are assuming that our signal(data) is stationary in physical time. Secondly, we will make the even stronger assumption of **logical independence**, which implies that our the data is stationary in information time as well. Logical independence necessarily implies that we cannot update the assigned probabilities dynamically as we learn new data points, or in other words that the model is assumed fixed.

Notice that logical independence implies exchangeability but the converse is not true. Exchangeability will become far more important than logical independence in the later chapters. Applying logical independence to our data set gives us

$$p(\mathcal{D}|\boldsymbol{\theta},\mathcal{K}) = \prod_j^J p(x_j,y_j|\boldsymbol{\theta},\mathcal{K}). \tag{3.3.6}$$

Using our Gaussian error model and logical independence we find

$$p(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{\theta},\mathcal{K}) = \prod_j^J \frac{1}{2\pi\sigma_x\sigma_y}\exp\left[-\frac{(x_j-x_j^*)^2}{2{\sigma_x}^2}-\frac{(y_j^*-y_j)^2}{2{\sigma_y}^2}\right] \tag{3.3.7}$$

our **likelihood** function, which we will also write as $L[\boldsymbol{\theta}]$ when we consider our data fixed. Using our straight line model: we can either write our likelihood function in terms of x_j^*,

$$p(\mathcal{D}|\boldsymbol{\theta},\mathcal{K}) = \prod_j^J \frac{1}{\sqrt{2\pi}\sigma_x\sigma_y}\exp\left[-\frac{(x_j-x_j^*)^2}{2{\sigma_x}^2}-\frac{(\beta x_j^*+\alpha-y_j)^2}{2{\sigma_y}^2}\right] \tag{3.3.8}$$

or in terms of y^*,

$$p(\mathcal{D}|\boldsymbol{\theta},\mathcal{K}) = \prod_j^J \frac{1}{\sqrt{2\pi}\sigma_x\sigma_y}\exp\left[-\frac{(x_j-\frac{y^*-\alpha}{\beta})^2}{2{\sigma_x}^2}-\frac{(y_j^*-y_j)^2}{2{\sigma_y}^2}\right], \tag{3.3.9}$$

which in both cases reduces our unknown parameters.

Our second logical proposition in this example is the **prior** which is supposed to capture mathematically what we know of the parameters before we have seen the data. Looking first at x_j^* and y_j^*, they are both locations and thus we consider the whole real line as equally likely,

$$p(x^*|\mathcal{K}) = p(y^*|\mathcal{K}) \propto 1, \tag{3.3.10}$$

which is an improper prior meaning that it is not properly normalised. We will first discuss only one parameter and then deal with the others as they arise. Using the product rule we combine the two logical statements about parameters and data,

$$p(\boldsymbol{\theta}, \mathcal{D}|\mathcal{K}) = p(\boldsymbol{\theta}|\mathcal{K})p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{K}) \tag{3.3.11}$$

which as previously noted is called the joint probability distribution.

In this example the x_j^* or y_j^* are **nuisance parameters** as they are only used to set up the model, are then integrated out and play no further role in the analyses. Marginalising over all the x_j^*,

$$\begin{aligned} p(\mathcal{D}|\alpha, \beta, \sigma_x, \sigma_y, \mathcal{K}) &\propto \prod_{j=1}^{J} \int p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{K}) dx_j^* \\ &\propto \left[2\pi(\beta^2\sigma_x^2 + \sigma_y^2)\right]^{-\frac{J}{2}} \exp\left[-\frac{JQ[x,y]}{2(\beta^2\sigma_x^2 + \sigma_y^2)}\right], \end{aligned} \tag{3.3.12}$$

with

$$Q[x,y] = \frac{1}{J}\sum_{j=1}^{J}(\beta x_j + \alpha - y_j)^2. \tag{3.3.13}$$

There is also another nuisance parameter: Examining the likelihood function (3.3.12), σ_x and σ_y only appear in the combination $\beta^2\sigma_x^2 + \sigma_y^2$, thus we can transform from $\{\sigma_x, \sigma_y\} \rightarrow \{\sigma, \lambda\}$ using

$$\sigma^2 = \beta^2\sigma_x^2 + \sigma_y^2 \qquad\qquad \lambda = \frac{\sigma_x}{\sigma_y} \tag{3.3.14}$$

with the Jacobian

$$\frac{\partial(\sigma_x, \sigma_y)}{\partial(\sigma, \lambda)} = \frac{\sigma}{\beta^2\lambda^2 + 1}. \tag{3.3.15}$$

This illustrates an important point: change of parameters does not affect our likelihood function but the Jacobian of the transformation does end up in our prior distribution. The unknown error parameters σ_x and σ_y we view as unknown scales in our problem and thus their prior distribution will be

$$p(\sigma_x, \sigma_y)d\sigma_x d\sigma_y \propto \frac{d\sigma_x}{\sigma_x}\frac{d\sigma_y}{\sigma_y}, \tag{3.3.16}$$

which we will motivate later. Transforming to our new set of parameters,

$$p(\sigma,\lambda)d\sigma d\lambda \propto \frac{\sqrt{\beta^2\lambda^2+1}}{\lambda\sigma}\frac{\sqrt{\beta^2\lambda^2+1}}{\sigma}\frac{\sigma}{\beta^2\lambda^2+1}d\sigma d\lambda \propto \frac{d\sigma}{\sigma}\frac{d\lambda}{\lambda}, \tag{3.3.17}$$

we conclude that our new parameters are also unknown scales. Also because the λ parameter does not appear in our likelihood function we can just marginalise it away, leaving us with the simplified form,

$$L[\alpha,\beta,\sigma] = \left(\sqrt{2\pi}\sigma\right)^{-J}\exp\left[-J\frac{Q[x,y]}{2\sigma^2}\right]. \tag{3.3.18}$$

It is a great advantage of Bayesian statistics that it can remove nuisance parameters through marginalising, which is not always so simple. For example if we chose a more complicated prior such as a Levy distribution for both σ_x and σ_y,

$$p(\sigma_x,\sigma_y|\mathcal{K}) = \frac{2}{\pi\sigma_y^2\sigma_x^2}\exp\left[-\frac{1}{2\sigma_x^2}-\frac{1}{2\sigma_y^2}\right] \tag{3.3.19}$$

transformed to the new variables,

$$p(\lambda,\sigma|\mathcal{K}) = \exp\left[-\frac{(1+\lambda^2)(1+\beta^2\lambda^2)}{2\lambda^2\sigma^2}\right]\frac{2\left(1+\beta^2\lambda^2\right)}{\pi\lambda^2\sigma^2}, \tag{3.3.20}$$

and integrated out the λ, we would obtain

$$p(\sigma|\mathcal{K}) = \sqrt{\frac{2}{\pi}}\frac{(1+\beta)}{\sigma^2}\exp\left[-\frac{(1+\beta)^2}{2\sigma^2}\right], \tag{3.3.21}$$

which shows that the functional invariance we observed in (3.3.17) is a property of that specific choice of prior distributions. In general marginalising over a nuisance parameter that does not appear in our likelihood function can change the prior distributions of other parameters.

Defining the sample moments,

$$\begin{aligned} \overline{x} &\equiv \sum_j^J \frac{x_j}{J} & \overline{y} &\equiv \sum_j^J \frac{y_j}{J} \\ \overline{x^2} &\equiv \sum_j^J \frac{x_j^2}{J} & \overline{y^2} &\equiv \sum_j^J \frac{y_j^2}{J} & \overline{xy} &\equiv \sum_j^J \frac{x_j y_j}{J}, \end{aligned} \tag{3.3.22}$$

sample variances and a correlation coefficient,

$$s_{yy}^2 \equiv \overline{y^2}-\overline{y}^2 \qquad s_{xx}^2 \equiv \overline{x^2}-\overline{x}^2 \qquad s_{xy}^2 \equiv \overline{xy}-\overline{x}\,\overline{y} \qquad \phi \equiv 1-\frac{s_{xy}^4}{s_{xx}^2 s_{yy}^2}, \tag{3.3.23}$$

we can rewrite after some algebra the quadratic function $Q[x,y]$ as,

$$Q[x,y] = (\alpha-\overline{\alpha})^2 + 2\overline{x}\,(\alpha-\overline{\alpha})\left(\beta-\overline{\beta}\right) + \overline{x}^2\left(\beta-\overline{\beta}\right)^2 + s_{yy}^2\phi, \tag{3.3.24}$$

with

$$\overline{\beta} = \frac{s_{xy}^2}{s_{xx}^2} \qquad\qquad \overline{\alpha} = \overline{y}-\overline{\beta}\,\overline{x}. \tag{3.3.25}$$

Hence the sample moments are **sufficient statistics**: according to our model they are the only properties of the data on which our inferences are based. Choosing $p(\alpha|\mathcal{K})$ constant allows us to marginalise out the α,

$$p(\mathcal{D}|\beta,\sigma,\mathcal{K}) \propto \frac{(\sqrt{2\pi}\sigma)^{1-J}}{\sqrt{J}} \exp\left[-J\frac{s_{xx}\left(\overline{\beta}-\beta\right)^2+s_{yy}\phi}{2\sigma^2}\right], \tag{3.3.26}$$

and for the slope β we also consider all values equally likely, $p(\beta|\mathcal{K}) \propto 1$, so that

$$p(\mathcal{D}|\sigma,\mathcal{K}) \propto \frac{\sigma\left(\sqrt{2\pi}\sigma\right)^{1-J}}{J\sqrt{2\pi s_{xx}}} \exp\left[-J\frac{s_{yy}\phi}{2s_{xx}\sigma^2}\right] \tag{3.3.27}$$

and finally with $p(\sigma|\mathcal{K}) \propto \frac{1}{\sigma}$ our evidence is

$$P(\mathcal{D}|\mathcal{K}) \propto \frac{\Gamma\left[\frac{J}{2}-1\right]\pi^{1-\frac{J}{2}}s_{yy}\phi}{2\sqrt{s_{xx}}}\left(Js_{yy}\phi\right)^{-\frac{J}{2}}. \tag{3.3.28}$$

Applying Bayes theorem (3.2.10) yields the posterior

$$p(\alpha,\beta,\sigma|\mathcal{D},\mathcal{K}) = \frac{2\sqrt{s_{xx}}}{\pi s_{yy}\phi\Gamma\left[\frac{j}{2}-1\right]}\left(\frac{Js_{yy}\phi}{2\sigma^2}\right)^{\frac{J}{2}}\exp\left[-J\frac{Q[x,y]}{2\sigma^2}\right]. \tag{3.3.29}$$

An observant reader would have noticed that we specified our prior distributions only proportionally, but our posterior is suddenly normalised. The logic is that the posterior is a ratio and our improper prior distribution contains a normalisation constant which cancels.

3.4 Summarising the posterior

If the posterior can be found analytically for all values of $\boldsymbol{\theta}$, it represents the answer of maximal information. Often, however it is easier to characterise the distribution in terms of a small number of measures, say location, dispersion and skewness instead of describing the whole posterior distribution. The problem of choosing a single measure of location is a well-known problem in descriptive statistics. Defining expectation of any given parameter θ as

$$\begin{aligned} \langle\theta\rangle &= \sum_b \theta_b p(\theta_b|\mathcal{D},\mathcal{K}) \\ \text{and} \quad \langle\theta\rangle &= \int \theta p(\theta|\mathcal{D},\mathcal{K})d\theta, \end{aligned} \tag{3.4.1}$$

in the discrete case or continuous case respectively.

The single measure of location we will call a point estimate. Our first example of a location measure is the posterior mean $\hat{\theta} = \langle\theta\rangle$. We chose the point estimate $\hat{\theta}$ to minimize the expectation of a loss or risk function $R[\hat{\theta},\theta]$

$$\min_{\hat{\theta}}\left\langle R[\hat{\theta},\theta]\right\rangle = \min_{\hat{\theta}}\int R[\hat{\theta},\theta]p(\theta|\mathcal{D},\mathcal{K})d\theta \tag{3.4.2}$$

and for the posterior mean we minimize the quadratic loss. Of course we are assuming $\left\langle R[\hat{\theta},\theta]\right\rangle$ is finite and that the minimum exists. Since for a given $p(\hat{\theta}|\mathcal{D},\mathcal{K})$ the variation of θ in the second term of the expected loss is fixed,

$$\begin{aligned} R[\hat{\theta},\theta] &= (\theta-\hat{\theta})^2 \\ \left\langle (\theta-\hat{\theta})^2\right\rangle &= \left\langle\theta^2\right\rangle - 2\hat{\theta}\left\langle\theta\right\rangle + \hat{\theta}^2 = \left(\hat{\theta}-\left\langle\theta\right\rangle\right)^2 + \left(\left\langle\theta^2\right\rangle - \left\langle\theta\right\rangle^2\right). \end{aligned} \tag{3.4.3}$$

$\left\langle (\theta-\hat{\theta})^2\right\rangle$ is therefore minimized by setting,

$$\hat{\theta} = \left\langle\theta\right\rangle = \int \theta p(\theta|\mathcal{D},\mathcal{K})d\theta. \tag{3.4.4}$$

The expected quadratic loss of the posterior mean is also called the variance of θ,

$$\mathrm{var}(\theta) = \left\langle\theta^2\right\rangle - \left\langle\theta\right\rangle^2. \tag{3.4.5}$$

But this may not be what we really want. There are valid arguments against using the mean values as the squared error criterion considers errors twice as great as four times as serious and thus focuses its attention on large but not very probable errors. We could also have used Laplace's original criterion namely minimizing the absolute loss,

$$\begin{aligned} R[\hat{\theta},\theta] &= \left|\theta-\hat{\theta}\right| \\ \left\langle\left|\hat{\theta}-\theta\right|\right\rangle &= \int_{-\infty}^{\hat{\theta}} (\hat{\theta}-\theta)p(\theta|\mathcal{D},\mathcal{K})\,\mathrm{d}\theta + \int_{\hat{\theta}}^{\infty} (\theta-\hat{\theta})p(\theta|\mathcal{D},\mathcal{K})\,\mathrm{d}\theta \end{aligned} \tag{3.4.6}$$

and upon setting the derivative to zero

$$\frac{\mathrm{d}}{\mathrm{d}\hat{\theta}}\left\langle\left|\hat{\theta}-\theta\right|\right\rangle = \int_{-\infty}^{\hat{\theta}} p(\theta|\mathcal{D},\mathcal{K})\,\mathrm{d}\theta - \int_{\hat{\theta}}^{\infty} p(\theta|\mathcal{D},\mathcal{K})\,\mathrm{d}\theta = 0, \tag{3.4.7}$$

results in $\hat{\theta}$ being the median, for which

$$p(\theta > \hat{\theta}|\mathcal{D},\mathcal{K}) = {}^1\!/_2. \tag{3.4.8}$$

So for the absolute loss function $\hat{\theta}$ is the *median* of the distribution. The median only considers errors twice as great to be twice as serious and is less sensitive to the tails of the distribution. Thus it is more *robust* against variation in the tails of the distribution, which is generally held to be a desirable property. It also implies that our estimation is less sensitive to outliers. Thus there is a trade-off between sensitivity and robustness. If we make the non-linear transformation $\phi(\theta)$ and suppose $\phi(\theta)$ is a strictly monotonic increasing function of θ, so that θ is a single-valued and invertible function of ϕ, then clearly

$$\phi(\hat{\theta}) = \hat{\phi}. \tag{3.4.9}$$

In fact all the percentiles have this invariance property, so we can give our point and interval estimates as the median and the interquartile span of the distribution. This is an especially good idea if we believe our parameter should be invariant under monotonic transformations for example in the case of scale parameters.

The final loss function we will consider is,

$$R[\hat{\theta},\theta] = \begin{cases} 0 & \hat{\theta} = \theta \\ 1 & \text{otherwise} \end{cases}, \tag{3.4.10}$$

which is basically a delta function $\delta(\theta - \hat{\theta})$ i.e. it is one for a specific value and zero for the rest. This loss function is a minimum when we choose $\hat{\theta}$ as the most probable value or the *mode* of the posterior distribution.

Using the most probable value of the posterior distribution is intimately connected to the method of maximum likelihood or the approximation of the posterior distribution with a Gaussian distribution (Laplace's method). The most probable set of parameters solves the equations,

$$\frac{\mathrm{d}}{\mathrm{d}\theta_k} \log \frac{L[\boldsymbol{\theta}]\pi[\boldsymbol{\theta}]}{\text{const.}}\bigg|_{\theta_k=\theta_k^*} = 0, \qquad \forall k \tag{3.4.11}$$

or

$$\frac{1}{L[\boldsymbol{\theta}]}\frac{\mathrm{d}L[\boldsymbol{\theta}]}{\mathrm{d}\theta_k} + \frac{1}{\pi[\boldsymbol{\theta}]}\frac{\mathrm{d}\pi[\boldsymbol{\theta}]}{\mathrm{d}\theta_k}\bigg|_{\theta_k=\theta_k^*} = 0, \qquad \forall k. \tag{3.4.12}$$

The prior $\pi(\boldsymbol{\theta})$ is independent of the number of observations J and the $\log L[\boldsymbol{\theta}]$ in general increases like J. Thus if we expand our parameters around the *most probable value of the likelihood function*, $\theta_k = \theta_k^* + \theta'$, the corrections will only be of order $\frac{1}{J}$. In the single variate case, expanding the posterior around the mode of the likelihood function gives a contribution from the prior distribution

$$\pi[\theta] = \pi[\theta^*]\left(1 + \theta'\frac{\pi'(\theta^*)}{\pi(\theta^*)} + \frac{\theta'^2}{2}\frac{\pi''(\theta^*)}{\pi(\theta^*)} + \ldots\right), \tag{3.4.13}$$

and a contribution from the likelihood function

$$\begin{aligned} L[\theta] &= L[\theta^*]\exp\left[-\frac{\theta'^2}{2}\frac{\mathrm{d}^2}{\mathrm{d}\theta^{*2}}\log L[\theta^*] + \frac{\theta'^2}{6}\frac{\mathrm{d}^3}{\mathrm{d}\theta^{*3}}\log L[\theta^*] + \ldots\right] \\ &\propto \exp\left[-\frac{\theta'^2}{2}\frac{\mathrm{d}^2}{\mathrm{d}\theta^{*2}}\log L[\theta^*]\right]\left(1 + \frac{\theta'^3}{6}\frac{\mathrm{d}^3}{\mathrm{d}\theta^{*3}}\log L[\theta^*] + \ldots\right), \end{aligned} \tag{3.4.14}$$

and combining the two contributions we have

$$\begin{aligned} p(\theta|\mathcal{D},\mathcal{K}) \propto &\exp\left[-\frac{1}{2}\left(\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log L[\theta^*]\right)\theta'^2\right] \\ &\times\left(1 + \theta'\frac{\pi'(\theta^*)}{\pi(\theta^*)} + \frac{\theta'^2}{2}\frac{\pi''(\theta^*)}{\pi(\theta^*)} + \frac{\theta'^3}{6}\frac{\mathrm{d}^3}{\mathrm{d}\theta^3}\log L[\theta^*]\right). \end{aligned} \tag{3.4.15}$$

The leading term of this approximation is called the normal form: It is a gaussian centered at the maximum likelihood value with unit variance over the second derivative of the log likelihood function at the maximum likelihood point

$$\text{var}\,(p)\,(\theta|\mathcal{D},\mathcal{K}) = \left[-\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log L[\theta]\right]^{-1}_{\theta=\theta^*}. \tag{3.4.16}$$

Similarly for the multivariate case we can also write the posterior in normal form

$$p(\boldsymbol{\theta}'|\mathcal{D},\mathcal{K}) \propto \exp\left[\frac{\boldsymbol{\theta}'^T\boldsymbol{H}\boldsymbol{\theta}'}{2}\right], \tag{3.4.17}$$

where $\boldsymbol{H}$ is the Hessian of the approximation,

$$\boldsymbol{H} = \left[\begin{array}{cccc} \frac{d^2\log L[\boldsymbol{\theta}]}{d\theta_1^2} & \frac{d^2\log L[\boldsymbol{\theta}]}{d\theta_1 d\theta_2} & \cdots & \frac{d^2\log L[\boldsymbol{\theta}]}{d\theta_1 d\theta_K} \\ \frac{d^2\log L[\boldsymbol{\theta}]}{d\theta_2 d\theta_1} & \frac{d^2\log L[\boldsymbol{\theta}]}{d\theta_2^2} & \cdots & \frac{d^2\log L[\boldsymbol{\theta}]}{d\theta_2 d\theta_K} \\ \vdots & & \ddots & \vdots \\ \frac{d^2\log L[\boldsymbol{\theta}]}{d\theta_K d\theta_1} & \frac{d^2\log L[\boldsymbol{\theta}]}{d\theta_K d\theta_2} & \cdots & \frac{d^2\log L[\boldsymbol{\theta}]}{d^2\theta_K} \end{array}\right]\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}, \tag{3.4.18}$$

and upon normalising

$$p(\boldsymbol{\theta}|\mathcal{D},\mathcal{K}) = \frac{\sqrt{\det\boldsymbol{H}}}{\sqrt{2\pi}^K}\exp\left[\frac{(\boldsymbol{\theta}'-\boldsymbol{\theta}^*)^T\boldsymbol{H}(\boldsymbol{\theta}'-\boldsymbol{\theta}^*)}{2}\right]. \tag{3.4.19}$$

Since $\log p(\theta|\mathcal{D},\mathcal{K}) \sim J$ the variance decreases with growing J and thus the difference $(\boldsymbol{\theta}^* - \boldsymbol{\theta}')$ is of order $J^{-1/2}$. It follows that the corrections $(\theta' - \theta^*)\frac{\pi'(\theta_k^*)}{\pi(\theta_k^*)}$ and $\frac{(\theta_k'-\theta_k^*)^3}{6}$ $\frac{\mathrm{d}^3}{\mathrm{d}\theta_k^3}\log L[\theta_k^*]$ is of order $J^{-1/2}$ which gives the error of the overall approximation as $J^{-1/2}$. This is the method of **Maximum Likelihood**, which has been advocated Fisher. It entails approximating the posterior with a Gaussian distribution. It is also called Laplace's method and interestingly it is parametrisation dependent, see MacKay (1998). It is also extensively used in Bayesian statistics, see Tierney *et al.* (1986).

In summary we have:

- The **Maximum likelihood** as a point estimate and the second derivative of the log-likelihood as its error estimate has the virtue of being easy to calculate. The point estimate is invariant under change of parameters but its error estimate is not because to define an invariant interval requires a prior distribution, which is usually ignored when using the maximum likelihood principal. Notice as well that in general the mode of the posterior will not coincide with the maximum likelihood mode. The point and dispersion estimate contains only local information and can thus be very misleading.

- **Percentiles**: The percentiles are invariant under monotone transformations and contain the prior. However if we are going through the effort of computing percentiles we might as well have plotted the posterior and be done with it. It is usually too much computational effort to provide the percentiles.

- **Posterior moments**: The posterior moments are not naturally invariant under change of parameters, but only invariant under those transformations that leave the prior unchanged. Thus we can view the prior distribution as choosing a class of transformations under which our inference should be invariant. We will explore this idea in more detail later. We also know how to transform posterior moments. If we convolve many posterior distributions together we know that

$$\left\langle \sum_{j=k}^{K} \theta_k \right\rangle = \sum_{k=1}^{K} \langle \theta_k \rangle \tag{3.4.20}$$

and

$$\text{var}\left(\sum_{k=1}^{K} \theta_j\right) = \sum_{k=1}^{K} \text{var}\left(\theta_k\right) + \sum_{i \neq k} \text{cov}\left(\theta_i, \theta_k\right), \tag{3.4.21}$$

where the covariance is defined by

$$\text{cov}\left(\theta_i, \theta_k\right) = \langle (\theta_i - \langle \theta_i \rangle)(\theta_k - \langle \theta_k \rangle) \rangle \,. \tag{3.4.22}$$

So that if we are given the mean and variance instead of the whole posterior we still know how to compute different quantities that depend on our parameters. This property is not shared by the other estimators so we will use the posterior moments and cumulants.

j	x_j	y_j	j	x_j	y_j	j	x_j	y_j	j	x_j	y_j
1	1.420	3.695	2	6.268	6.925	3	8.854	8.923	4	8.532	14.043
5	−5.398	−0.836	6	13.776	16.609	7	5.278	4.405	8	6.298	9.823
9	9.886	12.611	10	11.362	10.174	11	1.964	4.987	12	1.406	6.647
13	0.002	2.873	14	3.212	4.015	15	9.042	10.204	16	1.474	1.953
17	8.528	10.672	18	7.348	9.157	19	6.690	8.552	20	5.796	10.250

Table 3.1: Straight line data set, with $\overline{x} = 5.878$, $\overline{y} = 7.784$, $s_{xx} = 19.332$, $s_{xy} = 17.945$, $s_{yy} = 16.925$

3.5 Straight line fitting: numerical

Let us show the difference between the point estimates by substituting in some numbers in our straight line fitting example. Take the data set from Zellner (1971), shown in Table 3.1 and plotted in Figure 3.1. In Figure 3.2 we show an example comparing the different summary statistics which for this data are as follows.

(a) The Maximum likelihood estimates and corresponding standard deviation are,

$$\begin{aligned} \alpha^* &= 2.893 \quad \sqrt{\text{var}(\alpha^*)} = 0.395 \\ \beta^* &= 0.875 \quad \sqrt{\text{var}(\beta^*)} = 0.056 \\ \sigma^* &= 1.768 \quad \sqrt{\text{var}(\sigma^*)} = 0.280. \end{aligned} \tag{3.5.1}$$

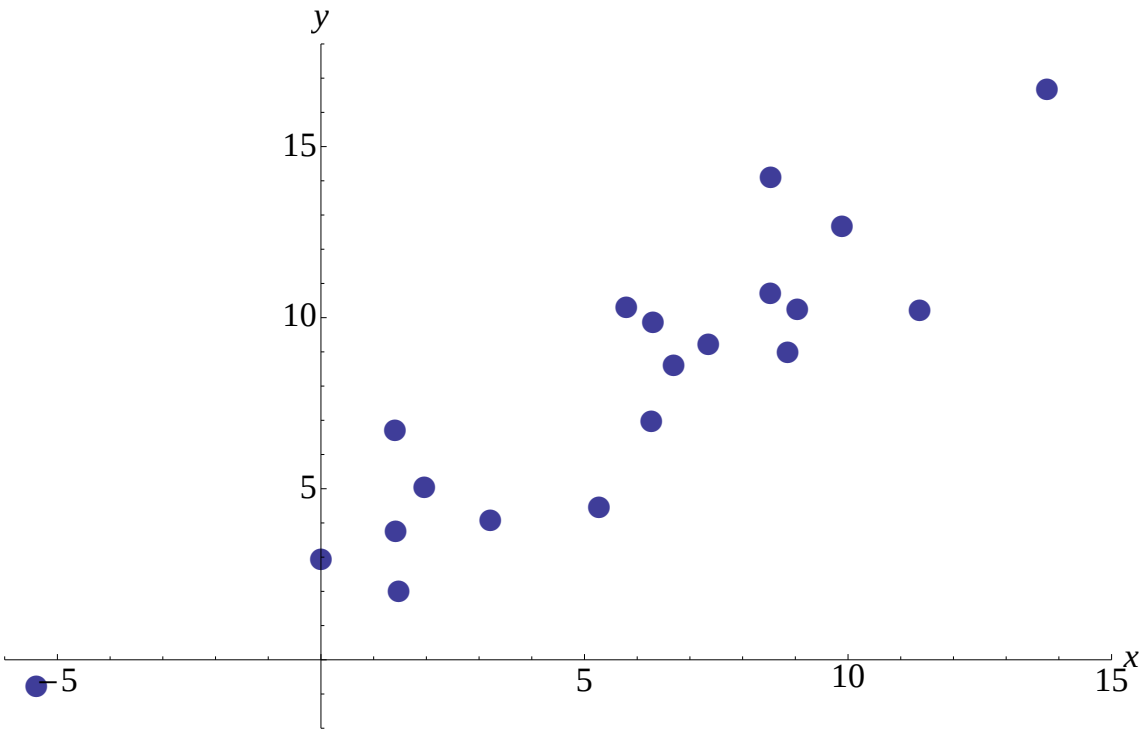


Figure 3.1: Plot of generated data from Zellner (1971), $y = 2 + x$ with $\sigma_x = 4$ and $\sigma_y = 1$.

(b) Median plus percentiles, which we choose to be $\{0.1587, 0.5, 0.8414\}$ in cumulative probability values so that for a Gaussian distribution that is centered on zero with unit variance would give $\{-1, 0, 1\}$ thus making it easier to compare with the other estimators. We obtain

$$\begin{aligned} \alpha &: \{2.43, 2.89, 3.36\} \\ \beta &: \{0.810, 0.875, 0.901\} \\ \sigma &: \{1.70, 1.90, 2.14\}. \end{aligned} \tag{3.5.2}$$

(c) And finally the posterior mean and standard deviation,

$$\begin{aligned} \langle\alpha\rangle &= \overline{y} - \frac{s_{xy}\overline{x}}{s_{xx}} = 2.89 & \sqrt{\mathrm{var}(\alpha)} &= 0.715 \\ \langle\beta\rangle &= \frac{s_{xy}}{s_{xx}} = 0.875 & \sqrt{\mathrm{var}(\beta)} &= 0.101 \\ \langle\sigma\rangle &= \sqrt{\frac{J}{2}\left(s_{yy} - \frac{s_{xy}}{s_{xx}}\right)}\frac{\Gamma\left[\frac{J-3}{2}\right]}{\Gamma\left[\frac{J}{2}-1\right]} = 1.95 & \sqrt{\mathrm{var}(\sigma)} &= 0.347 \end{aligned} \tag{3.5.3}$$

From the Figure we can see that Maximum likelihood gives the worst estimate, the posterior mean is in the middle and the percentiles is the best, but is arguable using three values instead of two as the other estimators. Using a single point estimate is only useful if the posterior is unimodel which will hold for all the posteriors in this dissertation, and if the posterior is unimodal all the estimates are mostly the same.

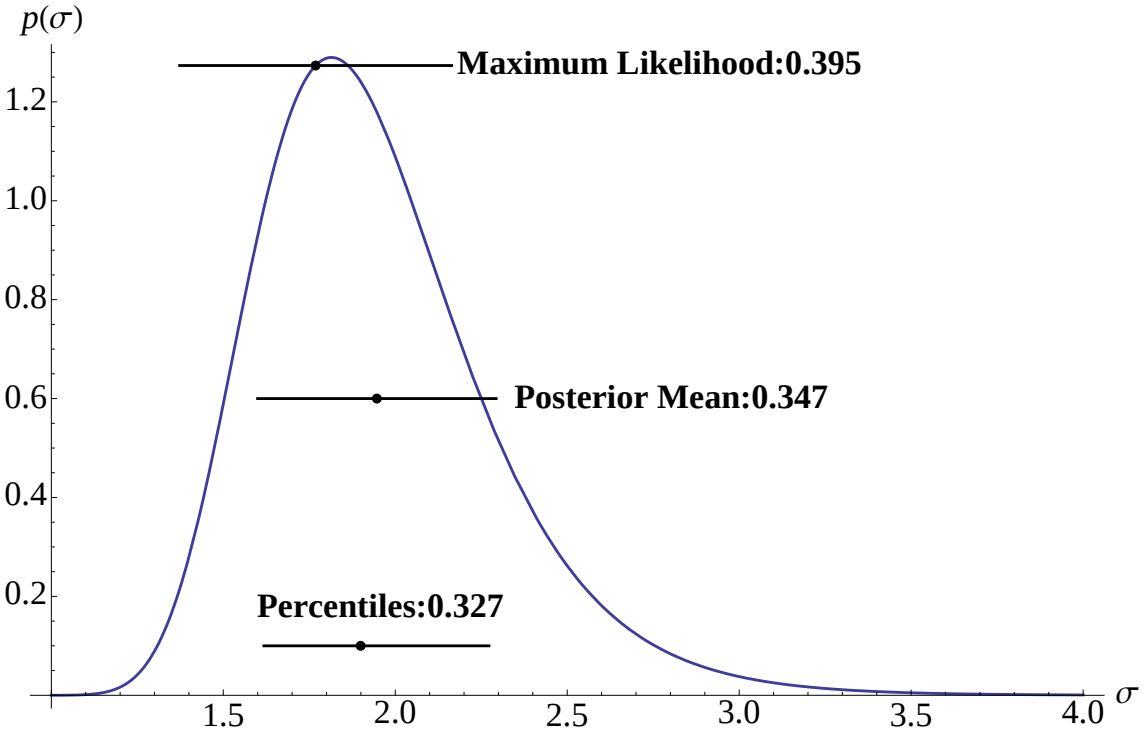


Figure 3.2: Comparison of the point estimate of σ in the straight line example. Maximum likelihood, Posterior mean and the percentiles are plotted with the width of their estimates.

3.6 Predictive distributions

In addition to parameter estimation, we also want to predict future events (in information time). Estimating parameters is only the halfway stop on what we should be doing with hypotheses, the other half being predicting events. So how do we use the posterior for prediction? The answer is simply that the posterior becomes the prior for our prediction. Defining the logical proposition $\mathcal{R}$ as "the next data point", then in general we would have,

$$p(\mathcal{R}|\mathcal{D},\mathcal{K}) = \int p(\mathcal{R}|\boldsymbol{\theta},\mathcal{K})p(\boldsymbol{\theta}|\mathcal{D},\mathcal{K})d\boldsymbol{\theta}. \tag{3.6.1}$$

In the straight line example let

$$\mathcal{R} \equiv \{x', y'\}, \tag{3.6.2}$$

to which we assign the probability,

$$p(\mathcal{R}|\alpha,\beta,\sigma,\mathcal{K}) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(\beta x' + \alpha - y')^2}{2\sigma^2}\right]. \tag{3.6.3}$$

The integral we have to perform is exactly the same as the first evidence integral but with an added data point $x_{J+1} = x'$,$y_{J+1} = y'$, so we define the updated sample moments as

$$\begin{aligned} \overline{x'} &\equiv \frac{J\overline{x} + x'}{1+J} & \overline{y'} &\equiv \frac{J\overline{y} + y'}{1+J} \\ \overline{x'^2} &\equiv \frac{J\overline{x^2} + x'^2}{1+J} & \overline{x'^2} &\equiv \frac{J\overline{y^2} + y'^2}{1+J} & \overline{x'y'} &\equiv \frac{J\overline{xy} + x'y'}{1+J}, \end{aligned} \tag{3.6.4}$$

and the updated sample variances

$$\begin{aligned} s'_{yy} &\equiv \frac{J(J+1)s_{yy} + J(\overline{y}-y')^2}{(J+1)^2} & s'_{xx} &\equiv \frac{J(J+1)s_{xx} + J(\overline{x}-x')^2}{(J+1)^2} \\ s'_{xy} &\equiv \frac{J(J+1)s_{xy} + J(\overline{x}-x')(\overline{y}-y')}{(J+1)^2} & \phi' &\equiv 1 - \frac{{s'_{xy}}^2}{s'_{xx}s'_{yy}}. \end{aligned} \tag{3.6.5}$$

The predictive distribution given the data $\mathcal{D}$ is then the ratio of two evidence terms,

$$p(\mathcal{R}|\mathcal{D},\mathcal{K}) = \frac{p(\mathcal{R},\mathcal{D}|\mathcal{K})}{p(\mathcal{D}|\mathcal{K})} = \frac{\frac{\Gamma\left[\frac{J+1}{2}-1\right]\pi^{1-\frac{J+1}{2}}s'_{yy}\phi}{2\sqrt{s'_{xx}}}\left((J+1)s'_{yy}\phi'\right)^{-\frac{J+1}{2}}}{\frac{\Gamma\left[\frac{J}{2}-1\right]\pi^{1-\frac{J}{2}}s_{yy}\phi}{2\sqrt{s_{xx}}}\left(Js_{yy}\phi\right)^{-\frac{J}{2}}} \tag{3.6.6}$$

In Figure 3.3 the predictive distribution (3.6.6) is plotted together with the posterior mean line. Interestingly, there are no parameters in the predictive distribution since we have taken all possible parameter values and their respective probabilities into account in the predictive distribution. For more about straight line fitting, see Gull (1989) and Press *et al.* (1992).

3.7 Model comparison

How do we make the best predictions that we can? Obviously we find the best model for the data. Intuitively the best model would be the model that assigns the highest probability on average to the data. Bayes Rule in discrete form is simply,

$$p(\mathcal{H}_m|\mathcal{D},\mathcal{K}) = \frac{p(D|\mathcal{H}_m,\mathcal{K})p(\mathcal{H}_m|\mathcal{K})}{p(\mathcal{D}|\mathcal{K})} = \frac{p(D|\mathcal{H}_m,\mathcal{K})p(\mathcal{H}_m|\mathcal{K})}{\sum_m p(D|\mathcal{H}_m,\mathcal{K})p(\mathcal{H}_m|\mathcal{K})}, \tag{3.7.1}$$

where $\mathcal{H}_m$ refers to the mth model instead of a parameter value. Usually there is no reason to prefer any model above any other a priori and thus we choose $p(\mathcal{H}_m|\mathcal{K}) = 1/M$ and we rewrite Bayes rule in its odds form by taking the ratio between two models m and m',

$$\frac{p(\mathcal{H}_m|\mathcal{D},\mathcal{K})}{p(\mathcal{H}_{m'}|\mathcal{D},\mathcal{K})} = \frac{p(\mathcal{H}_m|\mathcal{K})}{p(\mathcal{H}_{m'}|\mathcal{K})}\frac{p(\mathcal{D}|\mathcal{H}_m,\mathcal{K})}{p(\mathcal{D}|\mathcal{H}_{m'},\mathcal{K})}. \tag{3.7.2}$$

As the evidence for the set of models are equivalent they cancel out in the ratio. And because it is very rare to compare models that are indexed continuously we can write,

$$\frac{p(\mathcal{H}_m|\mathcal{D},\mathcal{K})}{p(\mathcal{H}_{m'}|\mathcal{D},\mathcal{K})} = \frac{p(\mathcal{D}|\mathcal{H}_m,\mathcal{K})}{p(\mathcal{D}|\mathcal{H}_{m'},\mathcal{K})}. \tag{3.7.3}$$

So to find the best model all we need to find is the model with the highest evidence in our set of models. This ratio of evidence terms is called the Bayes Factor. Also what we learn from this is that two hypotheses with the same evidence for all data sets $\mathcal{D}$ are in fact the same hypotheses, because they make exactly the same predictions. This has interesting consequences, for example a model with seven parameters that has highly informative prior information available can be equivalent to a model with six parameters but with less prior information available. In section (6.2) below we will show how Bayes factors play out for a concrete example from High Energy Physics.

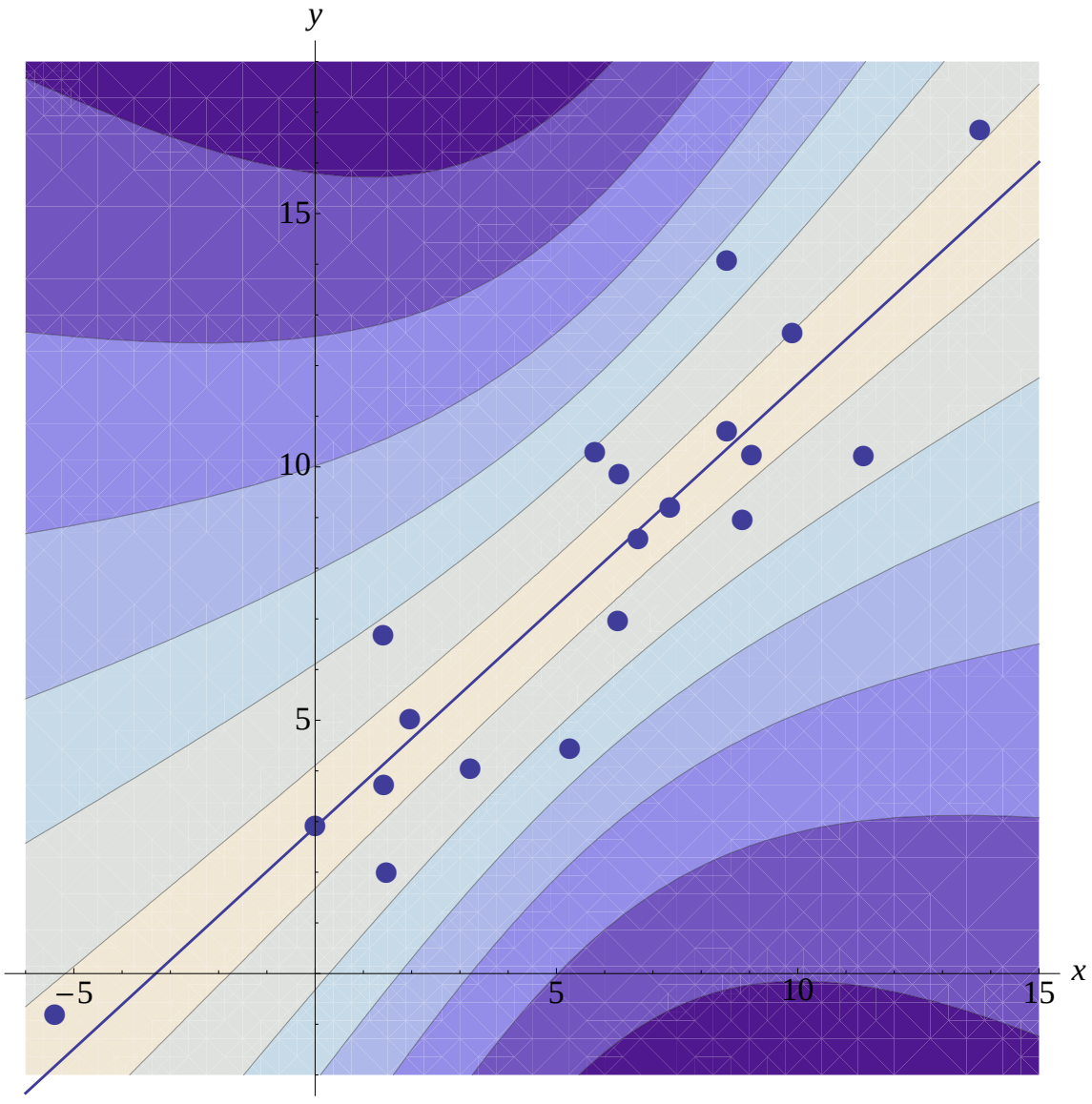


Figure 3.3: A plot of the line $y^* = \alpha x^* + \beta$ with the posterior mean estimates through the data from the straight line together with a contour plot of the logarithm of the probability of the predictive distribution of the next data point.

3.8 Asymptotic Behaviour

The evidence and predictive distributions are the goal in our data analysis. Assuming that we have accumulated lots of logically independent data and that our models are well-conditioned so that the maximum likelihood approximation is valid, our likelihood function tends asymptotically towards a product of delta functions:

$$\lim_{J\to\infty} p(\mathcal{D}|\boldsymbol{\theta},\mathcal{K}) \overset{\text{LI}}{=} \prod_k \delta(\theta_k - \theta_k^*), \tag{3.8.1}$$

where $\boldsymbol{\theta}$ is a set of parameters and $\boldsymbol{\theta}^*$ is the set of 'true' parameters that best describes the data set that we have observed if our model is appropriate. Under this large data limit the evidence is then just an evaluation of the prior for a specific value of $\boldsymbol{\theta}$,

$$\lim_{J\to\infty} p(\mathcal{D}|\mathcal{K}) \overset{\text{LI}}{=} p(\boldsymbol{\theta}^*|\mathcal{K}). \tag{3.8.2}$$

The posterior also tends towards a product of normalised delta functions and the predictive distribution becoming equal to the likelihood function with these "best" parameters,

$$\lim_{J\to\infty} p(\mathcal{R}|\mathcal{D},\mathcal{K}) \overset{\text{LI}}{=} p(\mathcal{R}|\boldsymbol{\theta}^*,\mathcal{K}). \tag{3.8.3}$$

The point of this discussion is that the prior and the likelihood functions are asymptotic forms of the evidence and predictive distributions respectively, which is why we place our focus on the finite forms instead of the asymptotic forms. Next we discuss how we assign probabilities in the first place. As we shall see, our guiding principal is to reason consistently.

Chapter 4

Assigning probabilities using logical independence

> The Probability for an event is the ratio of the number of cases favourable to it, to the number of all cases possible when nothing leads us to expect that any one of these cases should occur more than any other, which renders them, for us, equally possible.
>
> Pierre-Simon, marquis de Laplace

How should a logical proposition be translated into a prediction for future data? In other words, how do we assign probabilities to events? We will take the classical view that the "randomness" in a physical system is caused by our lack of knowledge or control of the system. Thus following Jaynes (2003) and d'Agostini (2003) we prefer to replace "random" with "uncertain" or "uncontrolled". Consequently we will exclude quantum mechanics from the discussion in this dissertation as it might require a physical "randomness". In this picture, classical physics is composed of *systematic* or *controllable* effects which are to be modelled, and *uncontrollable* effects that are not reproducible by the experimental technique and apparatus in use. Often the controllable effects are macroscopic while the uncontrollable ones are microscopic. What is uncontrolled depends on our current scientific expertise and the quality of the apparatus. If one or both improve, previously uncontrollable variables may become controllable and therefore "nonrandom".

In a slightly more mathematical formulation, we may say that the systematic effects may be adequately represented by a small number of *macroscopic variables* while the rest of the variables, usually the large majority, do not matter individually but only through their collective contribution to these few macroscopic variables. The classic example for this situation is, of course, that of statistical physics, where the $6N$ variables of N particles are relevant only insofar as they determine or modify the macroscopic variables of average energy (temperature), heat capacity, susceptibilities etc.

In this chapter, we shall construct a mathematical framework which strongly resembles that of traditional statistical mechanics and indeed encompasses it. Classical statistical

mechanics is the easiest example where all the randomness in a system is caused by our ignorance of the initial conditions of the system. We shall be more careful than most books in doing so, because the building blocks and results that appear in this chapter will form the point of departure or baseline from which later developments in this dissertation will be developed.

There is essentially only one problem in statistical thermodynamics: the distribution of energy E over N particles. In this chapter the number of particles N will be replaced by the number of trials R, the energy of a system E will be replaced by a generalised constraint G, the volume of the system V will play no role in this chapter and the thermodynamic limit ($N \to \infty$, while E/N is kept constant) will be called the *large data limit* or large prediction limit as in Section (3.8) depending on the context. The remarkable thing about this translation is that it changes nothing of the formalism of classical statistical mechanics, which we can then view as a procedure of constructing hypotheses or assigning probabilities.

There is one exception where we encounter conceptual difficulties and that is the idea of distinguishable and indistinguishable particles. We will derive **Maxwell-Boltzmann** statistics by using Logical Independence, which is usually associated with classical particles that have definite trajectories and are thus called **distinguishable** particles. Interestingly, Maxwell-Boltzmann and Bose-Einstein statistics can be derived for both distinguishable and indistinguishable particles, see Constantini (1987). Thus instead of using the physics definition we will define indistinguishable as exchangeable, which implies that there is no ordering of the trials and thus no trajectories. Obviously as we are discussing logical propositions and not particles it would be difficult to introduce trajectories in the first place. We hence assume our system is in "equilibrium" and that we are dealing with indistinguishable classical particles. We will use Chapter 9 of Jaynes (2003) as a basis for this chapter, but we hope to improve on that discussion. The real purpose of this chapter, however, is to show that all of the formalism flows from the logical independence assumption, which is philosophically problematic and which we wish to replace with a more solid foundation.

An important point that we will repeat many times: the framework below is continuous refinement of assigned probabilities; nowhere does data play a role here. All the assignments are based on logical propositions. The remainder of this chapter is structured as follows:

- In Section 4.1, the very high-dimensional "primordial" outcome space S^R is introduced, where each vector $\boldsymbol{x} \in S^R$ represents R distinguishable outcomes and the probability of any $\boldsymbol{x}$ is the same due to the *Principle of Indifference*.

- Section 4.2 explains the Principle of Indifference as a group invariance argument. We also point out that the concept of repetition plays no role when we apply the principle directly to the sample space.

- Section 4.3 traces a first projection from the R-dimensional S^R hypercube space to a one- dimensional partition $\{\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_B\}$ and corresponding probabilities $\boldsymbol{\rho} = \{\rho_1, \rho_2, \ldots, \rho_B\}$.
- Section 4.4 in turn projects this onto the set of *occupation numbers* $\boldsymbol{r} = \{r_1, \ldots, r_B\}$ with outcome space the *universal set* $U(\boldsymbol{r})$ and multinomial probability of (4.4.5).
- Section 4.5 will discuss how these concepts relate to physics.
- In Section 4.6 we develop the concept of waiting-time distributions which indicate that we could have used a different but equivalent normalisation of the sample space. The waiting-time identity itself will be used again later in the development.
- Section 4.7 develops the formalism of updating the probability from $\boldsymbol{\rho}$ to $\boldsymbol{\rho}'$ when we introduce a constraint like a fixed "energy" G.
- Section 4.8 computes the exact predictions from state of knowledge constructed in the previous section.
- Section 4.9 uses the saddlepoint approximation to compute an approximation to the exact formulas of the previous section.
- Section 4.11 introduces the principle of Maximum Entropy which is justified by our saddlepoint approximation.
- Section 4.12 solves three example using the methods developed.
- Section 4.13 connects the Grand Canonical Ensemble with the Principle of Minimum Relative Entropy.
- Section 4.14 discuss the problems with this formalism and what we will do next.

4.1 Primordial sample space

The strategy to formulating a prediction is to enumerate all the possibilities. A *trial* is a measurement or observation of a logical proposition. Each trial by definition will have one of M elementary *outcomes*, which are listed in the *sample space* S. This enumeration of the possible results is a theoretical construction: no event comes labelled with the sample space that should be used. By assumption these elementary outcomes will form an exclusive and exhaustive set of possibilities. Before the data is seen M would also be a prediction, but in the calculation below it is assumed a fixed number, for example $M = 6$ for the rolling of a die and $M = 2$ for the tossing of a coin. If the data contradicts this number our state of knowledge is updated and we redo the calculation.

An *experiment* is defined as R trials where every trial is a question or a measurement that can give the same M possible distinguishable answers, which is a point in a sample

space S. The extended sample space for the whole experiment consists of the direct product of R such spaces,

$$S^R = S \otimes S \otimes \cdots \otimes S. \tag{4.1.1}$$

A single trial in the space S is an outcome and the combination of R outcomes in the space S^R, an *result*. Importantly, "result" and "trial" here do not refer to any data obtained, but to the possible data that could occur in a given S^R. Denote the outcome of the rth trial by the integer x_r with $0 \leq x_r \leq M-1$ and $1 \leq r \leq R$. Any result can be specified by giving the vector $\boldsymbol{x} = \{x_1, \ldots, x_R\}$ and because these outcomes are mutually exclusive and exhaustive, based on our knowledge base $\mathcal{K}$ we need to assign the probability,

$$p(\boldsymbol{x}|\mathcal{K}) = p(x_1, \ldots, x_R|\mathcal{K}), \tag{4.1.2}$$

which is normalised according to

$$\sum_{\boldsymbol{x}} p(\boldsymbol{x}|\mathcal{K}) = \sum_{x_1=0}^{M-1} \cdots \sum_{x_R=0}^{M-1} p(\boldsymbol{x}|\mathcal{K}) = 1. \tag{4.1.3}$$

Note that we have "standardised" our outcome space to the integers $S = \{0, 1, \ldots, M-1\}$. In summary we have defined,

$$\begin{aligned} M &\equiv \texttt{Number of possible elementary results of one trial} \\ R &\equiv \texttt{Total number of trials} \\ x_r &\equiv \texttt{The outcome of trial } r. \end{aligned} \tag{4.1.4}$$

To use the sample space to translate logical statements into probability assignments, however we still need an essential ingredient.

4.2 Principle of indifference

Each of these results $\boldsymbol{x}$ in the sampling space S^R (which is equivalent to the ensemble in statistical mechanics), we will call a *microstate* and each of these states can be labelled uniquely by assigning a unique number to each vector $\boldsymbol{x}$. The number we assign to each vector is in base M and is a combination of the results at each trial. For example if there are 16 elementary results, which are labelled $\{0, \ldots, 9, A, \ldots, F\}$ and we observe the vector $\{A, 1, 3, F\}$ after four trials. The number we assign $\alpha[\boldsymbol{x}]$ could have been anything between 0 and 16^4 and in base 16, between 0000 and $FFFF$. For this specific $\boldsymbol{x}$ we assign $\alpha[\boldsymbol{x}]$ as $A13F$, which is just the natural mapping from a vector to a number. The **Principle of Insufficient Reason** then states that the observer has insufficient information to distinguish between the probability for different numbers he sees. If the observer sees $\alpha[\boldsymbol{x}]$ and $\alpha[\boldsymbol{x'}]$ he learns exactly the same amount, so that

$$p(\alpha[\boldsymbol{x}]|\mathcal{I}) = p(\alpha[\boldsymbol{x'}]|\mathcal{I}), \tag{4.2.1}$$

which will serve as an appropriate starting point for our calculation. The principle was introduced and named by Bernoulli (1713), but according to Keynes (1921), it would have made more sense to call it the 'principle of indifference': The observer is indifferent to the different labellings of the vectors and we would argue that the information available to the observer is *invariant* under the different numbers or permutation of the labels. Note that this is a basic invariance argument for a group transformation which is in this case permutation. The reason why we used the mapping of vectors to unique numbers is to show that there is no need for the concept of repetition in the sampling space when we use the principle of indifference; every number $\alpha[\boldsymbol{x}]$ is an unique object which is given its own probability assignment.

In counting we start from zero and similarly in probability theory we start from ignorance. In other words this describes our initial position of ignorance and our simplest knowledge base,

$\mathcal{I}$: `Each result in the sample space` S^R `is equally likely.`

Since we have a set of *elementary propositions* or microstates that are exclusive and exhaustive,

$$\sum_{\alpha} p(\alpha[\boldsymbol{x}]|\mathcal{I}) = 1. \tag{4.2.2}$$

Consequently there are M^R propositions each with the same probability,

$$p(\boldsymbol{x}|\mathcal{I}) = \frac{1}{M^R} \qquad \forall \boldsymbol{x} \in S^R. \tag{4.2.3}$$

4.3 Combining outcomes

Any compound proposition can be defined as a set of these elementary propositions on which it is true and a complementary set on which it is false. Consider a more complicated proposition where one of B different colours is associated with every elementary outcome. For example outcomes zero, one and two are labelled as red, three, four and five are labelled green and six, seven, eight and nine are labelled blue. Let red correspond to $b = 1$, green to $b = 2$ and blue to $b = 3$. Then by the Principle of Insufficient Reason we have to assign to the result of the first trial,

$$p(x_1|\mathcal{I}) = \frac{1}{10}, \qquad \forall x_1 \in \{0, \ldots, 9\} \quad M = 10, \quad R = 1. \tag{4.3.1}$$

The probability for seeing a red result is then captured in the projection onto variable b

$$\begin{aligned} p(b = 1|\mathcal{I}) &= p(x_1 = 1|\mathcal{I}) + p(x_1 = 2|\mathcal{I}) + p(x_1 = 3|\mathcal{I}) \\ &= \frac{3}{10} \end{aligned} \tag{4.3.2}$$

as intuition had told us already.

In general for a total of M outcomes of one trial: if $\mathcal{A}$ is the set of elementary proposition with $M_{\mathcal{A}}$ elements on which proposition $\mathcal{A}$ is true and false on the rest, then

$$M_{\mathcal{A}} = \sum_{x_1} \delta(x_1 \in \mathcal{A}) \qquad p(\mathcal{A}|\mathcal{I}) = \frac{M_{\mathcal{A}}}{M}. \tag{4.3.3}$$

In general the probability we assign to the vector $\boldsymbol{x}$, where the results are divided into B classes ($S = \mathcal{A}_1 \cup \mathcal{A}_2 \cup \cdots \cup \mathcal{A}_B$ and $\mathcal{A}_i \cap \mathcal{A}_j = \delta_{i,j}\mathcal{A}_i$) and each class contains m_b elementary results ($\sum_b m_b = M$) is

$$p(\boldsymbol{x}|\mathcal{I}) = \prod_{r=1}^{R} \frac{m_{b_r}}{M}, \tag{4.3.4}$$

where b_r is the compound result of the rth trial and m_{b_r} is the number of elementary results that corresponds to it. This statement is so intuitive that this was the original definition of probability used by James Bernoulli and by most writers for the next 150 years as we have seen from the quotation of Laplace. Again note that the sample space does not represent any actual data; we are counting all possibilities (hypothetical/theoretical) when we use the principle of indifference and predicting ratio's of sampling spaces based on our information.

The immediate goal of this construction is to build models, but what are the limitations of this methodology?

- Applying the principle to the sample space requires us to identify a *set of elementary propositions that would be applicable to our results.* In the case of flipping a coin it is easy to justify the use of two elementary equally likely outcomes, but if we measure particles in a complicated physics experiment we know that the results are compound and we do not know how many elementary propositions the results are made out of, only that the compound ones are not equally likely.

- We also assume permutation invariance of the results; thus if time symmetry is broken our models will fail. In other words we are only considering *exchangeable* data measurements and we cannot model any time evolution with this assumption.

4.4 Occupation numbers

Applying the principle of indifference assumes exchangeable events. Realising this can lead to considerable simplification if we recognise the fact that only the number of results $x_r \in \mathcal{A}_b$ is essential information while the order is trivial. Defining the general indicator function,

$$\delta(x_r, \mathcal{A}_b) = \begin{cases} 1 & \text{if } x_r \in \mathcal{A}_b \\ 0 & \text{if } x_r \notin \mathcal{A}_b \end{cases}, \tag{4.4.1}$$

then occupation numbers r_b count the number of times we predict a certain outcome b in a R trials,

$$r_b = \sum_{r=1}^{R} \delta(x_r, \mathcal{A}_b). \tag{4.4.2}$$

The assumption is that the ordering of the data is unimportant, so we can discard that information. Applying it to our representation of the equally likely microstates, removes the ordering of the vector and as we know this introduces factorials for all the permutations, leading to

$$\begin{aligned} p(\boldsymbol{r}|R,\mathcal{I}) &= \sum_{\boldsymbol{x}} p(\boldsymbol{x}|R,\mathcal{I}) \prod_b \delta\left[r_b - \sum \delta(x_r, \mathcal{A}_b)\right] \\ &= R! \prod_b \frac{1}{r_b!} \left(\frac{m_b}{M}\right)^{r_b}, \end{aligned} \tag{4.4.3}$$

where $\boldsymbol{r} = \{r_b\}_{b=1}^{B}$. Defining ρ_b as the ratio of the outcome space associated with the compound result b,

$$\rho_b = \frac{m_b}{M}, \quad 0 \leq \rho_b \leq 1, \quad \sum_b \rho_b = 1, \tag{4.4.4}$$

the probability assignment on our sample space is a **multinomial distribution**,

$$p(\boldsymbol{r}|\boldsymbol{\rho}, R, \mathcal{I}) = R! \prod_b \frac{\rho_b^{r_b}}{r_b!}. \tag{4.4.5}$$

Importantly we consider the set $\boldsymbol{\rho} = \{\rho_b\}$ to be both a probability and a parameter set of $\mathcal{I}$. Of course if we change our sampling space we must change the ratios ρ_b. Also notice that the $\boldsymbol{\rho}$ is a vector of *chances*, the probability assignment $p(\boldsymbol{r}|\boldsymbol{\rho}, R, \mathcal{I})$ factorises and the Principle of Indifference applied directly to the sample space is thus equivalent to Logical Independence.

4.5 Discussion

Consider (4.4.4) again,

$$\rho_b = \frac{m_b}{M}, \tag{4.5.1}$$

which is what the principle of indifference boils down to. This is our answer to the question of how to assign probabilities, but remember that this is a certain state of knowledge indicated by our knowledge base $\mathcal{I}$. Historically, according to Jeffreys (1961), there have been three different attempts to answer this question,

- If there are M possible alternatives, for m of which $\mathcal{A}$ is true, then the probability of $\mathcal{A}$ is defined to be $\frac{m}{M}$.

- If a proposition is true a large number of times, then the probability of $\mathcal{A}$ is the limit of the ratio of the number of times when $\mathcal{A}$ will be true to the whole number of trials, when the number of trials tends to infinity.

- An actual infinite number of trials is assumed. Then the probability of $\mathcal{A}$ is defined as the ratio of the number of cases where $\mathcal{A}$ is true to the whole number.

The first definition is from de Moivre (1718) and is used by Laplace (1812) (as in our opening quotation) and in modern works such as Neyman (1937). In these works it is usually introduced as a definition of a probability for an event. Introducing this statement as an additional axiom is unnecessary because it is a Logical proposition which identifies the set of elementary propositions. It is not a definition but a consequence of using probability as a degree of belief. As we have derived this definition from first principles we will not criticize its use too heavily. It has the definite advantage that it gives a numerical answer, but it might not be what we want at all.

At the very least, the sample spaces need to be constructed carefully, with different situations resulting in different spaces. Consider for example two boxes, one filled with a black ball and a white ball and the other with two black balls and a white ball. A box is chosen at random and then a ball is randomly drawn from that box. What is the probability of seeing a white ball? Simply lumping all five balls into a single outcome space would yield the easy and wrong answer, $p(\mathcal{A}|\mathcal{K}) = \frac{2}{5}$, but it is allowed if we think of the first statement as a definition. The correct approach would be to interpret that 'at random' implies that our state of knowledge is indifferent at that point thus applying the product rule twice and then the sum rule. In this case we find $p(\mathcal{A}|\mathcal{I}) = \frac{1}{2}\cdot\frac{1}{2} + \frac{1}{2}\cdot\frac{1}{3} = \frac{10}{24}$. Also if we have a biased dice this definition would be impossible to use as there is no equally likely elementary case. Here we have to point out that 'at random' is not a physical process but that we do not know which box or which ball will be chosen and is thus part of the information we have available. Some definition is written down and then never actually used. When it is applied probability is used as a degree of belief. There are only different assignments based on different knowledge bases, but the definition of probability stays the same. Thus we disagree with most of the literature (see James (2006)) that there are different definitions of probabilities.

The second proposal is the Venn limit or relative frequency used by von Mises (1939) and the third is the infinite population associated with Fisher (1922*b*). These proposals we will also not criticize too heavily as we will introduce an infinite population later as a limit of a finite one. But we have to point out that no probability has ever been assigned by counting an infinite number of trials or finding the limiting ratio of two infinite series, thus instead of giving a wrong answer, they give no answer at all. These are extraneous axioms that add nothing to the theory. All three these statements are different prior distributions and making any one of them a definition limits the applications of probability theory.

We believe that we can associate a fixed probability to each outcome and thus that the observation of a trial (or particle) does not teach us anything about future or past trials(or

other particles). In a sense we believe that the system we observe is an **ordered** system, and that these probabilities are fixed and varying initial conditions leads to the same probabilities. This is just the assumption of logical independence in different words, but this assumption is the major weakness of this analysis. So for every system to we apply this formalism we have to check this assumption as well as the absence of time evolution. To check the assumption of Logical Independence we obviously need a alternative hypotheses so that we can work out a Bayes Factor and see which assumption is preferred.

Summarising what we have done so far:

- Translating our logical statements into probabilities required the construction of a sampling space and applying the principle of indifference to it. The result is a multinomial distribution (Maxwell-Boltzmann statistics) for any prediction of discrete compound events.

- The sample space is a theoretical construct that describes the information available to the observer. The physical system itself is not part of a "random" process, because the random effects are caused by the lack of knowledge of the observer. Even if the future results are completely deterministic, the observer must still interpret the process as random if he is unable to predict these results in full. Unknown processes are thus synonymous with a random process.

- The sample space does not require the concept of repetition if we apply the principle of insufficient reason to it, because we consider every outcome as an unique event and even if the sample space leads us to a multinomial this is only a compression of the labelling of the sample space.

- The multinomial does not change its probabilities based upon previous results, because it represents a fixed set of chances on the sample space that counts result ratios as defined by the set $\{\mathcal{A}_b\}$ and is thus not connected to the data that we observe.

4.6 Stopping rules

We also have to mention that it is not necessary to condition the sample space on R, the total number of trials. There is an equivalent formulation where we condition on k, the number of occurrences of a specific outcome b. In the literature this is called a stopping rule: we usually stopped after conducting R trials, now we stop when we see k occurrences of a specific outcome. Notice that this marks a specific outcome as special and different from the others and we lose the symmetry that a multinomial has. Consider a binomial distribution (multinomial with only two outcomes) with probability ρ for success and r successes in R trials,

$$p(r|R,\rho,\mathcal{I}) = \binom{R}{r}\rho^r(1-\rho)^{R-r}. \tag{4.6.1}$$

The probability that we will see fewer than k successes in R trials is equal to the probability that it would take more than R trials to achieve k successes, thus

$$\sum_{r<k} p(r|R,\rho,\mathcal{I}) = \sum_{r>R} p(r|k,\rho,\mathcal{I}), \tag{4.6.2}$$

where R, k, ρ are fixed on both sides. Substituting our binomial distribution gives

$$\begin{aligned}\sum_{r<k} p(r|R,\rho,\mathcal{I}) &= \sum_{r=0}^{k-1} \binom{R}{r} \rho^r (1-\rho)^{R-r} \\ &= \binom{R}{k-1} \rho^{k-1}(1-\rho)^{R-k+1} + \binom{R}{k-2} \rho^{k-2}(1-\rho)^{R-k+2} + \cdots \\ &= \binom{R}{k-1} \rho^{k-1}(1-\rho)^{R-k+1} \sum_{j=0}^{k-1} \frac{(k-1)^{\underline{j}}}{(R-k+2)^{\overline{j}}} \left(\frac{1-\rho}{\rho}\right)^j .\end{aligned} \tag{4.6.3}$$

Using the properties of falling factorials (2.1.10)

$$\sum_{j=0}^{k-1} (k-1)^{\underline{j}} = \sum_{j=0}^{\infty} (k-1)^{\underline{j}} \tag{4.6.4}$$

the definition of a Gaussian Hypergeometric function (A.0.1) and remembering $1^{\overline{k}} = k!$, we have

$$\begin{aligned}\sum_{j=0}^{\infty} \frac{(k-1)^{\underline{j}}}{(R-k+2)^{\overline{j}}} \left(1-\frac{1}{\rho}\right)^j &= \sum_{j=0}^{\infty} \frac{(1-k)^{\overline{j}}}{(R-k+2)^{\overline{j}}} \left(1-\frac{1}{\rho}\right)^j \\ &= {}_2F_1 \left[\begin{matrix} 1, 1-k \\ R-k+2 \end{matrix} \middle| 1-\frac{1}{\rho} \right]\end{aligned} \tag{4.6.5}$$

and taking the first Euler transformation (A.0.5),

$$\begin{aligned}{}_2F_1 \left[\begin{matrix} 1, 1-k \\ R-k+2 \end{matrix} \middle| 1-\frac{1}{\rho} \right] &= \rho \cdot {}_2F_1 \left[\begin{matrix} 1, R+1 \\ R-k+2 \end{matrix} \middle| 1-\rho \right] \\ &= \rho \sum_{j=0}^{\infty} \frac{(R+1)^{\overline{j}}}{(R-k+2)^{\overline{j}}} (1-\rho)^j .\end{aligned} \tag{4.6.6}$$

Using (4.6.2) and after some manipulation, we find

$$\sum_{r=R+1}^{\infty} p(r|k,\rho,\mathcal{I}) = \sum_{r=R+1}^{\infty} \frac{(r-1)!\rho^k(1-\rho)^{r-k}}{(k-1)!(r-k)!}. \tag{4.6.7}$$

which yields the **negative binomial distribution**,

$$p(r|k,\rho,\mathcal{I}) = \frac{(r-1)!\rho^k(1-\rho)^{r-k}}{(k-1)!(r-k)!} = \binom{r-1}{k-1} \rho^k (1-\rho)^{r-k}, \tag{4.6.8}$$

which is the probability of requiring r independent trials for k successes, where r runs from k to infinity. With $r' = r - k$ the negative binomial can also be written as,

$$p(r'|k,\rho,\mathcal{I}) = \binom{r'+k-1}{k-1} \rho^k (1-\rho)^{r'} = (-1)^{k-1} \binom{-r'}{k-1} \rho^k (1-\rho)^{r'}. \tag{4.6.9}$$

4.7 Constraints for independent trials

Consider what happens if we add information to our multinomial state of knowledge described so far. The information that a scientist must translate into a model can come in many different forms. Traditionally scientists are interested in using symmetries (conservation) for model construction. These symmetries are constraints: they exclude certain cases from happening. Using our multinomial sampling space, we will try to build models with some linear function of the occupation numbers constrained. While there are many possible ways we could have constrained the sampling space, we will consider one mathematically similar to the microcanonical ensemble with a generalised energy; see for example Schrödinger (1952) and Jaynes (2003). We will use the method of average means to prove our results also called the Darwin-Fowler method, see Darwin and Fowler (1922*a*) and Darwin and Fowler (1922*b*) combined with the approach of Khinchin (1949). The method is basically to take the generating function of the constraints and then apply the saddle-point approximation to it i.e. the bread and butter calculation in statistical mechanics. The genius of Khinchin is to work out analytically as evidence ratios the predictions in which we are interested. Then we only need to apply the approximation once to find the solutions.

There are also two alternative methods for computing the same formulas, namely the method of most probable distributions due to Boltzmann and the method of Maximum Entropy used by Jaynes. The reason why we use the Darwin-Fowler method is that it also gives us an error estimate in the same calculation, which constitutes in this case a proof of the theorem of large numbers. The other two methods must be supplemented by additional calculations to supply the same result. After we have derived the main result we will discuss the other two methods and some additional results that we need later on.

Associate a fixed number g_b with every result x_r falling into $\mathcal{A}_b$. The set of numbers g_b is given by the physical model; privately we imagine this to be an energy, but formally it can be any linear functional of the results. If we demand that the average of this physical quantity remain the same in every trial then the total is also constrained. Defining the total value of the experiment as

$$G = \sum_{r=1}^{R} g_{b_r} = \sum_{b=1}^{B} g_b r_b. \tag{4.7.1}$$

This constraint of a fixed G, encoded as $\delta(G - \sum_b g_b r_b)$ restricts our sampling space and changes the occupation numbers that we expect to observe. We shall use the notation of $\boldsymbol{\rho}$ as the vector of prior probabilities, that represent a state of ignorance without constraint and the posterior vector $\boldsymbol{\rho}'$ as the posterior probabilities that incorporates constraint G.

Working out the average occupation number of a multinomial without the G-constraint gives,

$$\langle r_c \rangle = \sum_{U(\boldsymbol{r})} r_c R! \prod_b \frac{\rho_b^{r_b}}{r_b!} = R\rho_c, \tag{4.7.2}$$

While for our posterior probabilities we will use,

$$\rho'_c = \frac{\langle r_c|G\rangle}{R}. \tag{4.7.3}$$

We now compute the new average occupation number $\langle r_b|G\rangle$ in every bin given that our constraint G is known. Starting from the joint probability of our restriction and microscopic states (4.4.5),

$$p(G,\boldsymbol{r}|R,\boldsymbol{g},\boldsymbol{\rho},\mathcal{I}) = p(G|\boldsymbol{r},\boldsymbol{g},\mathcal{I})p(\boldsymbol{r}|R,\boldsymbol{\rho},\mathcal{I}) = \delta\left(G-\sum_b g_b r_b\right) R!\prod_b \frac{\rho_b^{r_b}}{r_b!}, \tag{4.7.4}$$

where we used the fact that $p(G|\boldsymbol{r},\mathcal{I}) = \delta\left(G-\sum_b g_b r_b\right)$ i.e. knowing all the occupation numbers implies that we know exactly the value of G. To keep the notation sane we will consider the vector $\boldsymbol{g}$ to be part of the knowledge base $\mathcal{I}$. Starting with the evidence for G,

$$p(G|R,\boldsymbol{\rho},\mathcal{I}) = \sum_{U(\boldsymbol{r})} p(G,\boldsymbol{r}|R,\boldsymbol{\rho},\mathcal{I}) = \sum_{U(\boldsymbol{r})} \delta\left(G-\sum_b g_b r_b\right) R!\prod_b \frac{\rho_b^{r_b}}{r_b!}, \tag{4.7.5}$$

we find the posterior (3.2.10),

$$p(\boldsymbol{r}|G,R,\boldsymbol{\rho},\mathcal{I}) = \frac{p(G,\boldsymbol{r}|R,\boldsymbol{\rho},\mathcal{I})}{p(G|R,\boldsymbol{\rho},\mathcal{I})}. \tag{4.7.6}$$

and estimate the mean occupation from it,

$$\langle r_c|G\rangle = \sum_{U(\boldsymbol{r})} r_c\, p(\boldsymbol{r}|G,\boldsymbol{\rho},R,\mathcal{I}). \tag{4.7.7}$$

The goal is then to take the limit $R\to\infty$, which will then give ρ' as our prior distribution for the multinomial. We also need to show that the variance in ρ' goes to zero in this limit so that eq. (3.8.1) holds. This large prediction limit is called the "Thermodynamic" limit and we will keep our parameter,

$$\gamma = \frac{G}{R}, \tag{4.7.8}$$

which is defined as a ratio fixed. The evidence for this constraint,

$$\Omega(R,G) = p(G|R,\boldsymbol{\rho},\mathcal{I}) \tag{4.7.9}$$

is called the **structure function** by Khinchin (1949); below, we will show how ρ'_c and its variance can be written as ratios of the structure function. Taking our joint probabilities (4.7.4) and computing the joint moment generating function of G as in Section 2.4,

$$\Phi\left[\lambda,G|p(G,\boldsymbol{r}|R,\boldsymbol{\rho},\mathcal{I})\right] = \sum_G p(G,\boldsymbol{r}|R,\boldsymbol{\rho},\mathcal{I})e^{-\lambda G} = R!\prod_b \frac{\rho_b^{r_b}e^{-\lambda g_b r_b}}{r_b!}. \tag{4.7.10}$$

Using the multinomial theorem (2.1.13) to sum over the universal set $U(\boldsymbol{r})$ i.e. over all $r_b \geq 0$ that add up to R, we find the moment generating function for $p(G|R,\boldsymbol{\rho},\mathcal{I})$

$$\begin{aligned}\Phi\left[\lambda, G|p(G|R,\boldsymbol{\rho},\mathcal{I})\right] &= \sum_{U(\boldsymbol{r})} \Phi\left[\lambda, G|p(G,\boldsymbol{r}|R,\boldsymbol{\rho},\mathcal{I})\right] \\ &= \sum_{U(\boldsymbol{r})} R! \prod_b \frac{\rho_b^{r_b} e^{-\lambda g_b r_b}}{r_b!} = \left(\sum_b \rho_b e^{-\lambda g_b}\right)^R\end{aligned} \tag{4.7.11}$$

and so we can rewrite (4.7.9) in the form

$$\left(\sum_b \rho_b e^{-\lambda g_b}\right)^R = \sum_G \Omega\left[G,R\right] e^{-\lambda G}. \tag{4.7.12}$$

In general, $\Omega\left(G,R\right)$ can be found by inverting this equation using (2.4.8) but in some simple cases it may be possible to read it off as the coefficient of $e^{-\lambda G}$. It is customary to define the **canonical partition function** as,

$$Z[\lambda] = \sum_b \rho_b e^{-\lambda g_b}. \tag{4.7.13}$$

The structure function $\Omega(G,R)$ gives the probability for finding a certain value of G in R trials. The partition function $Z[\lambda]$ on the other hand is the generating function for the probability that a single result contributed to that G and the R fold convolution of the partition function partitions the total G into the individual results i.e. result $b=1$ happened r_b times and so forth, hence the name partition function. The contributions of every trial is logically independent from every other trial and thus we end up with a product of a single trial distribution. To see the individual trial contributions we can for $R=1$ invert the partition function

$$p(G|R=1,\mathcal{I}) = \sum_b \rho_b\, \delta(G-g_b), \tag{4.7.14}$$

and for $R>1$, we convolve R such distributions, recursively

$$p(G|R,\mathcal{I}) = \sum_b \rho_b\, p(G-g_b|R-1,\mathcal{I}). \tag{4.7.15}$$

For small values of R we can use a computer to evaluate this sum directly but for our model construction we are more interested in the large R limit.

4.8 Predictions using the structure function

Starting from the definition of the occupation number moments

$$\begin{aligned}\langle r_c|G\rangle &= \sum_{U(\boldsymbol{r})}^{R} r_c \frac{p(\boldsymbol{r},G|R,\boldsymbol{\rho},\mathcal{I})}{p(G|R,\boldsymbol{\rho},\mathcal{I})} \\ &= \frac{1}{\Omega(R,G)} \sum_{U(\boldsymbol{r})}^{R} r_c R! \prod_b \frac{\rho_b^{r_b}}{r_b!} \delta\left[G - \sum_{b=1}^{B} g_b r_b\right],\end{aligned} \tag{4.8.1}$$

we absorb the r_c factor into the factorial, to get

$$\begin{aligned}\langle r_c|G\rangle &= \frac{R\rho_c}{\Omega(R,G)}\sum_{U(\boldsymbol{r})}^{R-1}(R-1)!\prod_b\frac{\rho_b^{r_b}}{r_b!}\delta\left[G-g_c-\sum_{b=1}^{B}g_b r_b\right]\\ &= R\rho_c\frac{\Omega(R-1,G-g_c)}{\Omega(R,G)}\end{aligned} \tag{4.8.2}$$

or

$$\rho_c' = \frac{\langle r_c|G\rangle}{R} = \rho_c\frac{\Omega(R-1,G-g_c)}{\Omega(R,G)}. \tag{4.8.3}$$

From (4.7.15) we can see that ρ_c' is correctly normalised, since

$$\sum_c \rho_c\Omega(R-1,G-g_c) = \Omega(R,G). \tag{4.8.4}$$

Using a similar construction we can prove the following collection of formulas:

$$\boxed{\begin{aligned}\langle r_c|G\rangle &= R\rho_c\frac{\Omega(R-1,G-g_c)}{\Omega(R,G)}\\ \langle r_c r_d|G\rangle &= R^2\rho_c\rho_d\frac{\Omega(R-2,G-g_c-g_d)}{\Omega(R,G)} \qquad c\neq d\\ \langle r_c^2-r_c|G\rangle &= R(R-1)\rho_c^2\frac{\Omega(R-2,G-2g_c)}{\Omega(R,G)}.\end{aligned}} \tag{4.8.5}$$

and hence the width in any given component r_b of $p(\boldsymbol{r}|G,I)$ divided by its mean

$$\frac{[\langle r_b^2|G\rangle-\langle r_b|G\rangle^2]^{1/2}}{\langle r_b|G\rangle} = \left(\frac{\Omega(R,G)\,\Omega(R-2,G-2g_b)}{\Omega^2(R-1,G-g_b)}-1\right)^{1/2}+\mathcal{O}(R^{-1/2}) \tag{4.8.6}$$

is small compared to R. The width of $p(\boldsymbol{r}|G,I)$ around its peak at $\langle r_b|G\rangle/R$ for large R is also so narrow that we can reasonably approximate it by

$$p(\boldsymbol{r}|\boldsymbol{\rho},G,\mathcal{I}) = \prod_b\delta(r_b-R\rho_b'). \tag{4.8.7}$$

where from (4.8.2)

$$\rho_b' = \rho_b\,\frac{\Omega(R-1,G-g_b)}{\Omega(R,G)} \tag{4.8.8}$$

also becomes asymptotically independent of R and G. In textbooks, this formula normally reads, Band (1955) and Grandy (1987):

$$\langle r_c|G\rangle = -\frac{1}{\lambda}\frac{\mathrm{d}}{\mathrm{d}g_c}\log Z[\lambda]. \tag{4.8.9}$$

To show that these formulae are equivalent we first have to apply the saddlepoint approximation to the former.

4.9 Saddlepoint approximation for independent trials

For large R it is not feasible to take λ derivatives of the Rth power of the partition function that (4.7.12) requires. Luckily we can replace the derivatives with complex integration, (2.4.8), which is much easier to approximate using the saddlepoint approximation as shown in section 2.6. First we define the saddlepoint of λ, the dual variable of the constraint, to be at λ^* the point where the logarithmic derivative of the generating function is equal to the parameter γ,

$$\frac{d}{d\lambda}\log Z[\lambda]\bigg|_{\lambda=\lambda^*} = -\frac{G}{R} = -\gamma. \tag{4.9.1}$$

From the properties of the generating function we know that this equation has one unique solution. The approximate solution is given by (2.6.12) with $n = R$, $x = \gamma$ and $K[\lambda^*] = \log Z[\lambda^*] + \gamma\lambda^*$,

$$\Omega(G,R) \approx \frac{Z^R[\lambda^*]\, e^{G\lambda^*}}{\sqrt{R2\pi V[\lambda^*]}}, \tag{4.9.2}$$

where

$$K''[\lambda^*] = V[\lambda^*] = \frac{d^2\log Z[\lambda]}{d\lambda^2}\bigg|_{\lambda=\lambda^*} = \frac{Z''[\lambda^*]}{Z[\lambda^*]} - \frac{Z'[\lambda^*]^2}{Z[\lambda^*]^2}. \tag{4.9.3}$$

Generally the saddlepoint approximation will destroy the normalisation of our structure function but this does not matter as the mean occupation number (4.8.2),

$$\langle r_c|G\rangle = R\rho_c\frac{\Omega(R-1,G-g_c)}{\Omega(R,G)}, \tag{4.9.4}$$

will be unaffected. Applying our saddlepoint approximation to the structure function we find,

$$\rho_c' = \frac{\langle r_c|G\rangle}{R} = \frac{\rho_c e^{-g_c\lambda^*}}{Z[\lambda^*]} \tag{4.9.5}$$

and

$$\langle r_c^2|G\rangle - \langle r_c|G\rangle^2 = R\left[\frac{e^{-g_c\lambda^*}}{Z[\lambda^*]} - \frac{e^{-2g_c\lambda^*}}{Z^2[\lambda^*]}\right] \tag{4.9.6}$$

or

$$\left\langle {\rho_c'}^2\right\rangle - \left\langle \rho_c'\right\rangle^2 = \frac{1}{R}\left[\frac{e^{-g_c\lambda^*}}{Z[\lambda^*]} - \frac{e^{-2g_c\lambda^*}}{Z^2[\lambda^*]}\right], \tag{4.9.7}$$

which tends to 0 as $R \to \infty$, which indicates that the fluctuations do indeed vanish in the thermodynamic limit and we have a well defined likelihood function. The same answer (4.9.5) results from the g_c-derivative prescription (4.8.9), so we conclude that textbooks usually derive the asymptotic version of the prediction formula and not the exact version. This is the standard method for constructing models given a single parameter G. Obviously this method can easily be generalised to deal with a set of parameters $\boldsymbol{\theta}$. In statistical mechanics this method is called the Darwin-Fowler method or the method of mean values. Its primary advantage is that it gives both the prediction and the accuracy of that prediction as embodied by (4.9.7).

4.10 Method of most probable distribution

There is also a simpler method for deriving exactly the same solution, namely the method of the most probable distribution. It is a more direct method but it yields less than the method of mean values as it provides only the most probable value but not its accuracy. Considering again our partition function

$$\sum_{U(\boldsymbol{r})} R! \prod_b \frac{\rho_b^{r_b}}{r_b!} e^{-\lambda g_b r_b} = Z[\lambda]^R, \tag{4.10.1}$$

we know from the method of mean values that the fluctuations die down in the limit $R \to \infty$ and thus there is a single term that contributes almost all the probability. Thus we replace the sum with the single most probable set of $\boldsymbol{r}$ which we label $\boldsymbol{r}^*$. This most probable set will also be identical with the set of expectation values $\langle r_c|G\rangle$ because in the limit the most probable and the average will coincide and both will obey the constraint, $G = \sum_b g_b r_b^* = \sum_b g_b \langle r_b|G\rangle$. Furthermore, all this will be true for a single value of λ, which with a moments thought we realise is the saddlepoint. We hence find directly

$$R! \prod_b \frac{\rho_b^{r_b^*}}{r_b^*!} e^{-\lambda^* g_b r_b^*} = Z[\lambda^*]^R. \tag{4.10.2}$$

We cannot however, vary the set $\{r_b\}$ independently as the sum may not add up to R. We must incorporate an extra Lagrange multiplier μ to enforce the extra constraint, following which we can vary each to find the most probable r_c

$$\nabla_{r_c} \log \left[\prod_b \frac{\rho_b^{r_b}}{r_b!} e^{-\mu r_b - \lambda g_b r_b} \right] = 0 \quad \forall c, \tag{4.10.3}$$

where $\nabla_{r_c} f[r_c] = f[r_c] - f[r_c - 1]$. The resulting set of formulas,

$$\log\left(\frac{\rho_c}{r_c^*}\right) - \mu - g_c\lambda = 0, \quad \forall c \tag{4.10.4}$$

has the solution

$$r_c^* = \rho_c \exp[-\mu^* - g_c\lambda^*], \tag{4.10.5}$$

where μ^* and λ^* are determined by the constraint equation

$$\sum_c r_c^* = \sum_c \rho_c \exp[-\mu^* - g_c\lambda^*] = R \tag{4.10.6}$$

or

$$Z[\lambda^*] = \sum_c \rho_c \exp[-g_c\lambda^*] = R \exp[\mu^*], \tag{4.10.7}$$

giving

$$\rho'^*_c = \frac{r_c^*}{R} = \frac{\rho_c \exp[-g_c\lambda^*]}{Z[\lambda^*]} \tag{4.10.8}$$

while λ^* is determined from $\gamma = \frac{G}{R} = \sum_c g_c \rho'_c$

$$\sum_c g_c \frac{\rho_c \exp[-g_c\lambda^*]}{Z[\lambda^*]} = \gamma. \tag{4.10.9}$$

4.11 Principle of Minimum Relative Entropy

A third derivation is based on a generalisation of the well-known Principle of Maximum Entropy: for a general set $\{\rho'_b\}$, the Shannon Entropy $H[\boldsymbol{\rho}'] = -\sum_b \rho'_b \log \rho'_b$ is maximised with undetermined Lagrange multipliers λ_0 for the normalisation constraint and λ for any other constraint formulated as $I(\boldsymbol{\rho}')$,

$$\delta\left[-\sum_b \rho'_b \log \rho'_b - \lambda_0 \sum_b \rho'_b - \lambda I(\boldsymbol{\rho}')\right] = 0 \tag{4.11.1}$$

thereby determining $\boldsymbol{\rho}'^* = \{\rho'^*_b\} = \{r^*_b/R\}$. Starting from the fixed-G constraint $p(G|\boldsymbol{r},\mathcal{I}) = \delta(G - \sum_b g_b r_b)$, many textbooks derive from this the Boltzmann probability $\rho'_b = e^{-\lambda g_b}/Z[\lambda]$ which corresponds to (4.9.5) for the special case of constant $\rho_b = 1/B$. In this section, we show how the above Principle of Maximum Entropy is a special case of a more general "Principle of Minimum Relative Entropy"

The principle of minimum relative entropy is equivalent to the method of most probable distribution when it is realised that the most probable distribution is contingent on the limit $R \to \infty$. Applying the limit first on the multinomial yields,

$$\lim_{R\to\infty} R! \prod_b \frac{\rho_b^{r_b}}{r_b!} = \exp\left[-R\sum_b \rho'_b \log\frac{\rho'_b}{\rho_b}\right], \tag{4.11.2}$$

with $\rho'_b = \frac{r_b}{R}$. So instead of trying to find the set of $\{r_b\}$ that has the maximum probability we can look for the set of ρ'_b that minimizes the Kullback-Leibler divergence or relative entropy,

$$H[\boldsymbol{\rho}'\|\boldsymbol{\rho}] \equiv \sum_b \rho'_b \log\frac{\rho'_b}{\rho_b}, \tag{4.11.3}$$

while satisfying the constraints. Let us first show that the Kullback-Leibler divergence can be used as a distance function: For an arbitrary probability distribution u_b, the relation

$$\sum_b \rho'_b \log\frac{u_b}{\rho_b} \leq \sum_b \rho'_b \left(\frac{u_b}{\rho'_b} - 1\right) = 0, \tag{4.11.4}$$

follows from the general inequality $\log(x) \leq (x-1)$ and $x = u_b/\rho'_b$ and summing over b, with equality only and only if $x = 1$; thus the relative entropy is zero if and only if $u_b = \rho'_b$. This gives us

$$H[\boldsymbol{\rho}'\|\boldsymbol{\rho}] \geq 0, \tag{4.11.5}$$

so that we can use the relative entropy as a variational principle:

$$\delta\left[H[\boldsymbol{\rho}'\|\boldsymbol{\rho}] - \mu\sum_b \rho'_b - \sum_b \lambda g_b \rho'_b\right] = \sum_b \left[1 + \log\frac{\rho'_b}{\rho_b} - \mu\rho'_b - \lambda g_b\right]\delta\rho'_b, \tag{4.11.6}$$

with solution

$$\rho'_b = \rho_b \exp\left[\mu^* + g_b\lambda^* - 1\right], \tag{4.11.7}$$

which is equivalent to the previous solution (4.10.5). We can also show that the variational argument always works by separating the expression (4.11.5) into

$$\sum_b \rho'_b \log \rho'_b \geq \sum_b \rho'_b \log u_b \tag{4.11.8}$$

and choosing for u_b (4.11.7) the inequality reduces to,

$$\sum_b \rho'_b \log \rho'_b \geq \sum_b \rho'_b \log \rho_b + \sum_b \rho'_b \mu^* + \sum_b g_b \rho'_b \lambda^* - 1 \tag{4.11.9}$$

or

$$H[\boldsymbol{\rho'}||\boldsymbol{\rho}] \geq \mu^* + \lambda^* \sum_b g_b \rho'_b - 1. \tag{4.11.10}$$

Varying over all possible $\boldsymbol{\rho'}$ that satisfies the constraint $\sum_b g_b \rho'_b = \gamma$ we see that the right hand side will remain constant. So to reach the minimum on the left hand side all the ρ'_b must equal u_b, in which case

$$H_{\min}[\boldsymbol{\rho'}||\boldsymbol{\rho}] = \mu^* + \lambda^* \gamma - 1, \tag{4.11.11}$$

showing that the minimum relative entropy solution is also the most probable and the mean value solution in turn. We must emphasise that the justification of the most probable and minimum relative entropy solution is that they coincide with the mean value solution and that all three solutions rely heavily on the assumption of logical independence. For the special case $\rho_b = 1/B$ and multiplying with (-1) we can reduce the minimum relative entropy principle to the maximum entropy principle.

4.12 Examples

Sometimes we have to deal with a large number of possible outcomes ($\lim B \to \infty$). The difficulty that this creates is that our model then becomes dependent on the specific details on how the limit is taken. The first approach to this problem is to assume equal prior probabilities ($\rho_b = 1/B$) and compute the partition function only up to a constant.

Consider for example **Planck's oscillator**: $g_b = b$ and $b = 0, 1, 2, \ldots$, where the average $G = \sum_{b=0}^{\infty} b r_b$ is kept fixed. The partition function is,

$$Z[\lambda] \propto \sum_{b=0}^{\infty} e^{-b\lambda} = \frac{1}{1 - e^{-\lambda}} \tag{4.12.1}$$

and using the negative binomial theorem (2.1.19) on the moment generating function, we find the structure function as the coefficient of $e^{-\lambda G}$ of Z^R,

$$Z^R[\lambda] \propto (1 - e^{-\lambda})^{-R} = \sum_{G=0}^{\infty} \binom{-R}{G} (-e^{-\lambda})^G \tag{4.12.2}$$

So that

$$\Omega(R,G) \propto \binom{R+G-1}{G}. \tag{4.12.3}$$

Applying (4.8.2), we find the exact answer to be a **negative hypergeometric distribution**,

$$\rho'_b = \frac{(R-1)G^{\underline{b}}}{(G+R-1)^{\underline{b+1}}} = \frac{\binom{-1}{b}\binom{-(R-1)}{G-b}}{\binom{-R}{G}}. \tag{4.12.4}$$

The negative hypergeometric distribution will play a very prominent role in the future chapters; thus it is very interesting that it appears here.

The saddlepoint for the same partition function is,

$$\frac{d}{d\lambda}\log\frac{1}{1-e^{-\lambda}}\bigg|_{\lambda=\lambda^*} = -\gamma \qquad \gamma = \frac{1}{e^{\lambda^*}-1} \qquad \lambda^* = \log\frac{1+\gamma}{\gamma} \tag{4.12.5}$$

and the approximate model is using (4.9.5)

$$\rho'_b = \frac{1}{1+\gamma}\left(1+\frac{1}{\gamma}\right)^{-b} \tag{4.12.6}$$

a **geometric distribution**.

A second approach is to assume that we are examining an interval that can be subdivided into sub-intervals in each of which only one event can occur independently of the interval size and other intervals. The result b is then that b such events occurred in the interval, so that the prior probability should be proportional to $\binom{B}{b}$. If we make B large, $\rho_b \propto \frac{1}{b!}$. The resulting partition function and structure function are

$$Z[\lambda] \propto \sum_{b=0}^{\infty}\frac{e^{-\lambda b}}{b!} = e^{e^{-\lambda}} \qquad \Omega(R,G) \propto \frac{R^G}{G!} \tag{4.12.7}$$

and applying (4.8.2) again,

$$\rho'_b = \binom{G}{b}\left(\frac{1}{R}\right)^G\left(1-\frac{1}{R}\right)^{G-b} \tag{4.12.8}$$

is a binomial distribution with $1/R$ acting as a probability. The saddlepoint approximation is

$$\frac{d}{d\lambda}e^{-\lambda}\bigg|_{\lambda=\lambda^*} = -\gamma \qquad \lambda^* = -\log\gamma \tag{4.12.9}$$

and the approximate model using (4.9.5) again yields

$$\rho'_b = e^{-\gamma}\frac{\gamma^b}{b!}, \tag{4.12.10}$$

which is a **Poisson Distribution**.

An observant reader would have noticed that there is a factor ρ_b that has seemingly disappeared from (4.11.4), (4.12.6), (4.11.7) and (4.12.10). So let us do a more elaborate

calculation under the heading of a **Generalized Planck Oscillator**: Let $g_b = b = 1, 2, \ldots$ and let the prior probability be negative binomial distributed with parameters k and $0 < \theta < 1$,

$$\rho_b = \binom{b+k-1}{b}\theta^k(1-\theta)^b \qquad\qquad b = 0, 1, 2, \ldots, \tag{4.12.11}$$

so that we introduce a stopping rule for the prior probability of our states. From the partition function

$$Z[\lambda] = \sum_{b=0}^{\infty}\binom{k+b-1}{b}\theta^k(1-\theta)^b e^{-b\lambda} = \frac{\theta^k}{[1-e^{-\lambda}(1-\theta)]^k} \tag{4.12.12}$$

we can again read off our structure function as the coefficient of $e^{-\lambda G}$, with the help of the negative binomial theorem (2.1.20)

$$\Omega[R, G] = \binom{Rk+G-1}{G}\theta^{Rk}(1-\theta)^G. \tag{4.12.13}$$

Using (4.8.2) our exact model is

$$\rho'_b = \frac{\binom{-k}{b}\binom{-(R-1)k}{G-b}}{\binom{-Rk}{G}}, \tag{4.12.14}$$

which is the **negative hypergeometric distribution** again. Our saddlepoint solution is,

$$\frac{k(1-\theta)}{e^{\lambda^*}+\theta-1} = \gamma \qquad\qquad \lambda^* = \log\left[\frac{(\gamma-k)(1-\theta)}{\gamma}\right] \tag{4.12.15}$$

and from (4.9.5)

$$\rho'_b = \binom{k+b-1}{b}\left(\frac{k}{k+\gamma}\right)^k\left(\frac{\gamma}{k+\gamma}\right)^b. \tag{4.12.16}$$

Importantly the value of θ is immaterial to our solution, thus the information contained in our g_b has in fact replaced it. Checking our calculation: we take the limit of our exact model using (2.1.6) and $G = \gamma R$,

$$\begin{aligned}\frac{\binom{-k}{b}\binom{-(R-1)k}{G-b}}{\binom{-Rk}{G}} &= \binom{k+b-1}{b}\frac{(R\gamma)!}{(R\gamma-b)!}\frac{(Rk)!}{(Rk-k)!}\frac{(Rk-k+R\gamma-b-1)!}{(Rk+R\gamma-1)!}\\ &\approx \binom{k+b-1}{b}(R\gamma)^b(Rk)^k(Rk+R\gamma)^{-k-b}\\ &= \binom{k+b-1}{b}\left(\frac{k}{k+\gamma}\right)^k\left(\frac{\gamma}{k+\gamma}\right)^b.\end{aligned} \tag{4.12.17}$$

Choosing $k = 1$ reduces this answer to the geometric distribution (4.12.6) while the large $k \to \infty$ asymptotic approximation is, using (2.1.6),

$$\begin{aligned}\log\rho'_b &= \log\binom{k+b-1}{b} + b\log\left(\frac{\gamma}{k+\gamma}\right) + k\log\left(\frac{k}{k+\gamma}\right)\\ &\sim \log\frac{k^b}{b!} + b\log\left(\frac{\gamma}{k+\gamma}\right) - \gamma\end{aligned} \tag{4.12.18}$$

so we end up with a **Poisson Distribution**,

$$\rho'_b = e^{-\gamma}\frac{\gamma^b}{b!}. \tag{4.12.19}$$

Another example where we can do all the calculations explicitly is the **Fermi Oscillator**, see Jaynes (1968). We have two states; one contributes one energy unit and the other nothing. The state that contributes is labelled a success and the state that does not contribute a failure. G is then interpreted as the total number of successes and the partition function becomes,

$$Z[\lambda] = \rho_f + \rho_s e^{-\lambda}. \tag{4.12.20}$$

The generating function is

$$\begin{aligned}\Phi[\Omega(R,G),G,\lambda] &= \left(\rho_f + \rho_s e^{-\lambda}\right)^R \\ &= \sum_{G=0}^{R}\binom{R}{G}\rho_s^G \rho_f^{R-G} e^{-\lambda G}.\end{aligned} \tag{4.12.21}$$

So the structure function is simply

$$\Omega(R,G) = \binom{R}{G}\rho_s^G(1-\rho_s)^{R-G}, \tag{4.12.22}$$

which is easy to understand. Applying (4.8.2) we find the model assignment is

$$\rho'_s = \frac{G}{R} \qquad\qquad \rho'_f = 1 - \frac{G}{R}, \tag{4.12.23}$$

for a specific trial. So splitting the sample space in two gives us a Bernoulli trial and the model assignment is again independent of the prior ρ_b. The saddlepoint equation is

$$\frac{d}{d\lambda}\log Z\left[\lambda\right] = \frac{d}{d\lambda}\log\left[\rho_f + \rho_s e^{-\lambda}\right] = -\frac{\rho_s}{\rho_s + \rho_f e^{\lambda}} = -\gamma, \tag{4.12.24}$$

with solution

$$\lambda^* = \log\left[\frac{\rho_s(1-\gamma)}{\rho_f\gamma}\right], \qquad\qquad Z[\lambda^*] = \frac{\rho_f}{1-\gamma} \tag{4.12.25}$$

which we substitute in (4.9.5) to find

$$\rho_s = \gamma \qquad\qquad \rho_f = 1-\gamma. \tag{4.12.26}$$

In this case the saddle point equation gives the exact answer.

The third example is traditionally used to illustrate the principle of maximum entropy and is called the Kangaroo example taken from Gull and Skilling (1984): the information is that a third of all kangaroos have blue eyes and one-third of all kangaroos are left-handed. The question is: what is the probability of a kangaroo being both left-handed

and blue-eyed? Of course we do not have enough information to give the correct answer but we should still be able to give a consistent reply. Let

$$\begin{aligned}\mathcal{B} &\equiv \texttt{Kangaroo is blue-eyed.}\\ \mathcal{L} &\equiv \texttt{Kangaroo is left-handed.}\\ x &\equiv p(\mathcal{B},\mathcal{L}|\mathcal{I})\end{aligned} \tag{4.12.27}$$

then the set of all possible solutions are, with $x = p(\mathcal{B},\mathcal{L}|\mathcal{I})$

Feasible Solutions — Left-handed

Blue-eyed	$\mathcal{L}$	$\overline{\mathcal{L}}$
$\mathcal{B}$	x	$\frac{1}{3}-x$
$\overline{\mathcal{B}}$	$\frac{1}{3}-x$	$\frac{1}{3}+x$

where $0 \leq x \leq \frac{1}{3}$. Assuming the prior where all cases are equally likely $\rho_{\mathcal{B},\mathcal{L}} = \rho_{\mathcal{B},\overline{\mathcal{L}}} = \rho_{\overline{\mathcal{B}},\mathcal{L}} = \rho_{\overline{\mathcal{B}},\overline{\mathcal{L}}} = \frac{1}{4}$ we calculate the different entropies for varying values of x seeking the maximum,

$$\begin{aligned}-\delta\left[\sum_b \rho'_b \log\frac{\rho'_b}{\rho_b}\right] &= -\delta\left[x\log 4x + 2\left(\tfrac{1}{3}-x\right)\log 4\left(\tfrac{1}{3}-x\right) + \left(x+\tfrac{1}{3}\right)\log 4\left(\tfrac{1}{3}+x\right)\right]\\ &= -\left[\log 4x - 2\log 4\left(\tfrac{1}{3}-x\right) + \log 4\left(\tfrac{1}{3}+x\right)\right]\delta x = 0.\end{aligned} \tag{4.12.28}$$

The maximum is attained at $x = \frac{1}{9}$, giving

Maximum Entropy Solution — Left-handed

Blue-eyed	$\mathcal{L}$	$\overline{\mathcal{L}}$
$\mathcal{B}$	$\frac{1}{9}$	$\frac{2}{9}$
$\overline{\mathcal{B}}$	$\frac{2}{9}$	$\frac{4}{9}$

The remarkable property of this answer is that this is the logical independent solution: The conditional probability a kangaroo being blue-eyed if we know it is left-handed is,

$$p(\mathcal{B}|\mathcal{L},\mathcal{I}) = \frac{p(\mathcal{B},\mathcal{L}|\mathcal{I})}{p(\mathcal{L}|\mathcal{I})} = \frac{1}{3} = p(\mathcal{B}|\mathcal{I}), \tag{4.12.29}$$

exactly the same as if we did not know it was left-handed. The principle of Minimum Relative Entropy was derived from a multinomial distribution, thus we should not be surprised that it should pick out the answer closest to the logical independent solution in any specific problem. Is this however the correct inference to make in this situation? Is it justified to pick out the logical independent answer if we are in a state of ignorance? We will try to answer this question later, but for now we know that the Principle of Minimum

Relative Entropy is in fact an approximation to the Darwin-Fowler method and after a moment's thought we realise that the Kangaroo example is just a two-dimensional Fermi oscillator. Write G_1 for the total number of left-handed kangaroos and G_2 for the total number of blue-eyed kangaroos and R for the total number of kangaroos. The partition function is,

$$Z[\lambda_1,\lambda_2]^R = \left(\frac{1}{4}+\frac{e^{-\lambda_1}}{4}+\frac{e^{-\lambda_2}}{4}+\frac{e^{-\lambda_1-\lambda_2}}{4}\right)^R \tag{4.12.30}$$

$$= \left(\frac{1}{4}\right)^R \left(1+e^{-\lambda_1}\right)^R \left(1+e^{-\lambda_2}\right)^R \tag{4.12.31}$$

the structure function,

$$\Omega\left[R,G_1,G_2\right] = \binom{R}{G_1}\binom{R}{G_2}\frac{1}{4^R} \tag{4.12.32}$$

and the exact solution is,

$$\rho'_{\mathcal{B},\mathcal{L}} = \frac{G_1}{R}\frac{G_2}{R} \tag{4.12.33}$$

and with $\frac{G_1}{R} = \frac{G_2}{R} = \frac{1}{3}$ this gives exactly the same answer as the maximum entropy principle. But we can also give an error estimate by using (4.8.5),

$$\mathrm{var}(\rho'_{\mathcal{B},\mathcal{L}}) = \frac{G_1G_2\left(G_1G_2-G_1-G_2-R+2\right)}{R(R-1)}, \tag{4.12.34}$$

which is not always appreciated in the literature of information theory.

4.13 Grand canonical ensemble

In the derivation of the principle of minimum relative entropy, we made a subtle switch at one point: Instead of considering the number of trials R to be fixed and conditioning on R, it was instead considered as an extra explicit constraint that must be enforced. We switched from an implicit constraint to an explicit constraint implying that there should be a framework without the number of trials fixed beforehand. Here we will derive such a framework based on a completely different assumption. Later we will show that in fact the formalism for minimum relative entropy is the same as for the grand canonical ensemble where the number of trials is not fixed. Let us start from a different knowledge base:

> Knowledge base $\mathcal{T}$:
> There exists a positive real number $\lambda\,dt$, namely the probability that an event or count will occur in the time interval $(t, t+dt)$. The knowledge of λ makes any information about the occurrence or non-occurrence in any other time interval irrelevant.

The motivation for this assumption is that we would require information to connect two different counts in time and we assume that this information is unavailable. We also

prominently assume logical independence of the counts in the intervals. Let $p(\Delta t|\lambda,\mathcal{T})$ be the probability that no event occurred in the interval $(0,\Delta t)$ and $\overline{\mathcal{E}}$ the proposition that no event occurred in the interval $(\Delta t,\Delta t+dt)$; the probability that no event occurred in the interval $(0,\Delta t+dt)$ is,

$$p(\Delta t+dt|\lambda,\mathcal{T}) \quad = p(\Delta t,dt|\lambda,\mathcal{T}) = p(\Delta t|\lambda,\mathcal{T})p(dt|\Delta t,\lambda,\mathcal{T}) = p(\Delta t|\lambda,\mathcal{T})(1-\lambda dt) \tag{4.13.1}$$

or after a Taylor expansion and taking the limit $dt\to 0$,

$$\frac{d}{dt}p(\Delta t|\lambda,\mathcal{T}) = -\lambda p(\Delta t|\lambda,\mathcal{T}). \tag{4.13.2}$$

Using the obvious initial condition $p(\Delta t=0|\lambda,\mathcal{T})=1$, we find,

$$p(\Delta t|\lambda,\mathcal{T}) = \lambda e^{-\lambda\Delta t} \qquad\qquad \lambda>0, \quad \Delta t\geq 0. \tag{4.13.3}$$

Assuming that a distribution is memoryless i.e. the probability of an event in any given interval is independent of the probability in any other interval is a strong enough property to characterise the exponential distribution as the unique solution for continuous intervals.

Defining the logical proposition with the parameter n as the nth event in the time interval Δt and τ_n as the time we waited for that event, then using the waiting-time identity (4.6.2) again,

$$\sum_{r=0}^{n-1} p(r|\lambda,\Delta t,\mathcal{T}) = \int_{\Delta t}^{\infty} p(\tau_n|\lambda,n,\mathcal{T})d\tau_n, \tag{4.13.4}$$

which reads that the probability of seeing fewer than n counts in the time interval Δt is the same as the probability that the nth event τ_n took longer than the interval Δt. The distribution for τ_n is, using independence and (4.13.3),

$$p(\tau_n=t_1+\cdots+t_n|\lambda,n,\mathcal{T}) = \prod_{j=1}^{n} p(t_j|\lambda,\mathcal{T})\delta\left(\tau_n-\sum_j t_j\right), \tag{4.13.5}$$

where each t_j is the time interval between the jth and $(j-1)$st event, which is unknown, but is distributed like an exponential distribution. Taking the Laplace transform in τ_n,

$$\mathcal{L}\left[s,\tau_n|p(\tau_n,\boldsymbol{t}|\lambda,n,\mathcal{T})\right] = \int_0^{\infty}\prod_{j=1}^{n} p(t_j|\lambda,\mathcal{T})\delta\left(\tau_n-\sum_j t_j\right)e^{-s\tau_n}d\tau_n \tag{4.13.6}$$

and integrating out the t_j intervals, we obtain

$$\mathcal{L}\left[s,\tau_n|p(\tau_n|\lambda,n,\mathcal{T})\right] = \prod_{j=1}^{n}\int_0^{\infty}\lambda e^{-\lambda t_j-st_j}dt_j = \left(\frac{\lambda}{s+\lambda}\right)^n \tag{4.13.7}$$

which is the Laplace Transform for the combined measurements τ_n. Inverting the Laplace transform gives us a **Gamma Distribution** for τ_n which follows directly from (4.13.5)

$$p(\tau_n|\lambda,n,\mathcal{T}) = \lambda e^{-\lambda\tau_n}\frac{(\lambda\tau_n)^{n-1}}{(n-1)!}, \tag{4.13.8}$$

that is the distribution for taking n consecutive measurements is distributed like a Gamma distribution or n convolutions of an exponential distribution. Integrating the right hand side of (4.13.4) repeatedly by parts we have

$$\begin{aligned}\int_{\Delta t}^{\infty} \lambda e^{-\lambda\tau_n}\frac{(\lambda\tau_n)^{n-1}}{(n-1)!}d\tau_n &= -\,e^{-\lambda\tau_n}\frac{(\lambda\tau_n)^{n-1}}{(n-1)!}\Big|_{\tau=\Delta t}^{\infty} + \int_{\Delta t}^{\infty}\lambda e^{-\lambda\tau_n}\frac{(\lambda\tau_n)^{n-2}}{(n-2)!}d\tau_n \\ &= e^{-\lambda\Delta t}\frac{(\lambda\Delta t)^{n-1}}{(n-1)!} + e^{-\lambda\Delta t}\frac{(\lambda\Delta t)^{n-2}}{(n-2)!} + \cdots + \lambda\Delta t e^{-\lambda\Delta t} + e^{-\lambda\Delta t} = \sum_{r=0}^{n-1} e^{-\lambda\Delta t}\frac{(\lambda\Delta t)^r}{r!},\end{aligned} \tag{4.13.9}$$

which shows that the probability of seeing exactly r counts in an interval Δt with rate λ is a Poisson Distribution,

$$p(r|\lambda,\Delta t,\mathcal{I}) = e^{-\lambda\Delta t}\frac{(\lambda\Delta t)^r}{r!}. \tag{4.13.10}$$

To show how the Poisson distribution connects with maximum entropy formalism we employ an artifice, introduced by Fisher (1922*a*), that a conditional distribution of independent Poisson distributions gives a multinomial distribution. Take a set of B independent processes one for each bin with their own rate parameter λ_b and ask what is the probability for seeing a total of R events in all the bins together during the time interval Δt

$$\begin{aligned}p(R|\boldsymbol{\lambda},\mathcal{T}) &= p(R = r_0 + \cdots + r_B|\Delta t,\boldsymbol{\lambda},\mathcal{T}) \\ &= e^{-\sum_b \lambda_b\Delta t}\frac{\left(\sum_b \lambda_b\Delta t\right)^R}{R!},\end{aligned} \tag{4.13.11}$$

because the convolution of Poisson Distributions is again a Poisson Distribution. Assuming that we count events until we have a total of R, the posterior is

$$p(\boldsymbol{r}|R,\boldsymbol{\lambda},\Delta t,\mathcal{T}) = \frac{p(\boldsymbol{r}|\boldsymbol{\lambda},\Delta t,\mathcal{T})p(R|\boldsymbol{r},\boldsymbol{\lambda},\Delta t,\mathcal{T})}{p(R|\boldsymbol{\lambda},\Delta t,\mathcal{T})} = \frac{\prod_b e^{-\lambda_b\Delta t}\frac{(\lambda_b\Delta t)^{r_b}}{r_b!}\delta\left(R-\sum_b r_b\right)}{e^{-\sum_b\lambda_b\Delta t}\frac{\left(\sum_b\lambda_b\Delta t\right)^R}{R!}}, \tag{4.13.12}$$

and defining a probability for seeing a event as the ratio of its rate towards the total rate,

$$\rho_b = \frac{\lambda_b}{\sum_b \lambda_b}, \tag{4.13.13}$$

gives us

$$p(\boldsymbol{r}|R,\boldsymbol{\lambda},\Delta t,\mathcal{T}) = R!\prod_b \frac{\rho_b^{r_b}}{r_b!} = p(\boldsymbol{r}|\boldsymbol{\rho},R,\mathcal{T}), \tag{4.13.14}$$

from which we learn that fixing R makes the time interval Δt redundant and changes our rate parameters into probabilities. Thus we could have derived exactly the same formalism by starting from a product of Poisson distributions.

We now derive the Principle of Minimum Relative Entropy in its Grand Canonical form: Taking a product of Poisson Distribution

$$p(\boldsymbol{r}|\boldsymbol{s},\mathcal{T}) = \prod_b e^{-s_b}\frac{s_b^{r_b}}{r_b!} \tag{4.13.15}$$

the large prediction limit is

$$-\log p(\boldsymbol{r}|\boldsymbol{s},\mathcal{K}) = \sum_b s_b - \sum_b r_b\left(1+\log\frac{s_b}{r_b}\right). \tag{4.13.16}$$

Now recall that the generic relative entropy for $\boldsymbol{\rho}'$ and $\boldsymbol{\rho}$ is

$$H[\boldsymbol{\rho}'||\boldsymbol{\rho}] = \sum_b \rho'_b \log\frac{\rho'_b}{\rho_b}. \tag{4.13.17}$$

The above strongly suggests a "generalised divergence" for any nonnegative $\boldsymbol{s},\boldsymbol{r}$ without normalisation constraints should be

$$H[\boldsymbol{r}||\boldsymbol{s}] = -\sum_b r_b\left(1+\log\frac{s_b}{r_b}\right). \tag{4.13.18}$$

Adding a constraint and varying the $\boldsymbol{r}$ we find,

$$\nabla_{r_c}\left[H[\boldsymbol{r}||\boldsymbol{s}] - \sum_b \lambda r_b g_b\right] = [\log r_c - \log s_c - r_c g_c] = 0 \qquad \forall c. \tag{4.13.19}$$

with solution

$$r_c = s_c e^{-\lambda g_c}. \tag{4.13.20}$$

4.14 From Logical Independence to Exchangeability

We end this chapter by pointing out the weaknesses of Logical Independence, which has played a fundamental role so far. The remedy for these weaknesses lies in the replacement of Logical Independence with *Exchangeability*, which will likewise be central for the remainder of this dissertation.

While there are probably many ways to illustrate the issue, let us concentrate on the incompatibility of Logical Independence with what we would call "learning" or, technically speaking, the updating of the evidence as data becomes available. As defined in (3.2.5), Logical Independence of two logical statements can be defined as the factorisation

$$p(\mathcal{A},\mathcal{B}|\mathcal{I}) = p(\mathcal{A}|\mathcal{I})p(\mathcal{B}|\mathcal{I}). \tag{4.14.1}$$

For N data points $\boldsymbol{x} = \{x_i\}_{i=1}^N$ assumed to be logically independent, this translates into the factorisation of the joint likelihood

$$p(\boldsymbol{x}|\mathcal{I}) = \prod_{i=1}^N p(x_i|\mathcal{I}) \tag{4.14.2}$$

and therefore obviously also

$$p(x_{N+1}, \boldsymbol{x}|\mathcal{I}) = \prod_{i=1}^{N+1} p(x_i|\mathcal{I}) \tag{4.14.3}$$

from which follows, catastrophically, that

$$p(x_{N+1}|\boldsymbol{x}, \mathcal{I}) = \frac{p(x_{N+1}, \boldsymbol{x}|\mathcal{I})}{p(\boldsymbol{x}|\mathcal{I})} \;=\; p(x_{N+1}|\mathcal{I}). \tag{4.14.4}$$

Using Logical Independence therefore is inescapably equivalent to ignoring existing data $\boldsymbol{x}$ for the purposes of predicting x_{N+1}. Put differently, Logical Independence means that *no learning can occur*.

The example used by Jaynes (2003) is a simple illustation of the point. Throw a six sided die twice. The probability of seeing result x_1=1 on the first trial is, using (4.3.3),

$$\begin{aligned} p(x_1 = 1|\mathcal{I}) &= \sum_{x_2=1}^{6} p(x_1 = 1, x_2|R = 2) \\ &= \frac{6}{36} = \frac{1}{6}, \end{aligned} \tag{4.14.5}$$

where we summed all the vectors that start with a one. So what probability will we assign to the second trial if we know the first result? The conditional probability is

$$\begin{aligned} p(x_2 = b|x_1 = 1, \mathcal{I}) &= \frac{p(x_1 = 1, x_2 = b|\mathcal{I})}{p(x_1 = 1|\mathcal{I})} \\ &= \frac{1/36}{1/6} = \frac{1}{6}. \end{aligned} \tag{4.14.6}$$

Knowing the first result does not change the probability assignment for the second trial. If we throw a dice ten thousand times and it shows the same face every time our assumption of Logical Independence absurdly forces us to assign a probability of $1/6$ to the next trial x_{N+1}, while any rational observer would have realised the trials are not independent and abandoned his assumption of Logical Independence. Of course, the same rational observer would then have to admit that his assumption of Logical Independence had been wrong from the start, and redo all his calculations. You cannot change assumptions mid-way during a calculation.

Logical Independence is demonstrably a very restrictive assumption which only applies to special cases. Clearly, we need an alternative assumption or hypothesis which differs minimally from Logical Independence in the sense that we can construct a formalism similar to that developed in this chapter, but which does have a mechanism to update evidence with data.

A clue to finding such a new assumption lies in the fact that Logical Independence is a factorisation of the joint likelihood, which ignores priors and their role. Indeed, it is quite possible to construct self-updating evidences based on factorising likelihoods simply by introducting a nontrivial prior; generically

$$\text{(updating evidence)} = \int \text{(logically independent likelihood)} \times \text{(nontrivial prior)}\, d\boldsymbol{\theta}\,.$$

The "trivial prior" that leads to conventional Logical Independence calculations is one where the entire set of parameters $\boldsymbol{\theta}$ is *assumed perfectly known*, i.e. the prior is a product of delta functions, in which case the likelihood and evidence are identical and no learning can occur. The moment we use a nontrivial prior, however, the evidence no longer factorises and updating of the evidence is possible. This would seem to be one of the strongest arguments in favour of priors in general.

The destruction of the property of Logical Independence by means of a nontrivial prior is clearly necessary and desirable for nontrivial model building. One can, however, now ask: If Logical Independence does not and should not survive in the construction of the evidence, is there a property which does? The answer lies in the concept of *Exchangeability*, the property that a probability remains unchanged under permutation of the data points. Indeed, we have a "conservation law for exchangeability"

$$(\text{exchangeable evidence}) = \int (\text{exchangeable likelihood}) \times (\text{prior})\, d\boldsymbol{\theta}\,.$$

In the second part of this dissertation, starting with Chapter 5, we shall explore the mathematical and conceptual consequences of replacing Logical Independence by Exchangeability. Very briefly, what we will show is the following:

- *All* exchangeable evidences can be written in terms of a hypergeometric likelihood and a prior. The central role played by the multinomial distribution in the logically independent case will now be played by the hypergeometric.
- While the hypergeometric distribution is not unique in satisfying "conservation of exchangeability", the other known exchangeable distributions are all limiting cases of the hypergeometric.
- The theory developed in the remainder of this dissertation, based on the hypergeometric likelihood, therefore forms a universal basis for all exchangeable evidences.

Chapter 5

Likelihoods and priors for exchangeable sequences

Here and below, we shall endeavour to progress from the traditional assumption of logical independence to the wider and, in our view, more appropriate concept of exchangeability. The former is defined as

> `Logical independence:`
> `A sequence of results` $\boldsymbol{x} = \{x_1, \ldots, x_N\}$ `is called logically independent if the joint probability of` $\boldsymbol{x}$ `fully factorises,`
>
> $$p(\boldsymbol{x}|\mathcal{I}) = \prod_{i=1}^{N} p(x_i|\mathcal{I}). \tag{5.0.1}$$

Indeed, there are many instances where logical independence is justified, for example if no causes are shared by the events or in the information-based view if we are certain of our model. However, as shown by the example of the hypergeometric distribution below, there are situations and cases where individual trials are fully exchangeable and occupation numbers fully appropriate, but where these trials are *not independent*. Hence we define

> `Exchangeability:`
> `A sequence of results` $\boldsymbol{x} = \{x_1, \ldots, x_N\}$ `is called exchangeable if the probability of every possibly permutation` $\boldsymbol{x}_\pi$ `of` $\boldsymbol{x}$ `is the same,`
>
> $$p(\boldsymbol{x}_\pi|\mathcal{H}) = p(\boldsymbol{x}|\mathcal{H}). \tag{5.0.2}$$

The occurrence and utility of occupation numbers (or *binning* in the language of experimentalists) is thus not a proof of independence at all. It should hence be no surprise if the wider scope of exchangeability were to result in important new insights and applications.

5.1 Overview

Throughout this dissertation the assumption of exchangeability is fundamental. All the systems we look at are based on exchangeable variables, so all the relevant calculations

can be reduced from ordered variables to occupation numbers. As a reminder of our notation: the sample space of *one* trial $\mathcal{S}$ with M elementary outcomes is partitioned such that there are $\boldsymbol{m} = \{m_b\}_{b=1}^B$ elementary outcomes in the partition $\{\mathcal{A}_b\}_{b=1}^B$ of $\mathcal{S}$, which common properties are mapped onto[1] $b \in \{0, 1, \ldots, B-1\}$. Further, $\boldsymbol{n} = \{n_b\}_{b=1}^B$ denotes the set of occupation numbers of N real measured data points $\{x_i\}_{i=1}^N$, while $\boldsymbol{r} = \{r_b\}_{b=1}^B$ is the set of predicted occupation numbers for R future measurements $\{x_i\}_{i=N+1}^{N+R}$. We shall also make use of the normalised $\boldsymbol{\rho} = \boldsymbol{m}/M$ for large M.

As mentioned in Section 3.8 the aim is to calculate the **posterior** distribution and subsequently the **predictive** distribution once the posteriors are known. Both rely on the calculation of the evidence

$$p(\boldsymbol{n}|\mathcal{K}) = \sum_{U(\boldsymbol{m})} p(\boldsymbol{n}|\boldsymbol{m}, \mathcal{K})\, p(\boldsymbol{m}|\mathcal{K}) \tag{5.1.1}$$

with $\sum_{U(\boldsymbol{m})}$ the sum or integral over all possible configurations of $\boldsymbol{m}$, in terms of which the posterior for $\boldsymbol{m}$ is determined by

$$p(\boldsymbol{m}|\boldsymbol{n}, \mathcal{K}) = \frac{p(\boldsymbol{n}, \boldsymbol{m}|\mathcal{K})}{p(\boldsymbol{n}|\mathcal{K})} = \frac{p(\boldsymbol{n}|\boldsymbol{m}, \mathcal{K})\, p(\boldsymbol{m}|\mathcal{K})}{p(\boldsymbol{n}|\mathcal{K})} \tag{5.1.2}$$

which then enters into the prediction

$$\begin{aligned} p(\boldsymbol{r}|\boldsymbol{n}, \mathcal{K}) &= \sum_{U(\boldsymbol{m})} p(\boldsymbol{r}|\boldsymbol{n}, \boldsymbol{m}, \mathcal{K})\, p(\boldsymbol{m}|\boldsymbol{n}, \mathcal{K}) \\ &= \sum_{U(\boldsymbol{m})} \frac{p(\boldsymbol{r}|\boldsymbol{n}, \boldsymbol{m}, \mathcal{K})\, p(\boldsymbol{n}|\boldsymbol{m}, \mathcal{K})\, p(\boldsymbol{m}|\mathcal{K})}{p(\boldsymbol{n}|\mathcal{K})}, \end{aligned} \tag{5.1.3}$$

which can be written as the ratio of "two evidences"

$$p(\boldsymbol{r}|\boldsymbol{n}, \mathcal{K}) = \frac{p(\boldsymbol{r}, \boldsymbol{n}|\mathcal{K})}{p(\boldsymbol{n}|\mathcal{K})} = \frac{\sum_{U(\boldsymbol{m})} p(\boldsymbol{r}, \boldsymbol{n}|\boldsymbol{m}, \mathcal{K})\, p(\boldsymbol{m}|\mathcal{K})}{\sum_{U(\boldsymbol{m})} p(\boldsymbol{n}|\boldsymbol{m}, \mathcal{K})\, p(\boldsymbol{m}|\mathcal{K})}. \tag{5.1.4}$$

Both posteriors and final predictions are based on knowledge or information, captured in constructing a specific model consisting of

(a) one or more priors $p(\boldsymbol{m}|\mathcal{K})$ or $p(\boldsymbol{\rho}|\mathcal{K})$ with $\rho_b = m_b/M$ possible becoming continuous, and

(b) one or more likelihoods $p(\boldsymbol{r}|\boldsymbol{m}, \mathcal{K})$ or $p(\boldsymbol{r}|\boldsymbol{\rho}, \mathcal{K})$, in which exact knowledge of $\boldsymbol{m}$ or equivalently $\boldsymbol{\rho}$ is hypothesised.

Later in this chapter we will turn to assigning prior distributions and in a later chapter will focus on constructing likelihood functions for exchangeable sequences.

[1] While mapping to $\{0, 1, \ldots, B-1\}$ rather than to $\{1, 2, \ldots, B\}$ may seem inconvenient, it yields the standard formulae for moments and cumulants of the binomial. In our convention, result b is therefore associated with occupation numbers r_{b+1}, m_{b+1} etc.

We use the same functional form or model for both likelihoods $p(\boldsymbol{r}|\boldsymbol{m},\mathcal{K})$ and $p(\boldsymbol{n}|\boldsymbol{m},\mathcal{K})$, so that all mathematical results apply to both. Conceptually, however, they differ radically: $p(\boldsymbol{r}|\boldsymbol{m},\mathcal{K})$ is a prediction of $\boldsymbol{r}$, where $\boldsymbol{r}$ can be any vector in $U(\boldsymbol{r})$, while $p(\boldsymbol{n}|\boldsymbol{m},\mathcal{K})$ should strictly speaking be written as

$$p(\boldsymbol{n}|\boldsymbol{m},\mathcal{K}) \equiv p(\boldsymbol{r}{=}\boldsymbol{n}\,|\,\boldsymbol{m},\mathcal{K}) \tag{5.1.5}$$

i.e. the *probability as predicted by the knowledge base $\mathcal{K}$, by our model and within that model by a particular hypothetical value of $\boldsymbol{m}$, that one of the results $\boldsymbol{r}$ which are possible within the model is exactly our measured data $\boldsymbol{n}$.*[2] As $p(\boldsymbol{r}|\boldsymbol{m},\mathcal{K})$ and $p(\boldsymbol{n}|\boldsymbol{m},\mathcal{K})$ are mathematically identical, we shall do most of our calculations using the notation of $p(\boldsymbol{r}|\boldsymbol{m},\mathcal{K})$, substituting $\boldsymbol{n}$ for $\boldsymbol{r}$ only when needed e.g. in (5.1.2) and (5.1.3). The **major points** in this chapter are:

(a) We will replace Logical Independence with Exchangeability, which is a much more natural assumption to make.

(b) The hypergeometric distribution will consequently replace the multinomial we used before.

(c) The prior that assumes the least is the negative hypergeometric distribution and correspondingly should be used in most circumstances.

(d) We will explore the use of symmetries to assign prior probabilities.

The sections can be summarised as follows:

- In Section 5.2 we illustrate the concepts of exchangeability and independence with a simple example.
- Section 5.3 derives the general hypergeometric distribution.
- Section 5.4 introduces the Heath-Sudderth and De-Finetti Representations, which connects the concepts of Logical Independence and Exchangeability.
- Section 5.5 we assume Johnson's postulate for our primordial sample space which implies that we should make linear predictions for future trials. This is equivalent to stating that we should choose the Pólya urn distribution as our prior. The most general Pólya urn is the negative hypergeometric distribution, which casts new light on the Laplace-De Finetti representation theorem.
- In Section 5.6 we then try to address the topic of choosing priors in general parameter spaces.

[2] The special status of $p(\boldsymbol{n}|\boldsymbol{m},\mathcal{K})$ is also indicated by the fact that we never sum over $\boldsymbol{n}$ since it is a single unchangeable vector.

5.2 Urn example

One important likelihood for discrete $\boldsymbol{m}$ is the multivariate hypergeometric distribution. To introduce it and the crucial concept of exchangeability, we first look at the following simple example:

> `Example knowledge base` $\mathcal{H}$`:`
> `Let there be an urn with five balls (`M`=5) of two different physical characteristics (`B`=2). Two of the balls are blue (`m_1`=2) and three are red (`m_2`=3). We have mapped the physical characteristics onto integers` $b \in \{0,1\}$`. Apart from their colour, the balls cannot be distinguished. Any ball drawn will not be placed back into the urn, so that a maximum of 5 draws can be performed.`

Consider an experiment under $\mathcal{H}$ where three balls are drawn (R=3) at times $j \in \{1,2,3\}$. Since we cannot tell balls of the same colour apart, there are 5 different balls to be drawn on the first trial, four on the second and three on the last. Also there are six different ways to permute this sequences thus a total of $5^{\underline{3}}/3! = 10$ different sequences could be drawn. Denote $\mathcal{B}_j$ as the result that a blue ball was drawn on the jth draw, and $\mathcal{R}_j$ that a red ball was drawn. By successive use of the product rule, the probability of typical sequence $\{\mathcal{R}_3, \mathcal{B}_2, \mathcal{R}_1\}$ is

$$\begin{aligned} p(\mathcal{R}_3, \mathcal{B}_2, \mathcal{R}_1|\mathcal{H}) &= p(\mathcal{R}_3|\mathcal{B}_2, \mathcal{R}_1\mathcal{H})\, p(\mathcal{B}_2|\mathcal{R}_1, \mathcal{H})\, p(\mathcal{R}_1|\mathcal{H}) \\ &= \left(\frac{2}{3}\right)\left(\frac{2}{4}\right)\left(\frac{3}{5}\right) \;=\; \left(\frac{m_2-1}{M-2}\right)\left(\frac{m_1}{M-1}\right)\left(\frac{m_2}{M}\right), \end{aligned} \tag{5.2.1}$$

The sequence $\{\mathcal{R}_3, \mathcal{R}_2, \mathcal{B}_1\}$ has probability

$$p(\mathcal{R}_3, \mathcal{R}_2, \mathcal{B}_1|\mathcal{H}) = \left(\frac{2}{3}\right)\left(\frac{3}{4}\right)\left(\frac{2}{5}\right) \;=\; \left(\frac{m_2-1}{M-2}\right)\left(\frac{m_2}{M-1}\right)\left(\frac{m_1}{M}\right). \tag{5.2.2}$$

The two sequences $\{\mathcal{R}_3, \mathcal{B}_2, \mathcal{R}_1\}$ and $\{\mathcal{R}_3, \mathcal{R}_2, \mathcal{B}_1\}$ are *exchangeable* under $\mathcal{H}$ since they result in the same probability

$$p(\mathcal{R}_3, \mathcal{B}_2, \mathcal{R}_1|\mathcal{H}) = p(\mathcal{R}_3, \mathcal{R}_2, \mathcal{B}_1|\mathcal{H}) = \frac{m_1 m_2 (m_2-1)}{M^{\underline{3}}}, \tag{5.2.3}$$

even though the individual draws are far from independent. A draw with replacement (knowledge base $\mathcal{H}'$) would imply that each draw's result has a probability independent of the previous results, yielding the factorisation of probability which defines logical independence; the two probabilities

$$\begin{aligned} p(\mathcal{R}_3, \mathcal{B}_2, \mathcal{R}_1|\mathcal{H}') &= p(\mathcal{R}_3|\mathcal{B}_2, \mathcal{R}_1\mathcal{H}')\, p(\mathcal{B}_2|\mathcal{R}_1, \mathcal{H}')\, p(\mathcal{R}_1|\mathcal{H}') \\ &= p(\mathcal{R}_3|\mathcal{H}')\, p(\mathcal{B}_2|\mathcal{H}')\, p(\mathcal{R}_1|\mathcal{H}') \\ p(\mathcal{R}_3, \mathcal{R}_2, \mathcal{B}_1|\mathcal{H}) &= p(\mathcal{R}_3|\mathcal{H}')\, p(\mathcal{R}_2|\mathcal{H}')\, p(\mathcal{B}_1|\mathcal{H}') \end{aligned} \tag{5.2.4}$$

are obviously equal to each other and to $m_1 m_2^2/M^3$. Sequences from $\mathcal{H}$ and from $\mathcal{H}'$ are hence both exchangeable, giving the same probabilities independent of the order of

the individual results but knowledge base $\mathcal{H}$ yields logically dependent sequences with correspondingly different probability assignments

$$\begin{aligned} p(\mathcal{R}_3, \mathcal{B}_2, \mathcal{R}_1|\mathcal{H}) &= \frac{1}{M^{\underline{3}}} \prod_{b=1}^{2} m_b^{\underline{r_b}} = \frac{1}{5}, \\ p(\mathcal{R}_3, \mathcal{B}_2, \mathcal{R}_1|\mathcal{H}') &= \frac{1}{M^3} \prod_{b=1}^{2} m_b^{r_b} = \frac{18}{125}. \end{aligned} \tag{5.2.5}$$

In order to project from ordered results $\boldsymbol{x} = (x_1, x_2, x_3)$ onto occupation numbers $\{r_1, r_2\}$, as in Chapter 2, we must add the probabilities for all possible orderings. In our little example there are three possible orderings of two red and one blue draw corresponding to the multinomial coefficient $3!/(1!\,2!)$, so the probability of drawing one blue and two red balls in any order is

$$p(r_1{=}1, r_2{=}2|\mathcal{H}) = 3\, p(\mathcal{R}_3, \mathcal{B}_2, \mathcal{R}_1|\mathcal{H}) = \frac{3!}{1!2!} \frac{m_1^{\underline{1}} m_2^{\underline{2}}}{M^{\underline{3}}} \tag{5.2.6}$$

which can be written in the form of a hypergeometric distribution

$$p(r_1, r_2|\mathcal{H}) = \frac{R!}{M^{\underline{R}}} \prod_{b=1}^{2} \frac{m_b^{\underline{r_b}}}{r_b!} = \binom{M}{R}^{-1} \prod_{b=1}^{2} \binom{m_b}{r_b}. \tag{5.2.7}$$

5.3 Hypergeometric distribution

The knowledge base for the generalised case is

> `Knowledge base` $\mathcal{H}$`:`
> `Let there be an urn with` M `balls exhibiting` B `different physical characteristics labelled by` $b \in \{0, 1, \ldots, B{-}1\}$`. Let` $\{m_b\}_{b=1}^{B}$ `be the set of known numbers of balls with characteristic` b `summing to` $\sum_b m_b = M$`. Balls once drawn are not replaced.`

The multivariate hypergeometric distribution follows directly from $\mathcal{H}$. Let $\mathcal{S}_{b,i}$ be the logical statement "$x_i = b$". By the *principle of indifference*, the probability of seeing an outcome c on the first trial is

$$p(\mathcal{S}_{c,1}|\mathcal{H}) = \frac{m_c}{M}. \tag{5.3.1}$$

Once $\mathcal{S}_{c,1}$ is known, the size of the outcome space decreases from M to $M{-}1$ and m_b decreases to $m_b - \delta_{b,c}$, so that on the second trial the principle of indifference yields

$$p(\mathcal{S}_{b,2}|\mathcal{S}_{c,1}, \mathcal{H}) = \frac{m_b - \delta_{b,c}}{M{-}1}. \tag{5.3.2}$$

Given that individual outcomes $x_i, i \in \{1, \ldots, R\}$ can be re-ordered at will, we place all the $x_i = 0$ outcomes first, followed by the $x_i = 1$ outcomes and so on, yielding occupation numbers $\boldsymbol{r} = \{r_1, r_2, \ldots, r_B\}$ with $\sum_b r_b = R$. Again iteratively expanding the joint

probability $p(x_1,\ldots,x_R|\mathcal{H})$ in terms of the product rule, the first subset of r_1 draws all with $x_i=0$ yield

$$p(\mathcal{S}_{0,1},\ldots,\mathcal{S}_{0,r_1}|\mathcal{H}) = \frac{m_1^{\underline{r_1}}}{M^{\underline{r_1}}}, \tag{5.3.3}$$

while the second subset of draws all with $x_i=1$ starts with an urn containing $M-m_1$ balls, so that

$$p(\mathcal{S}_{1,r_1+1},\ldots,\mathcal{S}_{1,r_1+r_2}|\mathcal{S}_{0,1},\ldots,\mathcal{S}_{0,r_1}\mathcal{H}) = \frac{m_2^{\underline{r_2}}}{(M-m_1)^{\underline{r_2}}}, \tag{5.3.4}$$

and $p(\mathcal{S}_{2,r_1+r_2+1},\ldots,\mathcal{S}_{2,r_1+r_2+r_3}|\mathcal{S}_{1,r_1+1},\ldots,\mathcal{S}_{1,r_1+r_2},\mathcal{S}_{0,1},\ldots,\mathcal{S}_{0,r_1}\mathcal{H})$ follows in the same way. This process is continued for all outcomes up to $b=(B-1)$. The joint probability for all draws is hence, for the ordered sequence $\boldsymbol{x}=\{x_1,\ldots,x_R\}$

$$p(\boldsymbol{x}|\boldsymbol{m},M,R,\mathcal{H}) = \left(\frac{m_1^{\underline{r_1}}}{M^{\underline{r_1}}}\right)\left(\frac{m_2^{\underline{r_2}}}{(M-m_1)^{\underline{r_2}}}\right)\cdots\left(\frac{m_{B-1}^{\underline{r_{B-1}}}}{(M-\sum_{b=0}^{B-2}m_b)^{\underline{r_{B-1}}}}\right) \tag{5.3.5}$$

and, projecting onto occupation numbers $\boldsymbol{r}$, we again add the probabilities of the $R!/\prod_b r_b!$ possible orderings of $\boldsymbol{x}$ to obtain the multivariate hypergeometric likelihood

$$p(\boldsymbol{r}|\boldsymbol{m},M,R,\mathcal{H}) = \frac{R!}{M^{\underline{R}}}\prod_b \frac{m_b^{\underline{r_b}}}{r_b!} = \binom{M}{R}^{-1}\prod_b\binom{m_b}{r_b} \tag{5.3.6}$$

where we have now made explicit the dependencies on R, M and $\boldsymbol{m}$.

5.4 Heath-Suddert and Laplace-de-Finetti representations

As indicated in (5.1.2) and (5.1.3), we need to calculate the *evidence* $p(\boldsymbol{n}|\mathcal{K})$. For the sake of generality and since they are mathematically the same, we shall calculate the prediction evidence $p(\boldsymbol{r}|\mathcal{K})$, or more accurately $p(\boldsymbol{r}|R,M,\mathcal{H})$ for the class of models we are treating, and substitute $\boldsymbol{n}$ for $\boldsymbol{r}$ when real data needs to be processed. For the multivariate hypergeometric likelihood (5.3.6), this evidence is

$$\begin{aligned} p(\boldsymbol{r}|R,\mathcal{H}) &= \sum_{U(\boldsymbol{m})}^{M} p(\boldsymbol{r},\boldsymbol{m}|R,M,\mathcal{H}) \;=\; \sum_{U(\boldsymbol{m})}^{M} p(\boldsymbol{r}|\boldsymbol{m},R,M,\mathcal{H})\; p(\boldsymbol{m}|R,M,\mathcal{H}) \\ &= \sum_{U(\boldsymbol{m})}^{M}\binom{M}{R}^{-1}\prod_b\binom{m_b}{r_b}\; p(\boldsymbol{m}|M,\mathcal{H}) \end{aligned} \tag{5.4.1}$$

This is de Finetti's theorem for finite sequences, which we wil call the *Heath-Sudderth* representation (HS) after Heath and Sudderth (1976). The theorem part consists of pointing out that we can use the HS representation for any finite exchangeable sequence, which was De Finetti's insight i.e. exchangeability is strong enough to characterise the hypergeometric distribution.

The de Finetti theorem proper follows on taking the urn populations to infinity with the outcome fractions $\rho_b = m_b/M$ tending to a constant limit. For $M \gg R$ and $m_b \gg r_b$, the hypergeometric reduces to the multinomial,

$$p(\boldsymbol{r}|\boldsymbol{m}, M, R, \mathcal{H}) = \frac{R!}{M^{\underline{R}}} \prod_b \frac{m_b^{\underline{r_b}}}{r_b!} \simeq \frac{R!}{M^R} \prod_b \frac{m_b^{r_b}}{r_b!} = R! \prod_b \frac{1}{r_b!} \left(\frac{m_b}{M}\right)^{r_b}, \tag{5.4.2}$$

in which case the above can be written as an weighted integral over multinomials

$$p(\boldsymbol{r}|R, \mathcal{H}) = \int R! \prod_b \frac{1}{r_b!} \left(\frac{m_b}{M}\right)^{r_b} dF_M(\boldsymbol{\rho}) \tag{5.4.3}$$

where $F_M(\boldsymbol{\rho})$ is the finite-M cumulative frequency distribution of $\boldsymbol{\rho}$ which increases by $p(\boldsymbol{m}|\mathcal{H})$ at the points $\boldsymbol{\rho} = \boldsymbol{m}/M$ (for more about taking such limit see Feller (1974)). The sequence of $F_M(\boldsymbol{\rho})$ is shown to tend uniformly to $F(\boldsymbol{\rho})$ and one obtains

$$p(\boldsymbol{r}|R, \mathcal{H}) = \int R! \prod_b \frac{\rho_b^{r_b}}{r_b!} dF(\boldsymbol{\rho}) \;=\; \int R! \prod_b \frac{\rho_b^{r_b}}{r_b!} p(\boldsymbol{\rho}|\mathcal{H})\, d\boldsymbol{\rho} \tag{5.4.4}$$

which is De Finetti's theorem for infinite exchangeable sequences and which we shall call the Laplace-de-Finetti (LdF) representation.

In the literature, the de Finetti theorem is considered significant as showing that every exchangeable sequence (as represented by $p(\boldsymbol{r}|\boldsymbol{R}, \mathcal{H})$) that is part of an infinitely long sequence can be written as a mixture of multinomial distributions weighted by an arbitrary normalised function (as represented by the prior $p(\boldsymbol{\rho}|\mathcal{H})$). The HS representation is similarly viewed as saying that every exchangeable sequence can be represented as a mixture of hypergeometric distributions weighted by a probability $p(\boldsymbol{m}|\mathcal{H})$. Indeed, it may be of interest to infer the priors $p(\boldsymbol{\rho}|\mathcal{H})$ or $p(\boldsymbol{m}|\mathcal{H})$ from known likelihoods $p(\boldsymbol{r}|\boldsymbol{\rho}, \mathcal{H})$ or $p(\boldsymbol{r}|\boldsymbol{m}, \mathcal{H})$. These theorems are usually celebrated because they ensure the existence of a prior distribution which was a historical criticism of Bayesian methods.

For our purposes, however, the HS and LdF representations are of practical interest in calculating the evidence by summing/integrating over all possible configurations of $\boldsymbol{m}$ or $\boldsymbol{\rho}$. We will return to the de Finetti theorem later in connection with the negative hypergeometric distribution.

5.5 Priors for exchangeable sequences

5.5.1 Johnson's postulate

Having made at least a preliminary determination of the likelihood $p(\boldsymbol{r}|\boldsymbol{m}, \mathcal{H})$, we now must tackle the huge topic of finding the priors $p(\boldsymbol{m}|\mathcal{H})$ and $p(\boldsymbol{\rho}|\mathcal{H})$. In this and the next section, we proceed to construct a prior and a primordial sample space for exchangeable trials in contrast to the prior and sample space for logically independent trials. Naturally, there will be not a single best prior but rather as many different priors as accurately reflect different states of knowledge.

A first step at constructing priors based on knowledge is to understand priors based on ignorance, the lack of knowledge. Historically, for his one-dimensional binomial likelihood $p(\boldsymbol{r}|\boldsymbol{\rho}) = \binom{R}{r}\rho^r(1-\rho)^{R-r}$, Bayes (1763) as quoted by Stigler (1982) argued that the evidence should be uniform and thus the prior distribution should also be uniform $p(\rho) = 1$, while Laplace (1774) again as quoted by Stigler (1986) used the same uniform prior invoking the principle of insufficient reason directly on the parameter space. The set $\boldsymbol{\rho} = \{\rho_1, \ldots, \rho_B\}$ entering the multinomial and multivariate hypergeometric lives in the $(B-1)$-dimensional simplex $0 < \rho_b < 1$, $\sum_b \rho_b = 1$ which can be shown to have a volume $1/\Gamma[B]$, so the uniform prior for a B-dimensional $\boldsymbol{\rho}$ is $p(\boldsymbol{\rho}|\mathcal{H}) = \Gamma[B]$. We will call the prior uniform on the simplex the Bayes-Laplace prior.

Uniformity does not, however, necessarily represent complete ignorance, as already pointed out by Edgeworth (1884) and Pearson (1907), who adopted the view that uniformity already contains knowledge of the limits of the parameter space. Jaynes (2003) and before him Jeffreys (1961), Haldane (1931) and others readily adopted and utilised this standpoint to define different sets of *ignorance priors*.

First, however, we must discuss the more general *Johnson's sufficiency postulate* as published by Johnson (1932) which states that the probability for a given trial to have a result b can depend only on the total existing occupation number for that result. For a state of ignorance this assumption is easy to motivate; for us to connect different bins to each other requires some information which for the primordial sample space is unavailable. If we have R data points $\boldsymbol{x} = \{x_1, \ldots, x_R\}$ with occupation number $r_b = \sum_{i=1}^{R} \delta(x_i, \mathcal{A}_b)$, then the probability for the $(R+1)$st result must obey

$$p(x_{R+1}{=}b|\boldsymbol{x}) = f_b(r_b), \tag{5.5.1}$$

where the $\{f_b\}$ can be any set of functions as long as they obey $0 \leq f_b(r_b)$ and, due to the normalisation condition for $\sum_b p(x_{R+1}{=}b|\boldsymbol{x})$,

$$\sum_b f_b(r_b) = 1. \tag{5.5.2}$$

It is easy to show that the f_b must be linear functions of the respective r_b: for a given occupation number define first

$$\boldsymbol{r} = \{r_1, \ldots, r_b, \ldots, r_c, \ldots, r_d, \ldots, r_B\}, \tag{5.5.3}$$

where b, c and d are arbitrary but fixed indices, and then

$$\begin{aligned}
\boldsymbol{r}^{(1)} &= \{r_1, \ldots, r_b{+}1, \ldots, r_c{-}1, \ldots, r_d, \ldots, r_B\}, \\
\boldsymbol{r}^{(2)} &= \{r_1, \ldots, r_b, \ldots, r_c{-}1, \ldots, r_d{+}1, \ldots, r_B\}, \\
\boldsymbol{r}^{(3)} &= \{r_1, \ldots, r_b{-}1, \ldots, r_c, \ldots, r_d{+}1, \ldots, r_B\},
\end{aligned} \tag{5.5.4}$$

so that $\sum_b r_b = R$ for each. Subtracting the sum of predictions of $\boldsymbol{r}$ from each of the other vectors in turn yields

$$f_b(r_b{+}1) - f_b(r_b) = f_c(r_c) - f_c(r_c{-}1) = f_d(r_d{+}1) - f_d(r_d) = f_b(r_b) - f_b(r_b{-}1), \tag{5.5.5}$$

meaning that

$$f_b(r_b) = \alpha_b^* + \beta^* r_b, \tag{5.5.6}$$

where $\alpha_b^* = f_b(0) \geq 0$ as f_b is a probability and $\beta^* = f_b(r_b{+}1) - f_b(r_b)$ must be independent of b.

5.5.2 Pólya urn distributions and their properties

Below, we shall use Pólya urn distributions in various forms and guises for priors, and hence briefly review their properties. In generic notation, the Pólya urn is filled with K balls where k_b balls have property b, and the Pólya urn distribution then specifies the probability for drawing $\boldsymbol{n} = \{n_1, \ldots, n_B\}$ balls,

$$p(\boldsymbol{n}|\boldsymbol{k}, c, \mathcal{H}) = \binom{-A/c}{N}^{-1} \prod_{b=1}^{B} \binom{-a_b/c}{n_b}. \tag{5.5.7}$$

The parameter c controls the method of sampling and sets the type of distribution, namely

c	Algorithm	Resulting distribution
-1	do not replace ball	hypergeometric
0	replace ball	multinomial
+1	replace ball and add another of the same b	negative hypergeometric

The multinomial is recovered as the limit $c \to 0$ while setting $\rho_b = a_b/A$,

$$\lim_{c\to 0} p(\boldsymbol{n}|\boldsymbol{a}, c, \mathcal{H}) = \lim_{c\to 0} \frac{N!}{n_1! \ldots n_B!} \frac{(a_1/c)^{\overline{n_1}} \ldots (a_B/c)^{\overline{n_B}}}{(A/c)^{\overline{N}}} = N! \prod_b \frac{\rho_b^{n_b}}{n_b!}. \tag{5.5.8}$$

The respective outcome space for the Pólya urn is $U(\boldsymbol{n})$, which for $c = -1$ implies that all the $a_b \geq n_b$.

5.5.3 Pólya urns and Johnson's postulate

We now show that a Pólya urn prior together with a hypergeometric likelihood satisfy Johnson's postulate. We must show that the prediction for x_{R+1}, given ordered outcomes $\boldsymbol{x}$,

$$p(x_{R+1}{=}d \,|\, \boldsymbol{x}, \boldsymbol{a}, c, \mathcal{H}) = \frac{p(x_{R+1}{=}d, \boldsymbol{x}|\boldsymbol{a}, c, \mathcal{H})}{p(\boldsymbol{x}|\boldsymbol{a}, c, \mathcal{H})} \tag{5.5.9}$$

for a hypergeometric likelihood $p(\boldsymbol{r}|\boldsymbol{m}, \mathcal{H})$ and Pólya prior (5.5.7) for occupation number $\boldsymbol{r}$ is a linear function of r_d, for all c. Rearranging and using the normalisation condition

in the second step, we obtain for the evidence another Pólya distribution

$$\begin{aligned} p(\boldsymbol{r}|\boldsymbol{a},c,\mathcal{H}) &= \sum_{U(\boldsymbol{m})} p(\boldsymbol{r}|\boldsymbol{m},\mathcal{H})\, p(\boldsymbol{m}|\boldsymbol{a},c,\mathcal{H}) \\ &= \sum_{U(\boldsymbol{m})} \left[\binom{M}{R}^{-1}\prod_b \binom{m_b}{r_b}\right]\left[\binom{-A/c}{M}^{-1}\prod_b \binom{-a_b/c}{m_b}\right] \\ &= \sum_{U(\boldsymbol{m})} \binom{-A/c}{R}^{-1}\binom{-A/c-R}{M-R}^{-1}\prod_b \binom{-a_b/c}{r_b}\binom{-a_b/c-r_b}{m_b-r_b} \\ &= \binom{-A/c}{R}^{-1}\prod_b \binom{-a_b/c}{r_b} \end{aligned} \tag{5.5.10}$$

The corresponding evidence for ordered outcomes $\boldsymbol{x}$ in the denominator of (5.5.10) is $p(\boldsymbol{r}|\boldsymbol{a},c,\mathcal{H})$ divided by the multinomial coefficient $R!/\prod_b r_b!$ or, using (2.1.18) and (2.1.16),

$$p(\boldsymbol{x}|\boldsymbol{a},c,\mathcal{H}) = \frac{1}{(A/c)^{\overline{R}}}\prod_b (a_b/c)^{\overline{r_b}}. \tag{5.5.11}$$

Since the numerator in (5.5.10) is just the same evidence with a lengthened vector $\boldsymbol{x}' = \{x_{R+1},\boldsymbol{x}\}$, the prediction for the next trial is,

$$f_d(r_d) = \frac{(A/c)^{\overline{R}}}{(A/c)^{\overline{R+1}}}\cdot\frac{(a_d/c)^{\overline{r_d+1}}}{(a_d/c)^{\overline{r_d}}} = \frac{(a_d/c)+r_d}{(A/c)+R}, \tag{5.5.12}$$

and hence Johnson's postulate (5.5.6) is satisfied with $\alpha_b^* = a_b/(A+cR)$ and $\beta^* = c/(A+cR)$.

The Pólya urn is thus an appropriate prior for multivariate hypergeometric likelihoods and hence for exchangeable distribution likelihoods. The status of Pólya urn distributions in relation to exchangeability does not contradict cases of logical independence but includes them, in the form of the $c\to 0$ limit taken in (5.5.8). Of course the assumption of Logical Independence is not wrong but it is not the general case, as we have already demonstrated.

The generic posterior is

$$\begin{aligned} p(\boldsymbol{m}|\boldsymbol{r},\boldsymbol{a},c,\mathcal{H}) &= \frac{p(\boldsymbol{r},\boldsymbol{m}|\boldsymbol{a},c,\mathcal{H})}{p(\boldsymbol{r}|\boldsymbol{a},c,\mathcal{H})} \\ &= \binom{-A/c-R}{M-R}^{-1}\prod_b \binom{-a_b/c-r_b}{m_b-r_b}. \end{aligned} \tag{5.5.13}$$

5.5.4 Pólya urn as conjugate prior for hypergeometric sampling

A further remarkable property of the Pólya urn is that it is closed under hypergeometric sampling; in other words, the posterior and prior have the same functional form. A prior closed under sampling is called a **conjugate** prior, see Raiffa and Schlaifer (1961). (This invariance may be called a "symmetry" under the transformation specified by Bayes theorem. For more about conjugate prior distributions see also O'Hagan and Forster (2004).) Usually a conjugate prior would lead to linear posterior mean estimates, see

Diaconis and Ylvisaker (1979), and thus Johnson's postulate also implies that we should pick a conjugate prior for hypergeometric sampling.

According to Hald (1960), there are three conjugate prior distributions for hypergeometric sampling, namely Hypergeometric, Negative Hypergeometric and the Multinomial. As we have seen, the Pólya Urn parametrises all three cases.

5.5.4.1 Case $c = -1$: hypergeometric prior and posterior

A hypergeometric prior with hypergeometric sampling

$$p(\boldsymbol{m}|\boldsymbol{a},\mathcal{H}) = \binom{A}{M}^{-1}\prod_b \binom{a_b}{m_b} \qquad p(\boldsymbol{r}|\boldsymbol{m},\mathcal{H}) = \binom{M}{R}^{-1}\prod_b \binom{m_b}{r_b}, \tag{5.5.14}$$

leads to a hypergeometric evidence and a hypergeometric posterior,

$$p(\boldsymbol{r}|\boldsymbol{a},\mathcal{H}) = \binom{A}{R}^{-1}\prod_b \binom{a_b}{r_b} \quad p(\boldsymbol{m}|\boldsymbol{r},\boldsymbol{a},\mathcal{H}) = \binom{A-R}{M-R}^{-1}\prod_b \binom{a_b-r_b}{m_b-r_b}. \tag{5.5.15}$$

The defining property of the Hypergeometric distribution is that the population is finite. If we try to take more trials than the population the hypergeometric distribution will start to assign zero probability to those trials. Consequently the hypergeometric prior represents a very specific state of knowledge where we know the population sizes and is hence a special case and not the general formulation that we are seeking at the moment.

5.5.4.2 Case $c = +1$: negative hypergeometric prior and posterior

A negative hypergeometric prior with hypergeometric sampling

$$p(\boldsymbol{m}|\boldsymbol{a},\mathcal{H}) = \binom{-A}{M}^{-1}\prod_b \binom{-a_b}{m_b} \qquad p(\boldsymbol{r}|\boldsymbol{m},\mathcal{H}) = \binom{M}{R}^{-1}\prod_b \binom{m_b}{r_b}, \tag{5.5.16}$$

leads to a negative hypergeometric evidence and negative hypergeometric posterior

$$p(\boldsymbol{r}|\boldsymbol{a},\mathcal{H}) = \binom{-A}{R}^{-1}\prod_b \binom{-a_b}{r_b} \quad p(\boldsymbol{m}|\boldsymbol{r},\boldsymbol{a},\mathcal{H}) = \binom{-A-R}{M}^{-1}\prod_b \binom{-a_b-m_b}{r_b}. \tag{5.5.17}$$

The negative hypergeometric prior distribution also has the remarkable property that we can decompose it into a Multinomial-Dirichlet mixture,

$$p(\boldsymbol{r}|\boldsymbol{a},\mathcal{H}) = \int \left[R!\prod_b \frac{\rho_b^{r_b}}{r_b!}\right]\left[\Gamma[A]\prod_b \frac{\rho_b^{a_b-1}}{\Gamma[a_b]}\right]\prod_b d\rho_b \;=\; \binom{-A}{R}^{-1}\prod_b \binom{-a_b}{r_b}. \tag{5.5.18}$$

Comparison with (5.4.3) shows that this equation is just the Laplace-de-Finetti representation with a Dirichlet prior distribution for $\boldsymbol{\rho}$. Choosing a distribution for $\boldsymbol{\rho}$ implies that we are uncertain about the values of $\boldsymbol{\rho}$. Contrast this with the multinomial choice where we specify a fixed $\boldsymbol{\rho}$ indicating certainty and the hypergeometric evidence where the populations numbers from which are $\boldsymbol{m}$ are drawn is known and fixed. Consequently the negative

hypergeometric distribution corresponds to greater uncertainty and thus greater generality than the other two choices of prior distributions. It also shows that the Laplace-De Finetti representation is more general than we would suppose. We do not need to assume that our sequences are subsets from an infinite sequence but only that we do not know whether they originate from a finite population. Operationally if we think the number of trials R can be extended to an arbitrarily large number then we can use the Laplace-De Finetti representation with a Dirichlet prior.

This still leaves us with the question of choosing the metaparameters $\boldsymbol{a}$ in the Dirichlet prior distribution,

$$p(\boldsymbol{\rho}|\boldsymbol{a},\mathcal{H}) = \Gamma\left[\textstyle\sum_b a_b\right] \prod_b \frac{\rho_b^{a_b-1}}{\Gamma[a_b]}, \qquad a_b > 0 \;\forall\, b. \tag{5.5.19}$$

This will be addressed in Section 5.6.

5.6 Priors for metaparameters

Let us recap: The previous chapter concerned the specification of forward probabilities of our theory. As we have seen, exchangeability allows us to always pick the likelihood or sampling to be a hypergeometric distribution. In the second part of the chapter, we then sought a general prior distribution with maximal ignorance, resulting in the choice of a negative hypergeometric prior. Equivalently, the Multinomial-Dirichlet mixture could be used to assign probabilities.

While we have thus chosen to fix $c = 1$ for the prior, we are still left with the unspecified metaparameters $\boldsymbol{a}$ which we can view in the LdF representation as initial counts for the multinomial distribution. Either way, the chain of reasoning of general formulae (5.1.1)–(5.1.3) forces us to choose a specific set of $\boldsymbol{a}$ as our initial values

$$p(\boldsymbol{n}|\boldsymbol{a},\mathcal{H}) = \sum_{U(\boldsymbol{m})} p(\boldsymbol{n}|\boldsymbol{m},\boldsymbol{a},\mathcal{H})\, p(\boldsymbol{m}|\boldsymbol{a},\mathcal{H}), \tag{5.6.1}$$

likewise the posterior and the prediction

$$p(\boldsymbol{m}|\boldsymbol{n},\boldsymbol{a},\mathcal{H}) = \frac{1}{p(\boldsymbol{n}|\boldsymbol{a},\mathcal{H})} p(\boldsymbol{n}|\boldsymbol{m},\boldsymbol{a},\mathcal{K})\, p(\boldsymbol{m}|\boldsymbol{a},\mathcal{H}), \tag{5.6.2}$$

$$p(\boldsymbol{r}|\boldsymbol{n},\boldsymbol{a},\mathcal{H}) = \sum_{U(\boldsymbol{m})} p(\boldsymbol{r}|\boldsymbol{m},\mathcal{H})\, p(\boldsymbol{m}|\boldsymbol{n},\boldsymbol{a},\mathcal{H}). \tag{5.6.3}$$

The process of finding priors can be a very mollusc-like argument: We analyse a parameter space in terms of metaparameters and can then in turn analyse these metaparameters in terms of metametaparameters and so forth, thereby creating a potentially infinite hierarchy of unknowns. As usual, we have to make choices at some point that stops the infinite recursion. In this section, we draw on the powerful principle of *invariance* or *symmetry* to provide such a basis. We define *invariance* and *symmetry* in information theoretic terms as *conservation of ignorance under change*, i.e. a change in parameter under some transformation leaves us none the wiser or the information invariant. We cannot distinguish

between one situation and the other. We attempt to provide a partial list of possible metaparameter priors based on some situations of ignorance or information symmetries that may occur in various applications.

Ignorance is, of course, not the only game in town. If pertinent information is available for constructing a parameter prior distribution then obviously that information should be used.

In the context of this chapter, all the variables used below such as angles, ρ's, location and scale parameters etc are special cases of $\boldsymbol{a}$. The symmetry arguments used below can, however, be used whenever a prior in any situation needs to be specified from first principles.

5.6.1 Unknown Probabilities or Chances

The first class of symmetries concerns ignorance of elementary probabilities $\boldsymbol{\rho}$ as used from the start of probability theory. Traditionally, four choices have been advocated in the literature:

The first choice is actually the baseline which assumes not ignorance but perfect knowledge,

$$p(\boldsymbol{\rho}|\mathcal{I})\, d\boldsymbol{\rho} = \prod_b \delta(\rho_b - C_b) d\rho_b, \tag{5.6.4}$$

where $\{C_b\}$ are fixed numbers, assumed known. Most elementary statistics books naively assume this to be self-evident — to the degree that the rather ridiculous assumptions underlying this choice, namely infinitely accurate knowledge even before any data is available, are never even discussed. This assumption of exactly known $\boldsymbol{\rho}$ formed the basis of Chapter 4, which is exactly why the present chapter and the next needed to be written.[3]

The second and superior choice is the Bayes-Laplace or uniform prior which is based on arguing that the evidence should be invariant under permutation; see Stigler (1982),

$$p(\boldsymbol{\rho}|\mathcal{I})\, d\boldsymbol{\rho} = (B-1)!\, d\boldsymbol{\rho},$$

which is the special $a_b = 1\ \forall\ b$ case of the Dirichlet prior (5.5.19).

The third choice is the Jaynes-Haldane prior, introduced by Haldane (1931) and motivated by Jaynes (1968). It is based on assumed invariance of odds under Bayes' theorem on going from prior to posterior (3.7.2),

$$\frac{p(\mathcal{H}|\mathcal{D},\mathcal{K})}{p(\overline{\mathcal{H}}|\mathcal{D},\mathcal{K})} = \frac{p(\mathcal{D}|\mathcal{H},\mathcal{K})}{p(\mathcal{D}|\overline{\mathcal{H}},\mathcal{K})}\, \frac{p(\mathcal{H}|\mathcal{K})}{p(\overline{\mathcal{H}}|\mathcal{K})}, \tag{5.6.5}$$

where $\mathcal{H}$ is the hypothesis and $\mathcal{D}$ is the data. Writing o for the prior odds for the hypothesis and o' for the posterior odds and ℓ for the likelihood ratio, this becomes

$$o' = \ell \cdot o. \tag{5.6.6}$$

[3] Of course traditional fitting algorithms and methods soften this assumption and even lead to good results in many cases. It is not the purpose of this dissertation, however, to repeat these conventional methods.

We start with the odds o for our hypothesis which is then updated with the likelihood ratio ℓ after seeing the data $\mathcal{D}$, giving the posterior odds o'. This gives us a transformation rule for invariance of odds. Transforming back into probabilities $\rho \equiv p(\mathcal{H}|\mathcal{K})$ and $\rho' = p(\mathcal{H}|\mathcal{D},\mathcal{K})$,

$$\frac{\rho'}{1-\rho'} = \ell \frac{\rho}{1-\rho}, \tag{5.6.7}$$

and solving, we obtain

$$\rho' = \frac{\ell\rho}{1-\rho+\ell\rho}. \tag{5.6.8}$$

Now imagine the same data $\mathcal{D}$ is considered by many individuals, all of whom independently assign a prior ρ and initial odds o and all consistently update their new odds o' according to the same mechanism (5.6.6). There is thus a "metaprobability" distribution $f(\rho)\,d\rho$ of individuals assigning a particular ρ. Total ignorance (or confusion) is then defined as saying that the information given to these individuals in the form of ℓ contains no consistent learning value for the population as a whole, leaving the distribution of individual knowledge the same state f as before. The distribution of priors $f[\rho]d\rho$ and of posteriors $f[\rho']d\rho'$ thus by hypothesis of ignorance have the same functional form f,

$$f[\rho]\,\mathrm{d}\rho = f[\rho']\,\mathrm{d}\rho'. \tag{5.6.9}$$

Combining this with the ρ-transformation results in

$$f[\rho](1+\ell\rho-\rho)^2\,\mathrm{d}\rho = f\left[\frac{\ell\rho}{1-\rho+\ell\rho}\right]\ell\,\mathrm{d}\rho \tag{5.6.10}$$

and taking the derivative in ℓ and setting $\ell = 1$, we obtain

$$2\rho = f[\rho] + (1-\rho)\rho f'[\rho], \tag{5.6.11}$$

and so

$$f[\rho] = \frac{\text{const.}}{\rho(1-\rho)}. \tag{5.6.12}$$

These arguments are similar to the geometric arguments as in Kendall and Moran (1963): we define a transformation and assert that our state of knowledge (in this case the odds) is invariant under that transformation. This leads to a functional equation which we then solve to find our prior distribution. This type of prior we will call an **invariance** prior. The arguments that lead to such priors are usually quite convincing but they do contain a serious flaw in that they can lead to improper priors, as in (5.6.12) above, in which the normalising constant cannot be evaluated. Fortunately, we already know from Eqs. (5.1.2)–(5.1.3) that the constant will cancel and that improper priors can, therefore, be used to good effect.

5.6.2 Unknown Angles

The fourth choice, which we will call the Jeffreys-Perks prior, is based on invariance on the circle or sphere i.e. ignorance of angles; see Jeffreys (1961) and Perks (1947). Consider assigning a probability to a binomial distribution with parameters ρ and $1-\rho$. Transform these to an angle

$$\rho = \cos^2\left(\frac{\phi}{2}\right) \qquad\qquad 1-\rho = \sin^2\left(\frac{\phi}{2}\right). \tag{5.6.13}$$

The motivation for this transformation is that it is much easier to specify a prior on a circle, because rotational invariance is such an obvious transformation. Invariance on the circle translates into

$$f[\phi] = f[\phi+\theta], \tag{5.6.14}$$

for arbitrary angle θ. Taking the derivative in θ immediately results in the uniform-angle prior

$$f'[\theta] = 0 \qquad \Rightarrow \qquad f[\theta] = \text{constant}, \tag{5.6.15}$$

but unlike the Jeffreys-Haldane prior this can be normalised so that

$$f[\phi] = \frac{1}{2\pi}. \tag{5.6.16}$$

Transforming back to our probabilities yields

$$f[\rho] = \frac{1}{\pi\sqrt{\rho(1-\rho)}}, \tag{5.6.17}$$

which is the Jeffreys-Perks prior. It also represents a useful compromise with the invariant Jaynes-Haldane prior, but with the advantage that it is normalisable.

For the trinomial case, we parametrise on the 3-sphere to automatically fulfil the normalisation condition $\rho_3 = 1-\rho_1-\rho_2$,

$$\rho_1 = \cos^2\left(\frac{\phi}{2}\right)\sin^2\left(\frac{\theta}{2}\right) \qquad \rho_2 = \sin^2\left(\frac{\phi}{2}\right)\sin^2\left(\frac{\theta}{2}\right) \qquad \rho_3 = \cos^2\left(\frac{\theta}{2}\right) \tag{5.6.18}$$

where $0 \leq \theta \leq \pi$ and $0 \leq \phi \leq 2\pi$. Transforming the rotational invariant prior distribution on a sphere,

$$p(\theta,\phi) = \frac{1}{2\pi}\sin\left(\frac{\theta}{2}\right), \tag{5.6.19}$$

to probabilities we find,

$$p(\rho_1,\rho_2) = \frac{1}{2\pi\sqrt{\rho_1\rho_2(1-\rho_1-\rho_2)}}. \tag{5.6.20}$$

For the general multinomial case, again define each ρ_b in the simplex as the square of the b-component of the B-dimensional unit sphere and invert to get

$$p(\boldsymbol{\rho}|\mathcal{H}) = \Gamma\left[\frac{B}{2}\right]\prod_b \frac{\rho_b^{-1/2}}{\Gamma[1/2]}. \tag{5.6.21}$$

From here on we shall always choose the metaparameters of the Dirichlet prior distribution to be $a_b = 1/2$ for all b.

5.6.3 Unknown location

The next class of situations concern scale and location. When we are ignorant of location, the appropriate prior should be unchanged under arbitrary translation of the location parameter μ,

$$f[\mu] = f[\mu + c]. \tag{5.6.22}$$

Taking the derivative in c as in the case of angle invariance and setting c to zero, we see that $f'[\mu] = 0$ and the only possible assignment is a constant function,

$$f[\mu] \propto \text{constant}. \tag{5.6.23}$$

In many cases this assignment works beautifully, but if we need a prior on the entire real line $(-\infty, \infty)$ or any other unbounded outcome space, the uniform prior is improper. Let us try a different tack for such cases: Consider a population of experts $k = 1, \ldots, K$, each of whom voices an independent opinion μ_k on what the value for μ should be. Since location parameters are additive, we combine the expert opinions by taking the additive mean,

$$\mu = \frac{\mu_1 + \mu_2 + \cdots + \mu_K}{K}. \tag{5.6.24}$$

The corresponding distribution for μ would be

$$p(\mu|\mathcal{H}) = \int p(\mu_1|\mathcal{H})p(\mu_2|\mathcal{H}) \cdots p(\mu_K|\mathcal{H})\ \delta\left(K^{-1}\textstyle\sum_k \mu_k - \mu\right) \displaystyle\prod_k d\mu_k. \tag{5.6.25}$$

Taking the Fourier transform on both sides in μ we get the characteristic function relation,

$$\Phi[t] = \prod_k \Phi_k\left[\frac{t}{K}\right]. \tag{5.6.26}$$

Using our consistency argument again, we now argue that every expert's opinion was based on exactly the same information and that every characteristic function must therefore be the same,

$$\Phi[t] = \Phi_k[t]. \tag{5.6.27}$$

Combining the arguments we have the functional equation,

$$\log \Phi[Kt] = K \log \Phi[t] \qquad -\infty < t < \infty, \quad K > 0. \tag{5.6.28}$$

Evidently, this requires a linear relation on the real line, and, because of the normalisation of the distribution there is no constant term. Hence

$$\log \Phi[t] = Ct, \tag{5.6.29}$$

where C is some complex number. Examining the definition of a Fourier transform we see that the characteristic function must obey the existence condition $\Phi[-t] = \Phi^*[t]$ giving us

$$\log \Phi[t] = i\alpha t - \beta|t|, \qquad -\infty < \alpha < \infty, \quad \beta > 0. \tag{5.6.30}$$

The restriction on β ensures that we have a proper distribution. Inverting the Fourier transform gives

$$p(\mu|\mathcal{H}) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp\left[i\alpha t - \beta|t| - it\mu\right] dt = \frac{1}{\pi}\frac{\beta}{\beta^2+(\alpha-\mu)^2}, \tag{5.6.31}$$

a Cauchy distribution, with median α and interquartile $\alpha \pm \beta$. This is our reference prior for locations with unknown variance, as we will see below, and we will call it a **stable** prior.

To connect our Cauchy prior with the Jeffreys-Perks prior distribution, we make the observation that an unknown angle is the ratio of two unknown lengths. If we take the first length fixed and equal to β, then we can make the transformation

$$\phi = \beta \tan\frac{\phi}{2}, \tag{5.6.32}$$

which transforms our unknown angle into an unknown location. This is a purely heuristic argument, but operationally as shown in Jeffreys and Jeffreys (1950) all we need for a location prior is a symmetrical distribution which need not have any moments. Intuitively the moments represent information on the properties of the location parameter thus by choosing a distribution whose moments diverge we refuse to provide such information. The stable argument then provides additional support for our choice.

5.6.4 Unknown location with known variance

The second transformation we will consider for location parameters is commonly called standardisation. The appropriate transformation

$$\mu = \frac{\mu_1 + \mu_2 + \cdots + \mu_K}{\sqrt{K}}. \tag{5.6.33}$$

preserves the second cumulant or variance, $\kappa_2(\mu) = \kappa_2(\mu_k)$. The underlying motivation is that every distribution has a variance $\kappa_2(\mu_k) = \sigma_k^2$ and combining them gives a variance of roughly $K\sigma^2$ which then needs to be standardised and dividing gives a quadratic change in the second cumulant. Again using the consistency argument gives us the new functional equation,

$$\sqrt{K}\log\Phi[t] = K\log\Phi\left[t\right], \tag{5.6.34}$$

yielding a quadratic relation on the real line,

$$\log\Phi[t] = Ct^2 \qquad\qquad -\infty < t < \infty, \tag{5.6.35}$$

with C a complex number. Using the existence condition $\Phi[-t] = \Phi^*[t]$ again, we find

$$\log\Phi[t] = -\beta\frac{t^2}{2} \qquad\qquad 0 < \beta < \infty. \tag{5.6.36}$$

Inverting the Fourier transform gives,

$$p(\mu|\mathcal{H}) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\beta\frac{t^2}{2}-it\mu}dt = \frac{1}{\sqrt{\pi\beta}}\exp\left[-\frac{\mu^2}{2\beta}\right], \tag{5.6.37}$$

a Gaussian distribution located at zero with variance β.

5.6.5 Unknown variance

The third case is to assign a variance to an unknown location. We are given opinions on a location μ_k on which all experts agree but with different variances σ_k^2 e.g. in the form of a product of symmetric Gaussian that we combine,

$$p(\sigma_k|\mu_k,\mathcal{K}) = \prod_k \frac{1}{\sqrt{2\pi}\sigma_k}\exp\left[-\frac{\mu_k^2}{\sigma_k^2}\right] \to \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{\mu_k^2}{\sigma^2}\right], \tag{5.6.38}$$

implying the composition rule,

$$\sigma^2 = \frac{\sigma_1^2+\sigma_2^2+\ldots\sigma_K^2}{K^2}. \tag{5.6.39}$$

The consistency argument then gives the functional equation as

$$\log\Phi\left[K^2 t\right] = K\log\Phi\left[t\right], \tag{5.6.40}$$

which gives a square root relation on the real axis,

$$\log\Phi = C\sqrt{t}, \tag{5.6.41}$$

where C is a complex number. Applying the existence condition $\Phi[-t] = \Phi^*[t]$ gives in this case

$$\log\Phi = -\beta\sqrt{-i2t} \qquad\qquad 0<\beta<\infty, \tag{5.6.42}$$

and inverting the Fourier transform gives

$$p(\sigma^2|\mathcal{H}) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\beta\sqrt{-i2t}-it\sigma^2}dt = \frac{\beta}{\sqrt{2\pi}\sigma^3}\exp\left[-\frac{\beta^2}{2\sigma^2}\right], \tag{5.6.43}$$

a Levy distribution. The reason that we know the composition rule is that if we combine the Levy prior with the Gaussian likelihood, we get back the Cauchy distribution (5.6.31)

$$\begin{aligned} p(\mu|\mathcal{H}) &= \int p(\mu|\sigma^2,\mathcal{H})p(\sigma^2|\mathcal{H})d\sigma^2 \\ &= \int\left[\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{\mu^2}{2\sigma^2}\right]\right]\left[\frac{\beta}{\sqrt{2\pi}\sigma^3}\exp\left[-\frac{\beta^2}{2\sigma^2}\right]\right]d\sigma^2 \\ &= \frac{1}{\pi}\frac{\beta}{\beta^2+\mu^2}, \end{aligned} \tag{5.6.44}$$

but with α equal to zero. We could have also argued from symmetry that α must be zero because any other value would have preferred positive or negative values.

5.6.6 Unknown scale

Since scale transformations are always multiplicative,

$$\sigma \to c\sigma \qquad c,\sigma>0, \tag{5.6.45}$$

the invariant prior should be

$$f[\sigma]d\sigma = f[c\sigma]cd\sigma. \tag{5.6.46}$$

Taking the derivative in c and setting $c = 1$ gives

$$\sigma f'[\sigma] + f[\sigma] = 0 \qquad\qquad f[\sigma] = \frac{\text{const.}}{\sigma}. \tag{5.6.47}$$

This, too, is an improper prior on the interval $[0, \infty)$, so let us try to find a proper pre-limit prior for scale. Intuitively we think of a radius as specifying a scale in an experiment; thus we try to add up the squares of location parameters to form a squared radius. We also need a spherically symmetric construction if we wish to define a radius and therefore start with multivariate extension of the Cauchy distribution,

$$p(\boldsymbol{\mu}|\mathcal{H}) = \frac{\beta\Gamma\left[\frac{1+K}{2}\right]}{\pi^{\frac{K+1}{2}}}\left(\frac{1}{\beta^2 + \mu_1^2 + \cdots + \mu_K^2}\right)^{\frac{1+K}{2}}. \tag{5.6.48}$$

This specific generalisation of the Cauchy distribution is rotationally invariant and all its marginal distribution are Cauchy as well. Projecting onto one variable via $r^2 = \mu_1^2 + \cdots + \mu_K^2$ gives us

$$p(r^2|\mathcal{H}) = \frac{\beta\Gamma\left[\frac{1+K}{2}\right]}{\sqrt{\pi}\Gamma\left[\frac{K}{2}\right]}\left(\frac{r^2}{\beta^2 + r^2}\right)^{\frac{1+K}{2}} r^{-3} \tag{5.6.49}$$

and in one dimension,

$$p(r^2|\mathcal{H}) = \frac{\beta}{\pi}\frac{1}{r(\beta^2 + r^2)}, \tag{5.6.50}$$

which is a folded Cauchy distribution which we will use as a prior for unknown scale.

A different perspective on the scale prior is to consider the odds as a scale parameter. Transforming our arcsine law (5.6.17) to odds gives

$$\begin{aligned} \rho &= \frac{o}{1+o} & d\rho &= \frac{1}{(1+o)^2}do \\ p(o|\mathcal{H})do &= \frac{1}{\pi\sqrt{o}(1+o)}do, \end{aligned} \tag{5.6.51}$$

which is equivalent to our scale invariant prior if we choose $r^2 = o$. We can also decompose the scale prior into an informative scale prior with an uninformative prior,

$$\begin{aligned} p(o|\mathcal{H}) &= \int_0^\infty \left[\frac{e^{-o/\alpha}}{\sqrt{\pi o\alpha}}\right]\left[\frac{e^{-\beta/\alpha}}{\sqrt{\beta\pi}}\alpha^{-\frac{3}{2}}\right]d\alpha \\ &= \frac{\sqrt{\beta}}{\sqrt{o}(\beta+o)}. \end{aligned} \tag{5.6.52}$$

Additionally we observe that if a prior represents ignorance of scale then it should also hold for the inverse of that scale. In fact our prior should be invariant under the transformation $o \to 1/o'$ and thus,

$$p(o'|\mathcal{H}) = \frac{\sqrt{\beta}}{\pi\sqrt{o'}(1+o'\beta)}, \tag{5.6.53}$$

which implies that β indicates a preference for either o or o' depending if it is larger or smaller than 1. For this reason, we choose $\beta = 1$. The transformation from the stable scale prior distribution to the unknown angle prior is simply the square of the location transform,

$$o = \tan^2 \frac{\phi}{2}. \tag{5.6.54}$$

5.6.7 Summary

We start from a fundamental prior choice, which in our case is the uniform distribution for an unknown angle,

$$p(\phi|\mathcal{H}) = \frac{1}{2\pi}. \tag{5.6.55}$$

Then we use a set of transformations to change into different classes of priors,

$$\begin{aligned} \texttt{Location:}\ \mu &= \tan \frac{\phi}{2} \\ \texttt{Probability:}\ \rho &= \frac{1-\cos\phi}{2} \\ \texttt{Scale:}\ o &= \tan^2\left[\frac{\phi}{2}\right], \end{aligned} \tag{5.6.56}$$

which leads to the class of stable prior distributions,

$$\begin{aligned} \texttt{Location:}\ p(\mu|\mathcal{H}) &= \frac{1}{\pi(1+\mu^2)} \\ \texttt{Probability:}\ p(\rho|\mathcal{H}) &= \frac{1}{\pi\sqrt{\rho(1-\rho)}} \\ \texttt{Scale:}\ p(o|\mathcal{H}) &= \frac{1}{\pi\sqrt{o}(1+o)}. \end{aligned} \tag{5.6.57}$$

Each of these distributions are stable under a certain type of convolution and can be decomposed into a mixture,

$$\begin{aligned} \texttt{Location:}\ p(\mu|\mathcal{H}) &= \int_0^\infty \left[\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{\mu^2}{2\sigma^2}\right]\right]\left[\frac{1}{\sqrt{2\pi}\sigma^3}\exp\left[-\frac{1}{2\sigma^2}\right]\right] d\sigma^2 \\ \texttt{Probability:}\ p(\rho|\mathcal{H}) &= \sum_{a,b=1/2}^{\infty}\left[\frac{\Gamma[a+b]}{\Gamma[a]\Gamma[b]}\rho^{a-1}(1-\rho)^{b-1}\right]\left[\frac{\delta(a-1)\delta(b)+\delta(a)\delta(b-1)}{2}\right] \\ \texttt{Scale:}\ p(o|\mathcal{H}) &= \int_0^\infty \left[\frac{e^{-o/\alpha}}{\sqrt{\pi o\alpha}}\right]\left[\frac{e^{-1/\alpha}}{\sqrt{\pi}}\alpha^{-\frac{3}{2}}\right] d\alpha. \end{aligned} \tag{5.6.58}$$

This provides us with a set of reference priors to use in most situations that occur in this dissertation. Operationally for prior distributions with unbounded support we prefer proper prior distributions with undefined moments.

This completes our knowledge base $\mathcal{H}$, which assumes exchangeability combined with stable prior distributions.

Chapter 6

Examples and applications

6.1 Poisson source strengths

To illustrate the use of our set of prior distributions we apply them to some simple problems.

6.1.1 Poisson Source Strength

Our first example is to estimate a set of source strengths $\boldsymbol{s} = \{s_b\}$ from a data count vector $\boldsymbol{n}$, basically a histogram. Starting from a product of unknown scale prior distributions,

$$p(\boldsymbol{s}|\mathcal{H}) = \prod_b \frac{1}{\pi\sqrt{s_b}(1+s_b)}. \tag{6.1.1}$$

we assume that dataset $\boldsymbol{n}$ follows a poissonian likelihood

$$p(\boldsymbol{n}|\boldsymbol{s},\mathcal{H}) = \prod_b e^{-s_b}\frac{s_b^{n_b}}{n_b!}, \tag{6.1.2}$$

so that the joint distribution is

$$p(\boldsymbol{s},\boldsymbol{n}|\mathcal{H}) = \prod_b \frac{e^{-s_b}}{\pi(1+s_b)}\frac{s_b^{n_b-1/2}}{n_b!}. \tag{6.1.3}$$

Integrating out $\boldsymbol{s}$, we find the evidence and the posterior,

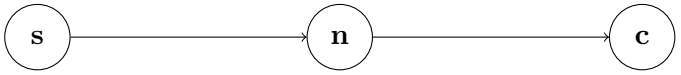
$$\begin{aligned} p(\boldsymbol{n}|\mathcal{H}) &= \prod_b \frac{\Gamma[n_b+1/2]}{\pi n_b!} U\left[\begin{matrix} n_b+1/2 \\ n_b+1/2 \end{matrix}\middle| 1\right] \\ p(\boldsymbol{s}|\boldsymbol{n},\mathcal{H}) &= \frac{p(\boldsymbol{s},\boldsymbol{n}|\mathcal{H})}{p(\boldsymbol{n}|\mathcal{H})} = \prod_b e^{-s_b}\, s_b^{n_b-1/2}\left\{(1+s_b)\,\Gamma[n_b+1/2]\; U\left[\begin{matrix} n_b+1/2 \\ n_b+1/2 \end{matrix}\middle| 1\right]\right\}^{-1} \end{aligned} \tag{6.1.4}$$

where $U[\dots]$ is a Confluent Hypergeometric function; see Appendix A. The mean and variance for $p(\boldsymbol{s}|\boldsymbol{n},\mathcal{H})$ are easily found from the generic posterior moments, of order j

$$\left\langle s_b^j \right\rangle = \frac{\Gamma[n_b + j + 1/2]}{\Gamma[n_b + 1/2]} \frac{U\begin{bmatrix} j + n_b + 1/2 \\ j + n_b + 1/2 \end{bmatrix} 1}{U\begin{bmatrix} n_b + 1/2 \\ n_b + 1/2 \end{bmatrix} 1}. \tag{6.1.5}$$

6.1.2 Poisson Source Strength with detector efficiency

In this example, taken Jaynes (2003), let us add an additional source of error to the data so that it is over-dispersed with greater variance than the Poisson distribution:



The above diagram illustrates the logical chain of our model: A source strength $\boldsymbol{s}$ generates a set of counts $\boldsymbol{n}$ but due to limited efficiency our detector sees only $\boldsymbol{c}$ of these counts. What can we infer about $\boldsymbol{s}$ from $\boldsymbol{c}$? Modelling measured counts with a binomial distribution with $\boldsymbol{\phi}$ the detector efficiency, assumed known in our example,

$$p(\boldsymbol{c}|\boldsymbol{n},\boldsymbol{\phi},\mathcal{H}) = \prod_b \binom{n_b}{c_b} \phi_b^{c_b} (1-\phi_b)^{n_b - c_b} \tag{6.1.6}$$

and combining it with our Poissonian likelihood yields

$$p(\boldsymbol{c}|\boldsymbol{s},\boldsymbol{\phi},\mathcal{H}) = \prod_b \sum_{n_b=c_b}^{\infty} p(\boldsymbol{n}|\boldsymbol{s},\mathcal{H}) p(\boldsymbol{c}|\boldsymbol{n},\boldsymbol{\phi},\mathcal{H}) = \prod_b e^{-s_b\phi_b} \frac{(s_b\phi_b)^{c_b}}{c_b!}. \tag{6.1.7}$$

Adding the prior (6.1.1) gives the corresponding evidence and posterior

$$\begin{aligned} p(\boldsymbol{c}|\boldsymbol{\phi},\mathcal{H}) &= \prod_b \frac{\Gamma[c_b + 1/2]\phi_b^{c_b}}{\pi c_b!} U\begin{bmatrix} c_b + 1/2 \\ c_b + 1/2 \end{bmatrix} \phi_b \\ p(\boldsymbol{s}|\boldsymbol{c},\boldsymbol{\phi},\mathcal{H}) &= \prod_b e^{-\phi_b s_b}\, s_b^{c_b - 1/2} \left\{ (1+s_b)\,\Gamma[c_b + 1/2]\; U\begin{bmatrix} c_b + 1/2 \\ c_b + 1/2 \end{bmatrix} \phi_b \right\}^{-1} \end{aligned} \tag{6.1.8}$$

and posterior moments

$$\left\langle s_b^j \right\rangle = \frac{\Gamma[n_b + j + 1/2]}{\Gamma[n_b + 1/2]} \frac{U\begin{bmatrix} j + n_b + 1/2 \\ j + n_b + 1/2 \end{bmatrix} \phi_b}{U\begin{bmatrix} n_b + 1/2 \\ n_b + 1/2 \end{bmatrix} \phi_b}. \tag{6.1.9}$$

6.2 Bayes Factors in high energy physics

In this section, we apply some of our results to data, taken in this case from a high-energy physics experiment. The *L3 Collaboration* was one of the four major experimental

Q (GeV)	$R_2(Q)$	Error	Q (GeV)	$R_2(Q)$	Error	Q (GeV)	$R_2(Q)$	Error	Q (GeV)	$R_2(Q)$	Error
0.030	1.362	0.051	1.020	0.962	0.005	2.020	0.995	0.008	3.020	0.993	0.011
0.064	1.424	0.021	1.060	0.967	0.005	2.060	0.998	0.008	3.060	0.990	0.011
0.102	1.362	0.013	1.100	0.965	0.005	2.100	0.993	0.008	3.100	1.005	0.011
0.141	1.270	0.009	1.140	0.962	0.005	2.140	0.990	0.008	3.140	0.983	0.011
0.181	1.210	0.007	1.180	0.969	0.005	2.180	0.998	0.008	3.180	0.991	0.011
0.221	1.154	0.006	1.220	0.962	0.005	2.220	0.985	0.008	3.220	0.998	0.011
0.260	1.118	0.005	1.260	0.988	0.006	2.260	1.000	0.009	3.260	1.001	0.011
0.300	1.091	0.005	1.300	0.982	0.006	2.300	0.975	0.008	3.300	0.994	0.011
0.340	1.062	0.005	1.340	0.985	0.006	2.340	0.998	0.009	3.340	0.981	0.011
0.380	1.028	0.004	1.380	0.979	0.006	2.380	0.998	0.009	3.380	0.980	0.011
0.420	1.023	0.004	1.420	0.979	0.006	2.420	0.978	0.009	3.420	0.983	0.012
0.460	1.005	0.004	1.460	0.992	0.006	2.460	0.986	0.009	3.460	0.982	0.012
0.500	0.991	0.004	1.500	0.975	0.006	2.500	0.985	0.009	3.500	0.989	0.012
0.540	0.979	0.004	1.540	0.988	0.007	2.540	1.000	0.009	3.540	0.990	0.012
0.580	0.968	0.004	1.580	0.990	0.007	2.580	0.989	0.009	3.580	0.996	0.012
0.620	0.965	0.004	1.620	0.983	0.007	2.620	0.988	0.009	3.620	1.001	0.012
0.660	0.962	0.004	1.660	0.991	0.007	2.660	0.999	0.010	3.660	0.999	0.012
0.700	0.959	0.004	1.700	0.995	0.007	2.700	0.981	0.010	3.700	0.992	0.012
0.740	0.959	0.004	1.740	0.984	0.007	2.740	0.983	0.010	3.740	0.988	0.012
0.780	0.959	0.004	1.780	0.990	0.007	2.780	1.001	0.010	3.780	0.999	0.013
0.820	0.958	0.004	1.820	0.982	0.007	2.820	1.003	0.010	3.820	0.986	0.013
0.860	0.957	0.005	1.860	0.993	0.007	2.860	1.006	0.010	3.860	0.997	0.013
0.900	0.963	0.005	1.900	0.979	0.007	2.900	0.981	0.010	3.900	0.979	0.013
0.940	0.958	0.005	1.940	0.986	0.008	2.940	1.010	0.011	3.940	0.992	0.013
0.980	0.965	0.005	1.980	0.991	0.008	2.980	1.018	0.011	3.980	1.009	0.010

Table 6.1: Table of data for the normalised Bose-Einstein correlation $R_2(Q)$ as a function of binned four-momentum difference Q for two-jet events, as published in Achard *et al.* (2011) by the L3 Collaboration. Standard errors quoted are statistical only. See text for details.

collaborations at the *Large Electron-Positron* (LEP) Collider at CERN in Geneva. While LEP has long since made way for the Large Hadron Collider, L3 and other collaborations have continued to analyse their data and publish results. For details on CERN, L3 and its many publications see the L3 homepage at `http://l3.web.cern.ch/l3/`

The data shown in Table 6.1 and displayed in Fig. 6.1 was published by L3 in 2011 in tabular form; see Fig. 6 and Table 8 of Achard *et al.* (2011). It was used by us in a preliminary analysis in the conference proceedings of de Kock *et al.* (2011*a*).

The focus in this dissertation falls not on L3 analysis, results or interpretation of the underlying physics; it is merely an exercise and a tool to assess in a realistic environment how Bayesian analysis might work. We therefore outline the underlying physics only very briefly. When electrons and positrons collide at very high energies (as measured in the standard high-energy physics unit of energy of GigaelectronVolts (GeV) also displayed in Table 6.1), they can form a Z-particle which then decays into many lighter particles. Among the most common decay products are pions, which, being bosonic quantum particles, obey quantum statistics and therefore tend to clump together, an effect variously termed *quantum correlations*, *Bose-Einstein correlations* or *femtoscopy*. The x-axis of Fig. 6.1 represents the invariant four-momentum difference $Q = \sqrt{(\boldsymbol{p}_1 - \boldsymbol{p}_2)^2 - (E_1 - E_2)^2}$ between pion 1 and pion 2, while the y-axis quantifies the correlations in terms of $R_2(Q)$ which is a ratio of the number of pion pairs with given Q resulting from a single Z decay to similar calculation made for pion pairs which, being taken from different Z-decays, are uncorrelated. The error bars displayed are the standard way to show how widely these

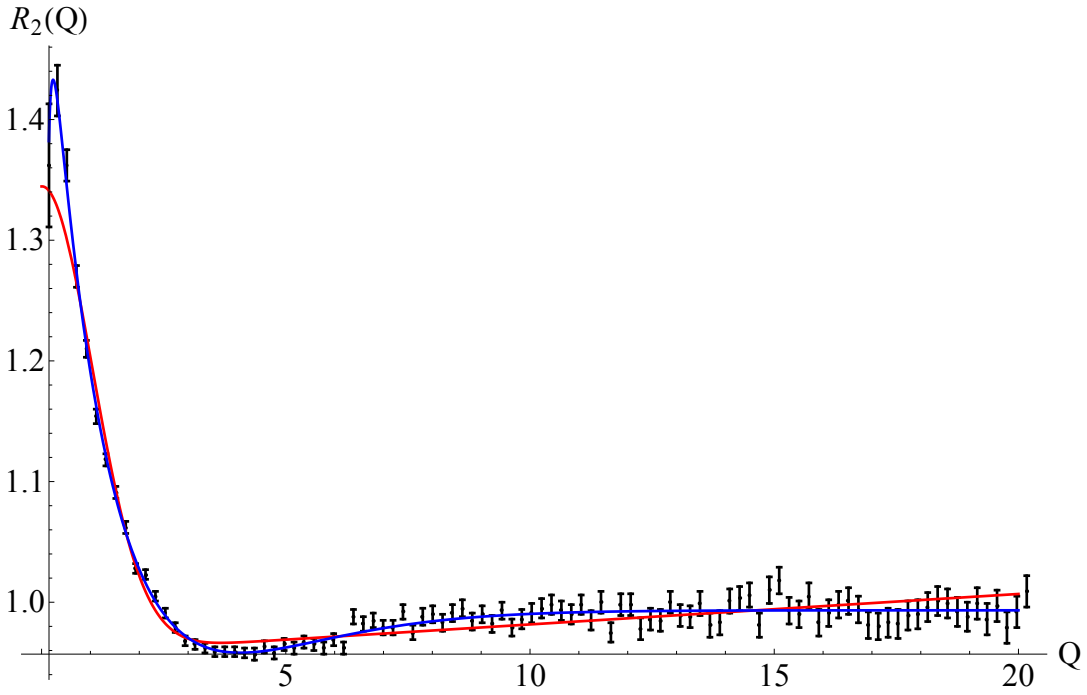


Figure 6.1: Plot of the normalised Bose-Einstein correlation $R_2(Q)$ as a function of binned four-momentum difference Q for two-jet events, as published in Achard *et al.* (2011) by the L3 Collaboration. See text for details. Also plotted in the figure is the worst fit namely the Gaussian Parametrisation $\mathcal{H}_G$ together with the best fit, the second derivative parametrisation $\mathcal{H}_{D3}$.

ratios fluctuate from Z-event to event.

In this section, $\boldsymbol{R}$ is literally the dataset of $R_2(Q)$ values of Table 6.1 and Fig. 6.1 listed as binned values $\{Q_b, R_b(Q_b), \sigma_b(Q_b)\}_{b=1}^{B}$, while the generic parameter notation $\boldsymbol{\theta}$ will translate into various lists of parameters depending on the parametrisation or model $\mathcal{H}$, which could therefore be written in cumbersome notation as $\mathcal{H}_m(\boldsymbol{\theta}_m)$.

Many experimental analyses of such data concern themselves with finding the best parametrisation, i.e. both a choice or selection of theoretical parametrisations containing free parameters $\boldsymbol{\theta}$, as well as finding the particular set of parameter values $\boldsymbol{\theta}^*$ which supposedly best reproduces the measured data. This is often done in terms of minimising a χ^2 criterion, p-values etc. Comparison of different models is then carried out by comparing respective χ^2 divided by the "degrees of freedom", usually taken as the number of data points minus the number of free parameters.

The Bayesian approach has a ready and more rigorous answer for both model comparison and for finding "best" parameter values given a particular model choice. As shown in Section 3.7, Bayesian model comparison would be comparison of posteriors

$$p(\mathcal{H}_m|\boldsymbol{R}) = \frac{\int d\boldsymbol{\theta}_m \, p(\boldsymbol{R}|\boldsymbol{\theta}_m, \mathcal{H}_m) \, p(\boldsymbol{\theta}_m|\mathcal{H}_m) \, p(\mathcal{H}_m)}{\sum_m p(\boldsymbol{R}|\mathcal{H}_m) \, p(\mathcal{H}_m)} \tag{6.2.1}$$

for different models $\mathcal{H}_m$, $\mathcal{H}_{m'}$, ... The denominator requires evaluation of all competing models simultaneously. As already shown in Section 3.7, a shorter but less complete analysis can be done with Bayes Factors, since taking ratios $p(\mathcal{H}_m|\boldsymbol{R})/p(\mathcal{H}_{m'}|\boldsymbol{R})$ cancels

out the denominators $\sum_m p(\boldsymbol{R}|\mathcal{H}_m)\,p(\mathcal{H}_m)$. In the notation of the present section, the Bayes Factor comparing model hypotheses $\mathcal{H}_m$ and $\mathcal{H}_{m'}$ reads

$$\begin{aligned} B_{mm'} &= \log \frac{p(\boldsymbol{R}|\mathcal{H}_m)}{p(\boldsymbol{R}|\mathcal{H}_{m'})} \;=\; \log \frac{\int d\boldsymbol{\theta}_m\; p(\boldsymbol{R}|\boldsymbol{\theta}_m,\mathcal{H}_m)\,p(\boldsymbol{\theta}_m|\mathcal{H}_m)}{\int d\boldsymbol{\theta}_{m'}\; p(\boldsymbol{R}|\boldsymbol{\theta}_{m'},\mathcal{H}_{m'})\,p(\boldsymbol{\theta}_{m'}|\mathcal{H}_{m'})} \\ &= h_{m'} - h_m \end{aligned} \tag{6.2.2}$$

with $h_m = -\log p(\boldsymbol{R}|\mathcal{H}_m)$ the negative log evidence. A Bayes factor larger than 0 favours $\mathcal{H}_m$ over $\mathcal{H}_{m'}$ and vice versa. Put differently, the "better" model will have the smaller h.

The integrals in (6.2.2) may be hard or impossible to calculate analytically, but fortunately accurate answers can be found in many cases with the help of the method of Maximum Likelihood or Laplace's Method as set out in Section 3.4. Below, we present two related methods both based on this. *Method A*, published in de Kock *et al.* (2011*a*), is based on gaussians for both the likelihood and the prior, while *Method B* assigns Cauchy distributions to parameter priors. For the sake of readability, we shorten the notation for the likelihood $p(\boldsymbol{R}\,|\,\boldsymbol{\theta}_m,\mathcal{H}_m)$ to $L(\boldsymbol{\theta}_m)$ or even $L(\boldsymbol{\theta})$.

6.2.1 Method A: Gaussian likelihoods, gaussian priors

Assuming independence of measurements (in this case the pion pair relative momenta), the integrand (likelihood)×(prior) becomes strongly peaked at a particular point $\tilde{\boldsymbol{\theta}}_m$ in the model parameter space, so that it can be expanded locally around the maximum

$$L(\boldsymbol{\theta}_m)\,p(\boldsymbol{\theta}_m\,|\,\mathcal{H}_m) \simeq L(\tilde{\boldsymbol{\theta}}_m)\,p(\tilde{\boldsymbol{\theta}}_m\,|\,\mathcal{H}_m)\exp\left[-\frac{1}{2}(\boldsymbol{\theta}_m-\tilde{\boldsymbol{\theta}}_m)\boldsymbol{A}_m^{-1}(\boldsymbol{\theta}_m-\tilde{\boldsymbol{\theta}}_m)\right] \tag{6.2.3}$$

where $\boldsymbol{A}_m^{-1}$ is the Hessian of the expansion

$$(\boldsymbol{A}_m^{-1})_{ij} = -\frac{\partial^2 \log[L(\boldsymbol{\theta}_m)\,p(\boldsymbol{\theta}_m\,|\,\mathcal{H}_m)]}{\partial\theta_{mi}\,\partial\theta_{mj}}\bigg|_{\boldsymbol{\theta}_m=\tilde{\boldsymbol{\theta}}_m} \tag{6.2.4}$$

and $\boldsymbol{A}_m$ is the parameter covariance matrix. As more data is accumulated, the peak narrows so that we can neglect the fact that parameters may have finite ranges and take the limits to infinity. Integrating (6.2.3) over $\mathbb{R}$, one obtains Laplace's result see Tierney *et al.* (1986) and Bleistein and Handelsman (1986),

$$p(\boldsymbol{R}|\mathcal{H}_m) = \int_{-\infty}^{+\infty} d\boldsymbol{\theta}_m\; L(\boldsymbol{\theta}_m)\,p(\boldsymbol{\theta}_m\,|\,\mathcal{H}_m) \;\simeq\; L(\tilde{\boldsymbol{\theta}}_m)\,p(\tilde{\boldsymbol{\theta}}_m\,|\,\mathcal{H}_m)\sqrt{(2\pi)^{N_m}\det\boldsymbol{A}_m}\,, \tag{6.2.5}$$

with N_m the number of parameters in $\boldsymbol{\theta}_m$. The negative log evidence for model $\mathcal{H}_m$ therefore neatly splits into

$$\begin{aligned} h_m &= -\log p(\boldsymbol{R}|\mathcal{H}_m) \;=\; L+F+P, \\ L &= -\log L(\tilde{\boldsymbol{\theta}}_m) && \text{the maximum likelihood contribution,} \\ F &= -\tfrac{1}{2}\log[(2\pi)^{N_m}\det\boldsymbol{A}_m] && \text{the determinant's contribution,} \\ P &= -\log p(\tilde{\boldsymbol{\theta}}_m|\mathcal{H}_m) && \text{the prior contribution,} \end{aligned} \tag{6.2.6}$$

whereby the difference with the usual maximum likelihood methods is also immediately apparent: the P- and F-contributions are not part of the non-Bayesian Maximum Likelihood repertoire. As we will see below, the contributions of P and F are, in fact, comparatively small for the specific L3 data set at hand, but of course this does not imply it will always be so.

6.2.2 Method B: Gaussian likelihoods, Cauchy priors

Method A as published in de Kock *et al.* (2011*a*) in our view already represents progress towards an approximation that is consistent with the Bayesian approach in that priors are taken seriously while still permitting analytical calculations. However, this requires us to specify these priors, which is far from easy. If we have pertinent information which determines the prior form and numerical values for its metaparameters, that is fine, but it may be hard to translate into mathematical form. The approach hence taken in de Kock *et al.* (2011*a*) was to assign Gaussians to $p(\tilde{\boldsymbol{\theta}}_m|\mathcal{H}_m)$ with width parameters determined from one of the available L3 data points and using the rest for the likelihood. While consistent, this hardly represents a full and final answer to the question of assigning or determining priors.

Method B is an attempt to improve on this situation in two ways, namely in setting priors that are as uninformative as possible, and secondly in doing a linear transform, where possible, to a parameter set with smaller correlations.

We deal first with the linear transform, starting again with the Laplace approximation but for the likelihood only. As in Section 3.4, we maximise only the likelihood

$$\frac{\partial}{\partial\theta_b}\log L(\boldsymbol{\theta}) = 0 \quad \forall\, b, \tag{6.2.7}$$

and label the solution as $\boldsymbol{\theta}^*$. The likelihood is then approximately

$$L(\boldsymbol{\theta}_m) \simeq L(\boldsymbol{\theta}^*_m)\exp\left[-\frac{1}{2}(\boldsymbol{\theta}_m-\boldsymbol{\theta}^*_m)\boldsymbol{H}^{-1}_m(\boldsymbol{\theta}_m-\boldsymbol{\theta}^*_m)\right], \tag{6.2.8}$$

with the Hessian

$$(\boldsymbol{H}^{-1}_m)_{ij} = -\left.\frac{\partial^2\log L(\boldsymbol{\theta}_m)}{\partial\theta_{mi}\,\partial\theta_{mj}}\right|_{\boldsymbol{\theta}_m=\boldsymbol{\theta}^*_m}. \tag{6.2.9}$$

Assuming our Hessian matrix is non-negative definite (part of our well-conditioned assumption), we make the orthogonal transformation,

$$\boldsymbol{\theta}_m - \boldsymbol{\theta}^*_m = \boldsymbol{B}\boldsymbol{\theta}'_m, \tag{6.2.10}$$

so that

$$L[\boldsymbol{\theta}_m] = L[\boldsymbol{\theta}^*_m]\exp\left[-\frac{1}{2}\boldsymbol{\theta}'_m\,\boldsymbol{I}\,\boldsymbol{\theta}'_m\right]|\boldsymbol{B}| \tag{6.2.11}$$

where $\boldsymbol{H}_m^{-1} = \boldsymbol{B}\boldsymbol{B}^T$, $\boldsymbol{I}$ is the identity matrix and $|\boldsymbol{B}|$ is the absolute value of the determinant of the matrix. Using the fact that,

$$|\boldsymbol{B}| = \sqrt{|\boldsymbol{B}\boldsymbol{B}^T|} = \sqrt{\boldsymbol{H}_m^{-1}} = |\boldsymbol{H}_m|^{-1/2}, \tag{6.2.12}$$

we get

$$L[\boldsymbol{\theta}] \simeq \frac{L[\boldsymbol{\theta}^*]}{\sqrt{|\boldsymbol{H}_m|}} \exp\left[-\frac{1}{2}\boldsymbol{\theta}'_m\boldsymbol{\theta}'_m\right]. \tag{6.2.13}$$

Next, we consider again the question of priors. Barring any width-reducing information, the most satisfactory prior would be one which least prejudices the results i.e. is as wide as possible, and so the use of improper priors of the type suggested in Section 5.6 seems the obvious answer. We cannot, however, use improper priors in calculating $p(\boldsymbol{R}|\mathcal{H}_m)$ and $p(\boldsymbol{R}|\mathcal{H}_{m'})$ in (6.2.2) since they represent different models with different parameter sets and therefore different evidence constants which therefore do not cancel.

Use of Laplace's method above has reduced our model to that of a Gaussian with known variance, which is determined directly by the parametrisation, but unknown location. We can therefore view all our parameters as unknown locations which, as shown in Section 5.6, motivates the use of the Cauchy prior[1]

$$p(\theta_b|\beta,\mathcal{H}) = \frac{\beta}{\pi\left(\beta^2+\theta_b^2\right)}, \tag{6.2.14}$$

and since we have standardised the likelihood gaussian in (6.2.11), we should also pick the standardised prior distribution. Hence we choose

$$p(\boldsymbol{\theta}'|\mathcal{H}) = \prod_{j=1}^{N_m} \frac{1}{\pi(1+\theta_j'^{\,2})}. \tag{6.2.15}$$

Putting together all the pieces, the evidence for model $\mathcal{H}_m$ in the Laplace approximation becomes

$$\begin{aligned} p(\boldsymbol{R}|\mathcal{H}_m) &= \int_{-\infty}^{+\infty} d\boldsymbol{\theta}'_m\, L(\boldsymbol{\theta}'_m)\, p(\boldsymbol{\theta}'_m\,|\,\mathcal{H}_m) \\ &\simeq \frac{L(\boldsymbol{\theta}^*_m)}{\sqrt{|H_m|}} \left(\frac{1}{\sqrt{\pi}}\, U \begin{bmatrix} 1/2 & \\ 1/2 & \end{bmatrix}^{1/2}\right)^{N_m}. \end{aligned} \tag{6.2.16}$$

This is a very general result as it is applicable any time the maximum likelihood solution, is well conditioned and is usually presented as the negative logarithm of the evidence,

$$\begin{aligned} h_m = -\log p(\boldsymbol{R}|\mathcal{H}_m) &= -\log L(\boldsymbol{\theta}^*) + \tfrac{1}{2}\log|\boldsymbol{H}_m| + N_m \cdot (0.647874\ldots) \\ &= L + D + C \end{aligned} \tag{6.2.17}$$

with L again the maximum likelihood contribution as before, D the determinant, and C a linear function of the number of parameters N_m.

[1] One could argue in the reverse direction, saying that the motivation for approximating likelihoods with Gaussian distributions is the fact that this eases the burden of choosing prior distributions!

6.2.3 Parametrisations

If raw data in the form of the complete sample of pion pair four-momentum differences would be available, it would be possible to derive relations between this sample and parametrisations. For the moment, we limit ourselves to the typical case where only binned data of the above form is provided which we treat as a histogram $\boldsymbol{n} = \{n_b\}_{b=1}^{B}$ with $\sum_b n_b = n$ over bin midpoints $\boldsymbol{Q} = \{Q_b\}$. The most general "parametrisation" of the histogram contents is then the multinomial

$$p(\boldsymbol{n} \,|\, \boldsymbol{\rho}, n) = n! \prod_{b=1}^{B} \frac{\rho_b^{n_b}}{n_b!}\,, \tag{6.2.18}$$

which on use of the Stirling approximation becomes, up to a normalisation constant,

$$p(\boldsymbol{n} \,|\, \boldsymbol{\rho}, n) = c \cdot \exp\left[-\sum_b n_b \ln \frac{n_b}{n\rho_b}\right]. \tag{6.2.19}$$

Expanding the free parameters $\boldsymbol{\rho}$ around the measured data $\boldsymbol{n}$ and truncating

$$\begin{aligned} p(\boldsymbol{n} \,|\, \boldsymbol{\rho}, n) &= c \cdot \exp\left[-\sum_b \left(\frac{(n\rho_b - n_b)^2}{2n_b} - \frac{(n\rho_b - n_b)^3}{3n_b^2} + \ldots\right)\right] \\ &\simeq c \cdot \exp\left[-\frac{1}{2}\sum_b \frac{(n\rho_b - n_b)^2}{n_b}\right], \end{aligned} \tag{6.2.20}$$

we can identify the multinomial quantities with the measured correlation functions at mid-bin points Q_b by setting[2] $n_b \to I\,R_2(Q_b)$, $C = \sum_b R_2(Q_b)$, and $n \to I\,C$. The n_b in the denominator is almost equal to the measured bin variances $n_b \simeq \sigma^2(n_b) = I^2\,\sigma^2(R_2(Q_b))$ so that the quadratic term is

$$\frac{(n\rho_b - n_b)^2}{2\,n_b} \simeq \frac{[R_2(Q_b) - f(Q_b \,|\, \boldsymbol{\theta}_m)]^2}{2\,\sigma(R_2(Q_b))^2}, \tag{6.2.21}$$

where $n\rho_b/I \to f(Q_b \,|\, \boldsymbol{\theta}_m)$, which includes all the constants, is the unnormalised parametrisation for $R_2(Q)$ in common use, sampled at bin midpoints Q_b, and $\sigma_b = \sigma(R_2(Q_b))$ are the standard errors provided by L3. On normalising, the likelihood is therefore the Gaussian

$$L(\boldsymbol{\theta}_m) = p(\boldsymbol{R}|\boldsymbol{\theta}_m, \boldsymbol{Q}, \boldsymbol{\sigma}, \mathcal{H}_m) = \prod_b \frac{1}{\sqrt{2\pi}\sigma_b}\, \exp\left[\frac{-\,[R_b - f(Q_b|\boldsymbol{\theta})]^2}{2\sigma_b^2}\right]. \tag{6.2.22}$$

This particular form of the likelihood unquestioningly assumes that the statistical errors $\boldsymbol{\sigma} = \{\sigma_b\}$ experimentally specified in Column 3 of the table are given; a more sophisticated analysis would consider these errors too.

6.2.4 Preliminary results for L3 data

The parametrisations $f(Q|\boldsymbol{\theta})$, shown in Table 6.2 that we are going to investigate are rather simple and not intended to exhaust the possibilities but only to illustrate the issues that we raised above. This is the same list as can be found in de Kock *et al.* (2011*a*). Table 6.3 and Table 6.4 list the negative log evidences h_m as calculated with Method A

[2] I is an arbitrary large integer to ensure that $I\,R_2(Q_b)$ is an integer. As it eventually cancels out, its size is immaterial.

Model	Name	N_m	Parametrisation $f(Q\|\boldsymbol{\theta})$
$\mathcal{H}_G$	Gaussian	4	$\gamma(1+\epsilon Q)\left(1+\lambda e^{-(rQ)^2}\right)$
$\mathcal{H}_{SE}$	Stretched Exponential	5	$\gamma(1+\epsilon Q)\left(1+\lambda e^{-(rQ)^\alpha}\right)$
$\mathcal{H}_T$	Simplified τ Model	5	$\gamma(1+\epsilon Q)\left(1+\lambda e^{-(rQ)^{2\alpha}}\cos\left[\tan(\alpha\pi/2)(rQ)^{2\alpha}\right]\right)$
$\mathcal{H}_{L_1}$	1st-order Lévy poly.	5	$\gamma\left(1+\lambda e^{-(rQ)^\alpha}\left[1+c_1L_1(Q\|\alpha,r)\right]\right)$
$\mathcal{H}_{L_3}$	3rd-order Lévy poly.	6	$\gamma\left(1+\lambda e^{-(rQ)^\alpha}\left[1+c_1L_1(Q\|\alpha,r)+c_3L_3(Q\|\alpha,r)\right]\right)$
$\mathcal{H}_{D_1}$	1st-order Derivative	5	$\gamma\left(1+\lambda e^{-(rQ)^\alpha}+c_1\frac{\mathrm{d}}{\mathrm{d}Q}e^{-(rQ)^\alpha}\right)$
$\mathcal{H}_{D_3}$	3rd-order Derivative	6	$\gamma\left(1+\lambda e^{-(rQ)^\alpha}+c_1\frac{\mathrm{d}}{\mathrm{d}Q}e^{-(rQ)^\alpha}+c_3\frac{\mathrm{d}^3}{\mathrm{d}Q^3}e^{-(rQ)^\alpha}\right)$

Table 6.2: List of parametrisations applied to L3 data.

and Method B respectively. The contributions $L = -\log L(\boldsymbol{\theta}^*_m)$ are the same for both methods. The determinant part $D = -\frac{1}{2}\log|\mathcal{H}_m|$ is affected by the prior distribution as we can see by comparing the methods: the greater width of the Cauchy prior increases the contribution of the determinant for all the models. The prior part P in Method A obviously also depends on the prior. Because the results of the two methods are similar, it increases our confidence in the stability of our answer and indicates that the results are not that sensitive to the specific procedure of constructing the priors as long as it treats all models fairly.

The idea, as stated in (6.2.2), is to compare any two models by taking the difference in negative log likelihoods $B_{mm'} = h_{m'} - h_m$, so that $\mathcal{H}_{D_3}$ "wins" in the present contest. From Table 6.4 we can conclude that the odds for $\mathcal{H}_{D_3}$ over $\mathcal{H}_{D_1}$ is $e^2 \approx 7:1$. Unlike the conventional χ^2, which was shown by Jaynes (2003) and others to be very inaccurate for "bad" parametrisations with large χ^2, the numbers can be trusted even for bad parametrisations.

The main conclusion of this exercise is that the log likelihood L dominates the overall h_m, but since all L values are large but similar in size, the determinant contributions may actually determine the final ranking.

Comparing our gauss-Cauchy Method B to similar information criteria in the literature by Akaike (1974) and Schwarz (1978),

$$\begin{aligned} \texttt{Akaike:} \quad & \log p(\boldsymbol{R}|\mathcal{H}) \approx \log L[\boldsymbol{\theta}^*] - B \\ \texttt{Schwarz:} \quad & \log p(\boldsymbol{R}|\mathcal{H}) \approx \log L[\boldsymbol{\theta}^*] - {}^1\!/_2 B \log N \end{aligned} \tag{6.2.23}$$

we see that these can be understood as approximations to the evidence that we have computed. If all the models have roughly the same parameter space size then the determinants would be the same and choosing the highest evidence would be equivalent to using the Akaike Information Criterion. The diagonal entries in the Hessian matrix will all be of order N (the sample size) thus the determinant will be roughly N^B and so the Schwarz criterion will be asymptotically equivalent. For more information about Bayes Factors see Kass and Raftery (1995).

Model	L	F	P	Total h_m
$\mathcal{H}_G$	516.0	22.3	2.5	540.8
$\mathcal{H}_{SE}$	488.7	23.6	-0.3	511.9
$\mathcal{H}_T$	440.0	25.6	-2.4	463.3
$\mathcal{H}_{L_1}$	438.6	21.0	1.5	461.1
$\mathcal{H}_{L_3}$	438.4	28.8	2.6	469.9
$\mathcal{H}_{D_1}$	439.4	20.2	2.9	462.4
$\mathcal{H}_{D_3}$	434.6	22.0	3.4	459.9

Table 6.3: Negative log likelihoods h_m as the sum of contributions L, F and P as defined in (6.2.6) using Method A.

Model	L	D	C	Total h_m
$\mathcal{H}_G$	516.0	26.0	2.6	544.6
$\mathcal{H}_{SE}$	488.7	28.2	3.2	520.1
$\mathcal{H}_T$	440.0	30.3	3.2	473.5
$\mathcal{H}_{L_1}$	438.6	25.6	3.2	467.4
$\mathcal{H}_{L_3}$	438.4	34.4	3.9	476.7
$\mathcal{H}_{D_1}$	440.3	24.5	3.2	468
$\mathcal{H}_{D_3}$	434.6	27.5	3.9	466

Table 6.4: Negative log likelihoods h_m as the sum of contributions L, D and C as defined in (6.2.17) using Method B.

6.3 Correlations

Using the Laplace-De Finetti theorem has some consequences, as we now illustrate with a simple example from Jaynes (1982). Consider the binary case again (4.12.22) and ask what two trial data ($R = 2$) can be represented by the LdF form? Setting,

$$A = \int_0^1 \rho^2 f(\rho) d\rho \qquad\qquad C = \int_0^1 (1-\rho)^2 f(\rho) d\rho. \tag{6.3.1}$$

we obtain after some algebra

$$\begin{aligned}(A-C)^2 &= \int_0^1\int_0^1 \left[\rho_1^2 - (1-\rho_1)^2\right]\left[\rho_2^2 - (1-\rho_2)^2\right] f(\rho_1)f(\rho_2)d\rho_1 d\rho_2 \\ &= \int_0^1\int_0^1 (4\rho_1\rho_2 - 2\rho_1 - 2\rho_2 + 1)\, f(\rho_1)f(\rho_2)d\rho_1 d\rho_2\end{aligned} \tag{6.3.2}$$

$$\begin{aligned}2(A+C) &= \int_0^1\int_0^1 \left[\rho_1^2 + (1-\rho_1)^2 + \rho_2^2 + (1-\rho_2)^2\right] f(\rho_1)f(\rho_2)d\rho_1 d\rho_2 \\ &= \int_0^1\int_0^1 \left[2\rho_1^2 + 2\rho_2^2 - 2\rho_1 - 2\rho_2 + 2\right] f(\rho_1)f(\rho_2)d\rho_1 d\rho_2,\end{aligned} \tag{6.3.3}$$

to find

$$2(A+C)-(A-C)^2-1=2\int_0^1\int_0^1(\rho_1-\rho_2)^2f(\rho_1)f(\rho_2)d\rho_1 d\rho_2 \qquad \geq 0, \tag{6.3.4}$$

because $f(\rho)\geq 0$. Hence we can only generate data sets for which (6.3.4) is non-negative.

In Fig. 6.2 any infinite exchangeable data set must fall into the region between the parabola and the straight line. Intuitively the right-hand side of (6.3.4) is the correlation between the outcomes of the first x_1 and second trial x_2; thus we can only represent non-negative correlated data sets with the LdF representation. Negative correlation between trials is thus a sign of finite data sequences.

It is worth making this distinction more vivid: Associate g_b with the outcome b and use the labels g_{br} for the outcome on trial r as previously. Consider the sum,

$$\left[\sum_{r=1}^{R}(g_{br}-\mu)\right]^2=\sum_{r=1}^{R}(g_{br}-\mu)^2+\sum_{s,r=1}^{R}(g_{bs}-\mu)(g_{br}-\mu), \tag{6.3.5}$$

where μ is the average contribution of a single trial. Summing over all possible g_{br} we find the variance of G, in terms of the variance of the single trials σ and the covariance between single trails ϕ to be

$$\mathrm{var}(G)=R\sigma^2+R(R-1)\phi. \tag{6.3.6}$$

Since the variance is positive we must have,

$$\phi\geq-\frac{\sigma^2}{R-1}. \tag{6.3.7}$$

A correlation coefficient can then be defined as the ratio $\frac{\phi}{\sigma}$ which in general can vary between minus one and one, but for infinite exchangeable sequences as this result shows some of the negative values are excluded. This result can also be interpreted geometrically, see Diaconis (1977).

6.3.1 Estimation of a correlation parameter

The final example will serve as additional motivation for the Jeffreys-Perks prior. Consider a data sample given in terms of occupation numbers $\{n_1,n_2,n_3,n_4\}$ in a 2x2 table as in the Kangaroo Example, where $\mathcal{B}$ and $\mathcal{L}$ are some arbitrary properties,

	$\mathcal{L}$	$\overline{\mathcal{L}}$
$\mathcal{B}$	n_1	n_2
$\overline{\mathcal{B}}$	n_3	n_4,

which we translate into a Dirichlet posterior with the use of our Jeffrey-Perks prior,

$$p(\boldsymbol{\rho}|\boldsymbol{n},\mathcal{H})=(N+1)!\frac{\rho_1^{n_1-1/2}}{(n_1-1/2)!}\frac{\rho_2^{n_2-1/2}}{(n_2-1/2)!}\frac{\rho_3^{n_3-1/2}}{(n_3-1/2)!}\frac{(1-\rho_1-\rho_2-\rho_3)^{n_4-1/2}}{(n_4-1/2)!}. \tag{6.3.8}$$

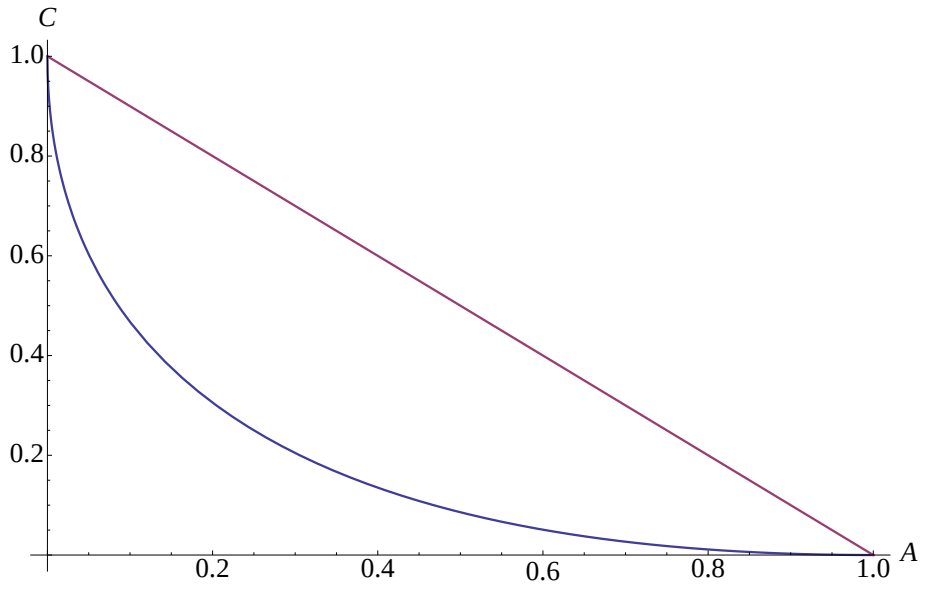


Figure 6.2: Plot of sequences allowed by the Laplace-de Finetti representation.

so that $\{\rho_1, \rho_2, \rho_3, \rho_4\}$ are chances associated with each occupation number. Defining a correlation parameter as,

$$\phi = \frac{\rho_1 + \rho_4 - \rho_2 - \rho_3}{2}, \tag{6.3.9}$$

we can easily estimate it from the posterior,

$$\langle \phi \rangle = \int \frac{\rho_1 + \rho_4 - \rho_2 - \rho_3}{2} p(\boldsymbol{\rho}|\boldsymbol{n}, \mathcal{H}) d\boldsymbol{\rho} = \frac{n_1 + n_4 - n_2 - n_3}{2(2+N)}, \tag{6.3.10}$$

but it is more informative to look at the posterior of ϕ than just its mean estimate. Transforming into marginals and correlation parameters,

$$\begin{aligned} \rho_1 &= \rho_{\mathcal{B}}\rho_{\mathcal{L}} + \phi \quad & \rho_2 &= \rho_{\mathcal{B}}\rho_{\overline{\mathcal{L}}} - \phi \\ \rho_3 &= \rho_{\overline{\mathcal{B}}}\rho_{\mathcal{L}} - \phi \quad & \rho_4 &= \rho_{\overline{\mathcal{B}}}\rho_{\overline{\mathcal{L}}} + \phi. \end{aligned} \tag{6.3.11}$$

In calculating the posterior we will focus on general methods and use a more scenic route to calculating in the hope that it would make the answer more digestible. First compute the general generating function where $\rho_1 + \rho_2 + \rho_3 + \rho_4 = \rho$ with the attention of setting ρ equal to one later,

$$p(\boldsymbol{\rho}|\boldsymbol{n}, \mathcal{H}) = (N+1)! \frac{\rho_1^{n_1 - 1/2}}{(n_1 - 1/2)!} \frac{\rho_2^{n_2 - 1/2}}{(n_2 - 1/2)!} \frac{\rho_3^{n_3 - 1/2}}{(n_3 - 1/2)!} \frac{\rho_4^{n_4 - 1/2}}{(n_4 - 1/2)!}. \tag{6.3.12}$$

We take the generating function in ρ to μ and in ϕ to ic. In general for $(0, \infty)$ support we use the Laplace transform and for the support $(-\infty, \infty)$ we use the Fourier transform. Hence, for the probabilities we use the Laplace transform and the Fourier transform for

the correlation parameter.

$$\begin{aligned}\Phi\left[\mu, ic\right] &= \int_0^\infty p(\boldsymbol{\rho}|\mathcal{H}) e^{-\mu(\rho_1+\rho_2+\rho_3+\rho_4)+ic(\rho_1+\rho_4-\rho_2-\rho_3)} d\boldsymbol{\rho} \\ &= (N+1)! \left(\mu - \frac{ic}{2}\right)^{-n_1-n_4-1} \left(\mu + \frac{ic}{2}\right)^{-n_2-n_3-1}\end{aligned} \tag{6.3.13}$$

From the integral we see that each of the factors is the generating function of a gamma distribution and the product is thus the convolution of two gamma distributions,

$$\begin{aligned}&\int_0^\rho \left[e^{\frac{ic}{2}s} \frac{s^{n_2+n_3-1}}{(n_2+n_3-1)!}\right] \left[e^{\frac{ic}{2}(\rho-s)} \frac{(s-\rho)^{n_1+n_4-1}}{(n_1+n_4-1)!}\right] ds \\ &= e^{\frac{ic}{2}\rho} \frac{\rho^{N+1}}{(N+1)!} {}_1F_1\left[\begin{matrix} n_2+n_3+1 \\ N+2 \end{matrix}\middle| ic\rho\right]\end{aligned} \tag{6.3.14}$$

and

$$\int_0^\infty e^{\frac{ic}{2}\rho} \frac{\rho^{N+1}}{(N+1)!} {}_1F_1\left[\begin{matrix} n_2+n_3+1 \\ N+2 \end{matrix}\middle| ic\rho\right] e^{-\mu\rho} d\rho = \left(\mu - \frac{ic}{2}\right)^{-n_1-n_4-1} \left(\mu + \frac{ic}{2}\right)^{-n_2-n_3-1}. \tag{6.3.15}$$

Setting $\rho = 1$ we have,

$$\Phi\left[ic\right] = e^{\frac{ic}{2}} {}_1F_1\left[\begin{matrix} n_2+n_3+1 \\ N+2 \end{matrix}\middle| ic\right], \tag{6.3.16}$$

which can easily be shown to be the Fourier transform of a Beta distribution,

$$p(\phi|\boldsymbol{n}, \mathcal{H}) = \frac{(N+1)!(\frac{1}{2}-\phi)^{n_2+n_3}(\frac{1}{2}+\phi)^{n_1+n_4}}{(n_2+n_3)!(n_1+n_4)!}. \tag{6.3.17}$$

We promised to give some additional arguments for the Jeffreys prior and this is it. Examine the no data case, $n_1 = n_2 = n_3 = n_4 = 0$: the posterior gives all possible correlations as equally likely. Contrast this with the Bayes-Laplace prior which states that all data sequences are equally likely but if you examine the posterior for the correlation parameter you find,

$$p(\phi|\mathcal{H}) = 3!\left(\tfrac{1}{2}-\phi\right)\left(\tfrac{1}{2}+\phi\right). \tag{6.3.18}$$

Thus the Jeffreys-Perks takes all correlations as equally likely which we prefer over all data sequences equally likely.

Chapter 7

Assigning probabilities for exchangeable sequences

This chapter will be a repeat of Chapter 4, trying to answer the question of how to assign probabilities. Assuming *exchangeability* will basically change our evidence assignment from formulae of **Maxwell-Boltzmann** to those resembling the corresponding **Bose-Einstein** and **Fermi-Dirac** statistics. In the Bose-Einstein case this is equivalent to using Laplace's rule of succession to assign probabilities and Von Mises's idea of relative frequency. While we will again follow the methodology of Khinchin (1960) and use the Darwin-Fowler method, a subtlety arises in that we are forced to switch to the grand canonical ensemble by introducing a prior for R to keep the mathematics sane. While the mathematical answer is well-known, the route taken in this chapter seems to be a new idea to base entropy on exchangeability; see Jaynes (1986).

This chapter is organised as follows:

- In Section 7.1 we derive the rule of succession which is a generalisation of the Principle of Insufficient Reason.
- In Section 7.2 we introduce the prior distribution for R, the total number of trials.
- In Section 4.7 we started from a set of logical independent prior probabilities $\boldsymbol{\rho}$ which we then updated to a set of posterior probabilities $\boldsymbol{\rho}'$ after imposing a constraint $\delta(G - \sum g_b r_B)$. In Section 7.3 below, we will do the same but within the wider class of exchangeability.
- Section 7.4 contains the calculation of the exact predictions for exchangeable trials in terms of the structure function.
- Section 7.5 applies the saddlepoint approximation to our structure function for exchangeable trials.
- In Section 7.6 we reconsider the Principle of Relative Entropy and thereby find a new distance measure.

- In Section 7.7 we solve a few simple examples.
- Section 7.8 introduces the opposite case of finite populations.
- In Section 7.9 we introduce the corresponding prior distributions for the number of trials.
- In Section 7.10 we define the relevant structure and partition functions.
- Section 7.11 is the calculation of the predictions for finite trials in terms of the structure function.
- Section 7.12 applies the saddlepoint approximation to our structure function for finite trials.
- In Section 7.13 we reconsider the Principle of Relative Entropy which leads us to a second entropy measure for finite populations.
- In Section 7.14 we again solve some simple example problems.

For orientation, we once again start with the key set of Bayesian equations. Essentially repeating Eqs. (5.6.1)–(5.6.3) but having already taken the limit $m_b/M \to \rho_b$, the evidence $p(\boldsymbol{n}|\boldsymbol{a},\mathcal{H})$ for data $\boldsymbol{n}$, posterior $p(\boldsymbol{\rho}|\boldsymbol{n},\boldsymbol{a},\mathcal{H})$ for Bernoulli probabilities $\boldsymbol{\rho}$ and prediction $p(\boldsymbol{r}|\boldsymbol{n},\mathcal{H})$ for future counts $\boldsymbol{r}$, given the data, are

$$p(\boldsymbol{n}|\boldsymbol{a},\mathcal{H}) = \int_{\mathcal{A}(\boldsymbol{\rho})} d\boldsymbol{\rho}\; p(\boldsymbol{n}|\boldsymbol{\rho},\boldsymbol{a},\mathcal{H})\, p(\boldsymbol{\rho}|\boldsymbol{a},\mathcal{H}), \tag{7.0.1}$$

$$p(\boldsymbol{\rho}|\boldsymbol{n},\boldsymbol{a},\mathcal{H}) = \frac{1}{p(\boldsymbol{n}|\boldsymbol{a},\mathcal{H})} p(\boldsymbol{n}|\boldsymbol{\rho},\boldsymbol{a},\mathcal{H})\, p(\boldsymbol{\rho}|\boldsymbol{a},\mathcal{H}), \tag{7.0.2}$$

$$p(\boldsymbol{r}|\boldsymbol{n},\boldsymbol{a},\mathcal{H}) = \int_{\mathcal{A}(\boldsymbol{\rho})} d\boldsymbol{\rho}\, p(\boldsymbol{r}|\boldsymbol{n},\boldsymbol{\rho},\mathcal{H})\, p(\boldsymbol{\rho}|\boldsymbol{n},\boldsymbol{a},\mathcal{H}), \tag{7.0.3}$$

where $\mathcal{A}(\boldsymbol{\rho})$ is the B-dimensional simple $0 \leq \rho_b \leq 1, \sum_b \rho_b = 1$.

7.1 Laplace's rule of succession for exchangeable sequences

A prediction of future counts based on past counts is called *Laplace's rule of succession*. We have shown that for exchangeable sequences two different approaches lead to a negative hypergeometric evidence,

$$\text{Neg. Hypergeometric}[\boldsymbol{n}|\boldsymbol{a}] = \text{Hypergeometric}[\boldsymbol{n}|\boldsymbol{m}] \wedge \text{Neg. Hypergeometric}[\boldsymbol{m}|\boldsymbol{a}] \tag{7.1.1}$$

or

$$\text{Neg. Hypergeometric}[\boldsymbol{n}|\boldsymbol{a}] = \text{Multinomial}[\boldsymbol{n}|\boldsymbol{\rho}] \wedge \text{Dirichlet}[\boldsymbol{\rho}|\boldsymbol{a}], \tag{7.1.2}$$

where we write $\wedge$ for a mixture i.e. we sum or integrate out the shared variable. As the two viewpoints are completely equivalent because they assign the same evidence to all data sequences, we will adopt the second approach, based on the Laplace-de-Finetti theorem with a Dirichlet prior

$$p(\boldsymbol{\rho}|\boldsymbol{a},\mathcal{H}) = \Gamma\big[\textstyle\sum_b a_b\big] \displaystyle\prod_b \frac{\rho_b^{a_b-1}}{\Gamma[a_b]}, \tag{7.1.3}$$

which together with a multinomial likelihood was shown in (5.5.18) also to result in the stated negative hypergeometric evidence, which in the present notation reads

$$p(\boldsymbol{n}|\boldsymbol{a},\mathcal{H}) = \int d\boldsymbol{\rho} \left[N! \prod_b \frac{\rho_b^{n_b}}{n_b!}\right] \left[\Gamma[A] \prod_b \frac{\rho_b^{a_b-1}}{\Gamma[a_b]}\right] = \binom{-A}{N}^{-1} \prod_b \binom{-a_b}{n_b}. \tag{7.1.4}$$

As argued in Section 5.6, there are essentially three defensible choices for the metaparameter prior $p(\boldsymbol{a}|\mathcal{H})$, all three of which are special cases of the Dirichlet prior, namely the uniform Bayes-Laplace prior with $a_b = 1\ \forall\, b$, the Jaynes-Haldane prior (5.6.12) with $a_b = 0$ and the Jeffreys-Perks prior (5.6.21) with $a_b = {}^1\!/_2$. In our view, the Bayes-Laplace prior is suboptimal since it is "too informative", while the Jaynes-Haldane prior is improper and therefore unusable in those cases where the evidence needs to be calculated. We hence make the fixed choice of a Jeffreys-Perks prior $\boldsymbol{a} = \{{}^1\!/_2, \ldots, {}^1\!/_2\}$,

$$p(\boldsymbol{\rho}|\mathcal{H}) = \Gamma\left[\frac{B}{2}\right] \prod_b \frac{\rho_b^{-1/2}}{\Gamma[1/2]} = \frac{[(B/2)-1]!}{\pi^{B/2}} \prod_b \rho_b^{-1/2} \tag{7.1.5}$$

throughout this chapter, and will correspondingly stop explicitly specifying $\boldsymbol{a}$ in the notation. The resulting posterior is

$$p(\boldsymbol{\rho}|\boldsymbol{n},\mathcal{H}) = (N + B/2 - 1)! \prod_b \frac{\rho_b^{n_b-1/2}}{(n_b - {}^1\!/_2)!}. \tag{7.1.6}$$

The probability assigned to a future set of counts $\boldsymbol{r}$

$$p(\boldsymbol{r}|\boldsymbol{\rho},\mathcal{H}) = R! \prod_b \frac{\rho_b^{r_b}}{r_b!} \tag{7.1.7}$$

combined with our posterior gives us, as we already know, a negative hypergeometric distribution for the prediction $\boldsymbol{r}$ of R future trials

$$p(\boldsymbol{r}|\boldsymbol{n},R,\mathcal{H}) = \binom{N + B/2 + R - 1}{R}^{-1} \prod_b \binom{r_b + n_b - {}^1\!/_2}{r_b}. \tag{7.1.8}$$

When predicting a single future trial R=1, only one of the B occupation numbers r_b can be nonzero. The single-trial prediction for this count to be in bin b is

$$p(r_b = 1|R = 1,\boldsymbol{n},\mathcal{H}) = \frac{n_b + {}^1\!/_2}{N + {}^1\!/_2 B}, \tag{7.1.9}$$

which is the **generalised Laplace's rule of succession**, see also Jaynes (2003). It is equivalent to Von Mises's idea of assigning predictive probabilities equal to their relative frequency. Compare this to the original rule of succession of Laplace (1812)

$$p(r_b = 1|R = 1, \boldsymbol{n}) = \frac{n_b + 1}{N + 2}, \tag{7.1.10}$$

who derived it using the uniform Bayes-Laplace prior. (For a history of the rule of succession see Zabell (1989).) To clarify the generalised rule of succession we investigate its large data limit $N \gg B, R$. Multiplying and dividing (7.1.8) by $(N + \frac{1}{2}B)^R = \prod_b (N + \frac{1}{2}B)^{r_b}$, we define

$$p(\boldsymbol{r}|\boldsymbol{n}, R, \mathcal{H}) = C \prod_b F_b \tag{7.1.11}$$

and setting $\rho_b = (n_b + \frac{1}{2})/(N + \frac{1}{2}B)$, the factors converge with the help of (2.1.18) to

$$\begin{aligned} F_b &= \lim_{N\gg B,R} \frac{1}{(N + \frac{1}{2}B)^{r_b}} \binom{r_b + n_b - \frac{1}{2}}{r_b} \\ &= \lim_{N\gg B} \frac{1}{r_b!} \left(\frac{n_b + \frac{1}{2}}{N + \frac{1}{2}B}\right) \left(\frac{n_b + \frac{1}{2} + 1}{N + \frac{1}{2}B}\right) \cdots \left(\frac{n_b + r_b - \frac{1}{2}}{N + \frac{1}{2}B}\right) = \frac{\rho_b^{r_b}}{r_b!}, \\ C &= \lim_{N\gg B} (N + \tfrac{1}{2}B)^R \binom{N + R + \frac{1}{2}B - 1}{R}^{-1} = R! \end{aligned} \tag{7.1.12}$$

so that we recover the logically independent answer

$$\lim_{N\gg B,R} p(\boldsymbol{r}|\boldsymbol{n}, R, \mathcal{H}) = R! \prod_b^B \frac{\rho_b^{r_b}}{r_b!}. \tag{7.1.13}$$

ie we have seen many counts we can assign exact probabilities to a specific outcome and we are back to our logical independence assumption. The other limit is the large prediction limit, where $R \gg N$ while keeping $\rho'_b = \frac{r_b}{R}$ fixed. Proceeding in the same way, we multiply and divide by $R^{N-\frac{1}{2}B} = \prod_b R^{n_b - \frac{1}{2}r_b}$,

$$\frac{1}{R^{n_b - \frac{1}{2}}} \binom{n_b + r_b - \frac{1}{2}}{r_b} = \frac{1}{(n_b - \frac{1}{2})!} \left(\frac{r_b + 1}{R}\right) \left(\frac{r_b + 2}{R}\right) \cdots \left(\frac{n_b + r_b + \frac{1}{2}}{R}\right), \tag{7.1.14}$$

so that

$$\begin{aligned} \frac{1}{R^{n_b - \frac{1}{2}}} \binom{n_b + r_b - \frac{1}{2}}{r_b} &\to \frac{{\rho'_b}^{n_b - \frac{1}{2}}}{(n_b - \frac{1}{2})!} \\ R^{N - \frac{1}{2}B} \binom{N + R + \frac{1}{2}B - 1}{R}^{-1} &\to \frac{(N + \frac{1}{2}B - 1)!}{R} \end{aligned} \tag{7.1.15}$$

or

$$\lim_{R\gg N,B} p(\boldsymbol{r}|R, \boldsymbol{n}, \mathcal{H}) = (N + \tfrac{1}{2}B - 1)! \prod_b \frac{{\rho'_b}^{n_b - \frac{1}{2}}}{(n_b - \frac{1}{2})!}, \tag{7.1.16}$$

a Dirichlet distribution. This shows again that the prior distribution is equal to the large prediction limit and the large data limit gives the likelihood function.

7.2 Outcome space and prediction for variable R

Using Laplace's rule of succession to change constraints into models requires us to change our procedure slightly. The method followed below is based on Khinchin (1960), but we must point out that our formulas are different and our interpretation is very different.

The first change is to consider the number of predicted trials R as a variable, possibly with an additional constraint like in the Maximum Entropy formalism. This change is basically forced upon us because there does not seem to be a corresponding multinomial theorem for the negative hypergeometric distribution. As we have seen in physics terms, this implies switching to a Grand Canonical Ensemble. The prior we shall use is a *Generalised Negative Binomial* for R which depends not on the detailed occupation numbers $\boldsymbol{n}$ but only on their sum $N = \sum_b n_b$,

$$p(R|\boldsymbol{n},\theta,\mathcal{H}) \;=\; p(R|N,\theta,\mathcal{H}) = \binom{N+R+{}^1\!/\!_2 B-1}{R}\theta^{N+{}^1\!/\!_2 B}(1-\theta)^R. \tag{7.2.1}$$

It is a negative binomial rather than a negative multinomial since there is only one meta-parameter θ. Combining with the rule of succession we find

$$p(\boldsymbol{r}|\boldsymbol{n},\theta,\mathcal{H}) = \prod_b^B \binom{n_b+r_b-{}^1\!/\!_2}{r_b}\theta^{n_b+{}^1\!/\!_2}(1-\theta)^{r_b} \tag{7.2.2}$$

$$= p(\boldsymbol{r}|R,\boldsymbol{n},\theta,\mathcal{H})p(R|\boldsymbol{n},\theta,\mathcal{H}). \tag{7.2.3}$$

The motivation for choosing a generalised negative binomial prior is that if we use Fisher's artifice on a set of negative binomial distributions,

$$p(R,\boldsymbol{r}|\boldsymbol{n},\theta,\mathcal{H}) = \; p(R|\boldsymbol{r},\mathcal{H})\,p(\boldsymbol{r}|\boldsymbol{n},\theta,\mathcal{H}) = \delta\big(R-\textstyle\sum_b r_b\big)\displaystyle\prod_b^B \binom{n_b+r_b-{}^1\!/\!_2}{r_b}\theta^{n_b+{}^1\!/\!_2}(1-\theta)^{r_b}, \tag{7.2.4}$$

then the total R would be distributed like our prior again,

$$p(R|\boldsymbol{n},\theta,\mathcal{H}) = \sum_{U(\boldsymbol{r})} p(R,\boldsymbol{r}|\boldsymbol{n},\theta,\mathcal{H}) = \binom{N+R+{}^1\!/\!_2 B-1}{R}\theta^{N+{}^1\!/\!_2 B}(1-\theta)^R, \tag{7.2.5}$$

and if we condition on it, we recover our negative hypergeometric distribution

$$p(\boldsymbol{r}|R,\boldsymbol{n},\theta,\mathcal{H}) = \frac{p(R,\boldsymbol{r}|\boldsymbol{n},\theta,\mathcal{H})}{p(R|\boldsymbol{n},\theta,\mathcal{H})} = \binom{N+R+{}^1\!/\!_2 B-1}{R}^{-1}\prod_b^B \binom{n_b+r_b-{}^1\!/\!_2}{r_b}. \tag{7.2.6}$$

The parameter θ is connected to the importance we assign to the past and the future: for example, the expected number of counts of our prior is,

$$\langle R\rangle = \sum_R R\,p(R|N,\theta,\mathcal{H}) = (N+{}^1\!/\!_2 B)\left(\frac{1-\theta}{\theta}\right), \tag{7.2.7}$$

so here we will make a definite choice to weight our predictions and prior equally by choosing $\theta = {}^1\!/_2$, so that (7.2.1) and (7.2.2) become

$$p(R|\boldsymbol{n}, {}^1\!/_2, \mathcal{H}) = p(R|N, {}^1\!/_2, \mathcal{H}) = \binom{N + R + {}^1\!/_2 B - 1}{R} 2^{-N-R-{}^1\!/_2 B} \tag{7.2.8}$$

$$p(\boldsymbol{r}|\boldsymbol{n}, {}^1\!/_2, \mathcal{H}) = \prod_b^B \binom{n_b + r_b - {}^1\!/_2}{r_b} 2^{-n_b - r_b - {}^1\!/_2}. \tag{7.2.9}$$

We can also connect this construction with the Poisson distribution we derived in Section 4.13, by simply adding a prior distribution. The logically independent assignment can be written as (4.13.15)

$$p(\boldsymbol{r}|\boldsymbol{s}, \mathcal{H}) = \prod_b e^{-s_b} \frac{s_b^{r_b}}{r_b!}, \tag{7.2.10}$$

and the uncertainty in the source strength can be modelled with a Gamma prior on each bin,

$$p(\boldsymbol{s}|\boldsymbol{n}, \mathcal{H}) = \prod_b e^{-s_b} \frac{s_b^{n_b - {}^1\!/_2}}{(n_b - {}^1\!/_2)!} \tag{7.2.11}$$

So that marginalising out the $\boldsymbol{s}$, we recover (7.2.9),

$$p(\boldsymbol{r}|\boldsymbol{n}, \mathcal{H}) = \prod_b \binom{n_b + r_b - {}^1\!/_2}{r_b} 2^{-n_b - r_b - {}^1\!/_2}. \tag{7.2.12}$$

The fact that a negative binomial can be decomposed into a Poisson-Gamma mixture was first used by Greenwood and Yule (1920). The formulation also implies that systems in general look like a collection of Planck Oscillators (4.12.16).

7.3 Constraints for exchangeable trials

This section essentially repeats the derivations of Section 4.7, comparing and contrasting what happens when the premise of logical independence is replaced by exchangeability. Adding constraints for G and for R to our product of oscillators (7.2.9), the joint probability is

$$\begin{aligned} p(R, G, \boldsymbol{r}|\boldsymbol{n}, \mathcal{H}) &= p(R|\boldsymbol{r}, \mathcal{H}) p(G|\boldsymbol{r}, \mathcal{H}) p(\boldsymbol{r}|\boldsymbol{n}, \mathcal{H}) \\ &= \delta\left(G - \textstyle\sum_b g_b r_b\right) \delta\left(R - \textstyle\sum_b r_b\right) \prod_b \binom{n_b + r_b - {}^1\!/_2}{r_b} 2^{-n_b - r_b - {}^1\!/_2}, \end{aligned} \tag{7.3.1}$$

where again we suppress $\boldsymbol{g}$ in the notation, considering it to be part of the background knowledge base $\mathcal{H}$. In order to marginalise over $\boldsymbol{r}$ to find the evidence

$$p(R, G|\mathcal{H}) = \sum_{U[\boldsymbol{r}]} p(R, G, \boldsymbol{r}|\mathcal{H}), \tag{7.3.2}$$

it is is easier to compute the joint moment generating function ($G \to \lambda$, $R \to \mu$) first as we did in (4.7.10), remembering $R = \sum_b r_b$ and $G = \sum_b g_b r_b$

$$\begin{aligned}\Phi[\mu, \lambda \,|\, p(R, G, \boldsymbol{r}|\boldsymbol{n}, \mathcal{H})] &= \sum_{U(\boldsymbol{r})} p(R, G, \boldsymbol{r}|\boldsymbol{n}, \mathcal{H}) e^{-\mu \sum_b r_b - \lambda \sum_b g_b r_b} \\ &= \prod_b \left(\frac{1/2}{1 - 1/2 e^{-\mu - \lambda g_b}} \right)^{n_b + 1/2}.\end{aligned} \tag{7.3.3}$$

Following the nomenclature of Khinchin (1960) we define a "volume"

$$V = N + 1/2 B \tag{7.3.4}$$

making explicit the large parameter we shall use in the saddlepoint approximation. The limit $V \to \infty$, often called the "bulk limit", is in our terminology just the large data limit $N \gg R$ which we for the purposes of constructing forward probabilities. From the generalised rule of succession (7.1.9), we shall take

$$\rho_b = \frac{n_b + 1/2}{N + 1/2 B} = \frac{n_b + 1/2}{V}. \tag{7.3.5}$$

As in (4.7.12), the structure function $\Omega[R, G] = p(R, G|\boldsymbol{n}, \mathcal{H})$ for exchangeable sequences is the inverse of the generating function,

$$\sum_{G,R} \Omega\left[R, G\right] e^{-\mu R - \lambda G} = \Phi[\mu, \lambda \,|\, p(R, G, \boldsymbol{r}|\boldsymbol{n}, \mathcal{H})] \tag{7.3.6}$$

and defining the partition function as

$$Z\left[\mu, \lambda\right] = \prod_b \left(\frac{1/2}{1 - 1/2 e^{-\mu - \lambda g_b}} \right)^{\rho_b}, \tag{7.3.7}$$

we have, again in analogy to (4.7.12),

$$\sum_{G,R} \Omega\left[R, G\right] e^{-\mu - \lambda G} = Z[\mu, \lambda]^V. \tag{7.3.8}$$

From (7.1.13) it is clear that the updated probabilities would have the same definition as before

$$\rho'_b = \frac{\langle r_b | G \rangle}{V}, \tag{7.3.9}$$

while G/V is again taken to be a constant ratio

$$\frac{G}{V} = \gamma. \tag{7.3.10}$$

7.4 Predictions using the exchangeable structure function

Continuing the generalisation, now in parallel with Section 4.8, the definition (4.8.1) of $\langle r_c|G\rangle$ becomes

$$\begin{aligned}\langle r_c|G\rangle &= \sum_{\boldsymbol{r}} r_c \frac{p(\boldsymbol{r},R,G|\mathcal{H})}{p(R,G|\mathcal{H})} \\ &= \frac{1}{\Omega[R,G]}\sum_{\boldsymbol{r}} r_c \prod_b^B \binom{n_b+r_b-\frac{1}{2}}{r_b} 2^{-n_b-r_b-\frac{1}{2}}\delta\left(G-\textstyle\sum_b g_b r_b\right)\delta\left(R-\textstyle\sum_b r_b\right)\end{aligned} \tag{7.4.1}$$

and using the identity

$$\frac{r}{n+\frac{1}{2}}\binom{n+r-\frac{1}{2}}{r} = \sum_{k=1}^{\infty}\binom{r+n-\frac{1}{2}-k}{r-k} \tag{7.4.2}$$

we obtain

$$\begin{aligned}\langle r_c|G\rangle = \frac{n_c+\frac{1}{2}}{\Omega[R,G]}\sum_{k,\boldsymbol{r}}\binom{r_c+n_c-\frac{1}{2}-k}{n_c-k}2^{-n_c-r_c-\frac{1}{2}} \\ \times\prod_{b\neq c}^{B}\binom{n_b-\frac{1}{2}+r_b}{r_b}2^{-n_b-r_b-\frac{1}{2}}\delta\left(G-\textstyle\sum_b g_b r_b\right)\delta\left(R-\textstyle\sum_b r_b\right).\end{aligned} \tag{7.4.3}$$

Relabelling the r_c term simplifies the equation of $\langle r_c|G\rangle$,

$$\langle r_c|G\rangle = \frac{n_c+\frac{1}{2}}{\Omega[R,G]}\sum_{k,\boldsymbol{r}}\prod_b^B\binom{n_b-\frac{1}{2}+r_b}{r_b}2^{-n_b-r_b-k-\frac{1}{2}}\delta\left(G-kg_c-\textstyle\sum_b g_b r_b\right)\delta\left(R-k-\textstyle\sum_b r_b\right) \tag{7.4.4}$$

or, in terms of structure functions and generalising (4.8.3),

$$\rho_c' \;=\; \frac{\langle r_c|G\rangle}{V} = \rho_c\sum_{k=1}^{\infty}\frac{2^{-k}\Omega[R-k,G-kg_c]}{\Omega[R,G]}. \tag{7.4.5}$$

Similarly we can show that

$$\boxed{\begin{aligned}\langle r_c|G\rangle &= V\rho_c\sum_{k=1}^{\infty}\frac{2^{-k}\Omega[R-k,G-kg_c]}{\Omega[R,G]} \\ \langle r_c,r_d|G\rangle &= V\rho_c V\rho_d\sum_{j,k=1}^{\infty}\frac{2^{-j-k}\Omega\left[R-j-k,G-g_c j-g_d k\right]}{\Omega\left[R,G\right]}, \quad \forall c\neq d \\ \langle r_c(r_c-1)|G\rangle &= V\rho_c(V\rho_c+1)\sum_{k=1}^{\infty}(k-1)\frac{2^{-k}\Omega\left[R-k,G-kg_c\right]}{\Omega\left[R,G\right]}.\end{aligned}} \tag{7.4.6}$$

7.5 Saddlepoint approximations for exchangeable trials

To approximate the structure function, we must do two saddlepoint calculations simultaneously, one for the λ parameter,

$$-\frac{\partial}{\partial\lambda}\log Z\left[\mu,\lambda\right]\Big|_{\lambda=\lambda^*} = \sum_b \frac{\rho_b g_b}{2e^{\mu+\lambda g_b}-1}\Bigg|_{\lambda=\lambda^*} = \frac{G}{V} = \gamma \tag{7.5.1}$$

and one for the μ parameter,

$$-\frac{\partial}{\partial\mu}\log Z\left[\mu,\lambda\right]\Big|_{\mu=\mu^*} = \sum_b \frac{\rho_b}{2e^{\mu+\lambda g_b}-1}\Bigg|_{\mu=\mu^*} = \frac{R}{V} = 1. \tag{7.5.2}$$

In general the ratio $\frac{R}{V}$ can be chosen to be less than one, but we choose the ratio equal to one based on symmetry. We wish to use the negative hypergeometric distribution as a distance function so we cannot weight one vector of counts as more important than the other.

Solving both these equations simultaneously is the major computational effort that goes into the approximation. The accuracy of the approximation is then described by the second derivative around the saddle point,

$$\begin{aligned}
\sigma_{\lambda,\lambda} &= \frac{\partial^2}{\partial\lambda^2}\log Z\left[\mu,\lambda\right]\Big|_{\mu=\mu^*,\lambda=\lambda^*} = \sum_b \frac{2\rho_b g_b^2 e^{\mu+\lambda g_b}}{\left(2e^{\mu+\lambda g_b}-1\right)^2} &&> 0\\
\sigma_{\mu,\mu} &= \frac{\partial^2}{\partial\mu^2}\log Z[\theta,\lambda]\Big|_{\mu=\mu^*,\lambda=\lambda^*} = \sum_b \frac{2\rho_b e^{\mu+\lambda g_b}}{\left(2e^{\mu+\lambda g_b}-1\right)^2} &&> 0
\end{aligned} \tag{7.5.3}$$

and the covariance between the two

$$\sigma_{\mu,\lambda} = \frac{\partial^2}{\partial\mu\partial\lambda}\log Z\left[\mu,\lambda\right]\Big|_{\mu=\mu^*,\lambda=\lambda^*} = \sum_b \frac{2\rho_b g_b e^{\mu+\lambda g_b}}{\left(2e^{\mu+\lambda g_b}-1\right)^2}. \tag{7.5.4}$$

The approximation to $\Omega\left[R,G\right]$ is then, using (2.6.10)

$$\begin{aligned}
&\frac{Z\left[\mu^*,\lambda^*\right]^V e^{\mu^* R+\lambda^* G}}{-4\pi^2}\\
&\times \int_{-i\infty}^{i\infty}\int_{-i\infty}^{i\infty} \exp\left[V\frac{\sigma_{\mu,\mu}(\mu-\mu^*)^2+2\sigma_{\mu,\lambda}(\mu-\mu^*)(\lambda-\lambda^*)+\sigma_{\lambda,\lambda}(\lambda-\lambda^*)^2}{2}\right]\mathrm{d}\mu\,\mathrm{d}\lambda\\
&\simeq \frac{Z\left[\mu^*,\lambda^*\right]^V e^{\mu^* R+\lambda^* G}}{2\pi V\sqrt{\sigma_{\mu,\mu}\sigma_{\lambda,\lambda}-\sigma_{\mu,\lambda}^2}}.
\end{aligned} \tag{7.5.5}$$

Applying this approximation to our three predictions (7.4.6) we find

$$\begin{aligned}
\frac{\langle r_c|G\rangle}{V} &= \frac{\rho_c}{2e^{\mu^*+g_c\lambda^*}-1},\\
\frac{\langle r_c, r_d|G\rangle}{V^2} &= \left(\frac{\rho_c}{2e^{\mu^*+g_c\lambda^*}-1}\right)\left(\frac{\rho_d}{2e^{\mu^*+g_d\lambda^*}-1}\right) \quad \forall c\neq d,\\
\frac{\langle r_c(r_c-1)|G\rangle}{V^2} &= \frac{\rho_c(\rho_c+1)}{\left(2e^{\mu^*+g_c\lambda^*}-1\right)^2}.
\end{aligned} \tag{7.5.6}$$

The prediction for the next trial would be,

$$\rho'_c = \frac{\langle r_c|G\rangle}{V} = \frac{\rho_c}{2e^{\mu^*+g_c\lambda^*}-1}, \tag{7.5.7}$$

which from equation (7.5.2) will sum to one and from equation (7.5.1) have the correct mean γ.

7.6 Reconsidering relative entropy

Again we seek an entropy function to maximize. First we try to find the most probable set of counts r_b^*. Applying the discrete difference operator ∇ to our product of oscillators,

$$\nabla_{r_c}\log\left[\prod_b \binom{n_b+r_b-{}^1\!/_2}{r_b} 2^{-r_b-n_b-{}^1\!/_2} e^{-\mu r_b-\lambda g_b r_b}\right] = 0 \quad \forall\, c, \tag{7.6.1}$$

the resulting set of formulas are

$$\log\left(\frac{n_b+r_b-{}^1\!/_2}{2r_b}\right) - g_b\lambda - \mu = 0 \quad \forall c \tag{7.6.2}$$

and solving, we obtain

$$r_b^* = \frac{V\rho_b - {}^1\!/_2}{2e^{\mu^*+g_b\lambda^*}-1}, \tag{7.6.3}$$

so for small V the most probable and mean values do not coincide, but the difference decreases as we increase V.

For a generalised principle of Maximum Entropy we first apply Stirling's approximation directly to the product to find

$$H[\boldsymbol{\rho}'||\boldsymbol{\rho}] = \sum_b\left[-\rho_b\log\left(\frac{\rho'_b+\rho_b}{2\rho_b}\right) - \rho'_b\log\left(\frac{\rho'_b+\rho_b}{2\rho'_b}\right)\right], \tag{7.6.4}$$

which we call a **Bose-Einstein** Divergence. From the inequality $-\log x \geq 1-x$ with equality if and only if $x=1$, we have,

$$\sum_b\left[-\rho_b\log\left(\frac{\rho'_b+\rho_b}{2\rho_b}\right) - \rho'_b\log\left(\frac{\rho'_b+\rho_b}{2\rho'_b}\right)\right] \geq \sum_b\left[\rho_b - \frac{\rho'_b+\rho_b}{2} + \rho'_b - \frac{\rho'_b+\rho_b}{2}\right] = 0, \tag{7.6.5}$$

so that the Bose-Einstein divergence is always positive and will be zero if and only if $\rho_b = \rho'_b$. This is our new entropy function which would be maximised to construct models corresponding to exchangeable trials with an ignorance prior. It seems to be an improvement over the Maxwell-Boltzmann (Kullback-Leibler) divergence because it is more general and it is symmetric in its arguments.

We now show that the Bose-Einstein Divergence is a convex function but not a distance (norm). Note that $f[t] = t\log t$ is a strictly convex function, since $f''[t] = \frac{1}{t}$ for all positive

t and rewriting Jensen's inequality as formulated in Cover and Thomas (2012) for convex functions f

$$\sum_j w_j f(x_j) \geq f\left[\sum_j w_j x_j\right], \qquad x_j \geq 0, \quad \sum_j w_j = 1, \tag{7.6.6}$$

with $\lambda' = 1 - \lambda$ and

$$\begin{aligned} \boldsymbol{w} &= \left\{ \frac{\lambda(\rho+f)}{\lambda(\rho+f)+\lambda'(\rho'+f')}, \frac{\lambda'(\rho'+f')}{\lambda(\rho+f)+\lambda'(\rho'+f')} \right\} \\ \boldsymbol{x} &= \left\{ \frac{\rho}{2(\rho+f)}, \frac{\rho'}{2(\rho'+f')} \right\} \quad \text{or} \quad \left\{ \frac{f}{2(\rho+f)}, \frac{f'}{2(\rho'+f')} \right\} \end{aligned} \tag{7.6.7}$$

gives us

$$\begin{aligned} &(\lambda\rho+\lambda'\rho')\log\frac{\lambda\rho+\lambda'\rho'}{\lambda(\rho+f)+\lambda'(\rho'+f')} + (\lambda f+\lambda' f')\log\frac{\lambda f+\lambda' f'}{\lambda(\rho+f)+\lambda'(\rho'+f')} \\ &\leq \lambda p\log\frac{\lambda p}{\lambda(p+f)} + \lambda f\log\frac{\lambda f}{\lambda(p+f)} + \lambda' p'\log\frac{\lambda' p'}{\lambda'(p'+f')} + \lambda' f'\log\frac{\lambda' f'}{\lambda'(p'+f')}, \end{aligned} \tag{7.6.8}$$

after multiplying with $\lambda(\rho+f)+\lambda'(\rho'+f')$. Replacing ρ, ρ', f, f' with $\rho_b, \rho_b', f_b, f_b'$ respectively and summing over all b we have

$$H\left[\lambda\boldsymbol{\rho}+(1-\lambda)\boldsymbol{\rho}'||\lambda\boldsymbol{f}+(1-\lambda)\boldsymbol{f}'\right] \leq \lambda H\left[\boldsymbol{\rho}||\boldsymbol{f}\right] + (1-\lambda)H\left[\boldsymbol{\rho}'||\boldsymbol{f}'\right]. \tag{7.6.9}$$

So the Bose-Einstein divergence is a convex function in the pair $\{\rho, f\}$, but it is still not a distance between distributions as it does not satisfy the triangle inequality as a simple counter example will show: Take,

$$\boldsymbol{f}_0 = \left\{\frac{1}{2}, \frac{1}{2}\right\} \qquad \boldsymbol{f}_1 = \left\{\frac{1}{3}, \frac{2}{3}\right\} \qquad \boldsymbol{f}_2 = \left\{\frac{1}{6}, \frac{5}{6}\right\}, \tag{7.6.10}$$

then

$$H[\boldsymbol{f}_0||\boldsymbol{f}_1] \approx 0.0287252 \qquad H[\boldsymbol{f}_1||\boldsymbol{f}_2] \approx 0.0375949 \qquad H[\boldsymbol{f}_2||\boldsymbol{f}_0] \approx 0.12932, \tag{7.6.11}$$

i.e. it is shorter to go from f_0 to f_2 via f_1 than to go directly from f_0 to f_2.

7.7 Examples

Consider again first the Planck oscillator $\rho_b = \text{const.}$, $g_b = b = 0, 1, 2, \ldots$, so that the partition function

$$\log Z[\mu, \lambda] = \text{const.} \sum_{b=0}^{\infty} \log \frac{1/2}{1 - 1/2 e^{-\mu - g_b\lambda}} \tag{7.7.1}$$

is approximated with the two saddlepoint approximations,

$$\begin{aligned} -\frac{\mathrm{d}}{\mathrm{d}\mu}\log Z[\mu,\lambda]\Big|_{\mu=\mu^*} &= \sum_{b=0}^{\infty}\frac{1}{2e^{\mu-b\lambda}-1}\Big|_{\mu=\mu^*} \approx \frac{-\log\left[1-1/2 e^{-\mu^*}\right]}{\lambda^*} = 1 \\ -\frac{\mathrm{d}}{\mathrm{d}\lambda}\log Z[\mu,\lambda]\Big|_{\lambda=\lambda^*} &= \sum_{b=0}^{\infty}\frac{b}{2e^{\mu-b\lambda}-1}\Big|_{\mu=\mu^*} \approx \frac{\mathrm{Li}_2\left[1/2 e^{-\mu^*}\right]}{{\lambda^*}^2} = \gamma \end{aligned} \tag{7.7.2}$$

where we have replaced the sum over b with a continuous integral to keep the answer analytical and $\mathrm{Li}_n[x]$ is the polylogarithm,

$$\mathrm{Li}_n[x] = \sum_{k=1}^{\infty} \frac{x^k}{k^n}. \tag{7.7.3}$$

The Fermi oscillator, secondly, has partition function

$$Z[\mu,\lambda]^V = \left(\frac{1/2}{1/2 - e^{-\mu}}\right)^{V/2} \left(\frac{1/2}{1/2 - e^{-\mu-\lambda}}\right)^{V/2}, \tag{7.7.4}$$

and inverting we find the structure function

$$\Omega[R,G] = 2^{-V-R} \binom{G+V/2-1}{V/2-1} \binom{V/2+R-G-1}{V/2-1} \qquad R \geq G. \tag{7.7.5}$$

Using (7.4.6), our predictions are

$$\rho'_s = \frac{G}{V} = \gamma \qquad\qquad \rho'_f = \frac{R-G}{V} = 1-\gamma. \tag{7.7.6}$$

Compare this to (4.12.23) which is essentially the same answer.

Thirdly, we return to the Kangaroo example as defined in (4.12.27) and first try to solve it analytically. Define G_1 as the number of blue-eyed kangaroos and G_2 as the number of left-handed kangaroos. The partition function is

$$Z[\mu,\lambda_1,\lambda_2]^V = \left[\left(2-e^{-\mu}\right)\left(2-e^{-\mu-\lambda_1}\right)\left(2-e^{-\mu-\lambda_2}\right)\left(2-e^{-\mu-\lambda_1-\lambda_2}\right)\right]^{-V/4}, \tag{7.7.7}$$

and after some algebra we find the Kangaroo structure function

$$\Omega\left[R,G_1,G_2\right] = 2^{-R-V}(-1)^R \sum_{g=0}^{G_1} \binom{-V/4}{g} \binom{-V/4}{G_1-g} \binom{-V/4}{G_2-G_1+g} \binom{-V/4}{R-g-G_2}. \tag{7.7.8}$$

Explaining what the structure function is doing is pretty simple. We label the number of kangaroos of a certain outcome with g specifically $(\mathcal{B},\overline{\mathcal{L}})$. The constraint on the total associated with $\mathcal{B}$ is G_1 thus g cannot exceed G_1 and the number of trials in $(\mathcal{B},\mathcal{L})$ must be G_1-g. The same holds for $(\overline{\mathcal{B}},\mathcal{L})$ with G_2-G_1-g which must add up to G_2. The last outcome $(\overline{\mathcal{B}},\overline{\mathcal{L}})$ must ensure that all the trials add up to R, thus $R-g-G_2$. Each of these trials is drawn from a negative binomial distribution and the structure function is then the sum of all possible values of g. If we were to use (7.4.6) the mathematics would become unwieldy for a problem that is conceptually very simple. So let us take a shortcut by focusing on the most probable value of g,

$$\begin{aligned} 0 &= \nabla_g \log \binom{-V/4}{g} \binom{-V/4}{G_1-g} \binom{-V/4}{G_2-G_1+g} \binom{-V/4}{R-g-G_2} 2^{-R-V}(-1)^R \qquad (7.7.9) \\ &= \log\left(\frac{-1+g+G_2-R}{g+G_2-R-V/4}\right)\left(\frac{-1+g+V/4}{g}\right)\left(\frac{1-g+G_1}{-g+G_1+V/4}\right)\left(\frac{-1+g-G_1+G_2+V/4}{g-G_1+G_2}\right) \end{aligned}$$

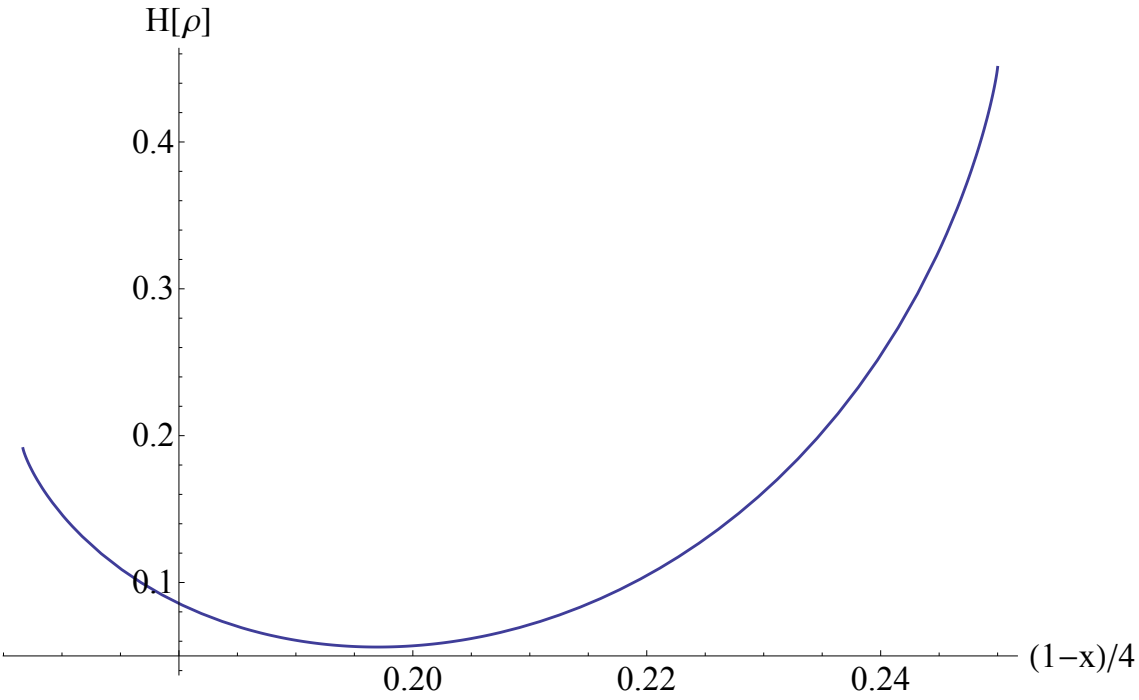


Figure 7.1: Bose-Einstein divergence (7.6.4) of the Kangaroo example as a function of the correlation $(1-x)/4$

which for the large V case with $x = \frac{g}{V}$, $\gamma_1 = \frac{G_1}{V}$, $\gamma_2 = \frac{G_2}{V}$ and $V = R$ results in

$$0 = 2\log\left(\frac{1+4x}{4x}\right) + \log\left(1+\frac{3}{12x-11}\right) + \log\left(1+\frac{3}{12x-7}\right) \tag{7.7.10}$$

The same equation could have been found by minimising the Bose-Einstein divergence. Plotting in Figure (7.1) the Bose-Einstein Entropy as a function of $(1-x)/4$, which is an indication of the amount of correlation between $\mathcal{B}$ and $\mathcal{L}$, we see that the presence of a constraint on the marginals removes some of the negative correlation as possibilities. In any case the negative binomial distribution prefers positive correlations as we must allow for infinite sequences and in infinite sequences there are no negative correlations. Solving the equation gives us

$$\begin{aligned} 0 &= 144x^3 - 108x^2 + 7x + 2, \quad 0 \leq x \leq 1/3 \\ x &= 0.212108. \end{aligned} \tag{7.7.11}$$

This solution differs from that found with the conventional maximum entropy maximisation, which we previously determined as $x = \frac{2}{9}$. In fact we expect positive correlation between two unknown propositions because it is far more likely than independence.

7.8 Finite populations

Having treated the case of infinite populations, we now turn back to the case of finite populations, which we shall indicate by the use of the symbol $\mathcal{F}$. Consequences of finiteness

are immediate. The Laplace's rule of succession, for example, does not hold because if the population size is known and fixed then the trials must be negatively correlated. Fixing the population size in each bin n_b or $V\rho_b$ we then add a constraint to the system that changes the trials r_b that we can see. In this case the data consists of us knowing the populations sizes in each bin.

The hypergeometric prior and likelihood are

$$p(\boldsymbol{m}|\boldsymbol{n},\mathcal{F}) = \binom{N}{M}^{-1}\prod_b \binom{n_b}{m_b}, \qquad p(\boldsymbol{r}|R,\boldsymbol{m},\mathcal{F}) = \binom{M}{R}^{-1}\prod_b \binom{m_b}{r_b}, \tag{7.8.1}$$

and the resulting evidence and posterior are

$$p(\boldsymbol{r}|R,\boldsymbol{n},\mathcal{F}) = \binom{N}{R}^{-1}\prod_b \binom{n_b}{r_b}, \quad p(\boldsymbol{m}|R,\boldsymbol{r},\boldsymbol{n},\mathcal{F}) = \binom{N-R}{M-R}^{-1}\prod_b \binom{n_b-r_b}{m_b-r_b}. \tag{7.8.2}$$

R and r_b now enter the single-trial prediction with minus signs,

$$p(r_c{=}1\,|\,R{=}1,\boldsymbol{n},\mathcal{F}) = \frac{n_c - r_c}{N-R}, \tag{7.8.3}$$

which is the opposite of the usual rule of succession: Seeing a particular outcome reduces the probability of the same outcome happening again. The prediction for general $\boldsymbol{r}$ is

$$p(\boldsymbol{r}|R,\boldsymbol{n},\mathcal{F}) = \frac{1}{N^{r_b}}\binom{n_b}{r_b} = \frac{1}{r_b!}\frac{n_b}{N}\frac{n_b-1}{N}\ldots\frac{n_b-r_b}{N} \tag{7.8.4}$$

and in the large population limit $N\to\infty$ with $\rho_b = \frac{n_b}{N}$ gives

$$\frac{1}{N^{r_b}}\binom{n_b}{r_b} \to \frac{\rho_b^{r_b}}{r_b!} \qquad\qquad N^{r_b}\binom{N}{R}^{-1} \to R!, \tag{7.8.5}$$

and thus the prediction again reduces to the logical independent limit

$$\lim_{N\to\infty} p(\boldsymbol{r}|R,\boldsymbol{n},\mathcal{F}) = R!\prod_b^B \frac{\rho_b^{r_b}}{r_b!}. \tag{7.8.6}$$

7.9 Prior distribution for R

As before we must introduce a prior for the number of trials to switch to the Grand Canonical view. In analogy with our previous prior (7.2.1) the obvious choice is a **Binomial** distribution,

$$p(R|N,\mathcal{F}) = \binom{N}{R}\theta^R(1-\theta)^{N-R}. \tag{7.9.1}$$

and combining it with the hypergeometric distribution, yields the evidence

$$p(R,\boldsymbol{r}|\boldsymbol{n},\mathcal{F}) = \sum_{R=0}^{\infty} p(\boldsymbol{r}|\boldsymbol{n},R,\mathcal{F})p(R|\mathcal{F}) = \prod_b \binom{n_b}{r_b}\theta^{r_b}(1-\theta)^{n_b-r_b} \tag{7.9.2}$$

The motivation for this choice is as before Fisher's artifice,

$$p(R,\boldsymbol{r}|\boldsymbol{n},\mathcal{F}) = \delta\left(R-\sum_b r_b\right)\prod_b^B \binom{n_b}{r_b}\theta^{r_b}(1-\theta)^{n_b-r_b}, \tag{7.9.3}$$

where the total trials are distributed like the individual trials,

$$p(R|\boldsymbol{n},\mathcal{F}) = \binom{N}{R}\theta^R(1-\theta)^{N-R} \tag{7.9.4}$$

and if we condition on it we recover our original hypergeometric distribution,

$$p(\boldsymbol{r}|R,\boldsymbol{n},\mathcal{F}) = \frac{p(\boldsymbol{r}|\boldsymbol{n},\mathcal{F})}{p(R|\boldsymbol{n},\mathcal{F})} = \binom{N}{R}^{-1}\prod_b \binom{n_b}{r_b}. \tag{7.9.5}$$

Applying the pinciple of indifference to our Binomial distribtuion we set $\theta = 1/2$,

$$p(\boldsymbol{r}|R,\boldsymbol{n},\mathcal{F}) = \prod_b \binom{n_b}{r_b}2^{-n_b}. \tag{7.9.6}$$

Our prior distribution for the amount of trials,

$$p(R|\boldsymbol{n},\mathcal{F}) = \binom{N}{R}2^{-N}, \tag{7.9.7}$$

gives us the expected amount of R,

$$\langle R\rangle = \frac{N}{2}. \tag{7.9.8}$$

We also define $\rho'_b = \frac{\langle r_b|G\rangle}{V}$ as before.

7.10 Constraints for finite trials

Adding a constraint for G and a constraint for R,

$$p(G,R,\boldsymbol{r}|\boldsymbol{n},\mathcal{F}) = \delta\left(G-\sum_b g_b r_b\right)\delta\left(R-\sum_b r_b\right)\prod_b\binom{n_b}{r_b}2^{-n_b} \tag{7.10.1}$$

gives us the exponential generating function ($G\to\lambda, R\to\mu$),

$$\begin{aligned}\Phi[\{\mu,\lambda\},\{R,G\}|p(G,R,\boldsymbol{r}|\boldsymbol{n},\mathcal{F})] &= \sum_{R,G,\boldsymbol{r}} p(R,G,\boldsymbol{r}|\boldsymbol{n},\mathcal{F})e^{-\mu R-\lambda G}\\ &= \prod_b\left(1/2+1/2e^{-\mu-\lambda g_b}\right)^{a_b}.\end{aligned} \tag{7.10.2}$$

The partition function is defined as,

$$Z[\mu,\lambda] = \left(1/2+1/2e^{-\mu-\lambda g_b}\right)^{\rho_b}, \tag{7.10.3}$$

and with $V=N$ and $\rho_b = \frac{n_b}{N}$, we have

$$\sum_{R,G}\Omega[R,G]e^{-\mu R-\lambda G} = Z[\mu,\lambda]^V. \tag{7.10.4}$$

7.11 Predictions using the finite structure function

Using the identity

$$\frac{r}{n}\binom{n}{r} = \sum_{k=1}^{\infty} -(-1)^k \binom{a}{r-k} \tag{7.11.1}$$

and from the definition of $\langle r_c|G\rangle = \sum_{\boldsymbol{r}} r_c p(\boldsymbol{r}|G,\mathcal{F})$,

$$\langle r_c|G\rangle = \frac{1}{\Omega[R,G]} \sum_{\boldsymbol{r}} r_c \prod_b^B \binom{n_b}{r_b} 2^{-n_b} \delta\left(G - \sum g_b r_b\right) \delta\left(R - \sum r_b\right) \tag{7.11.2}$$

$$= \frac{n_c}{\Omega[R,G]} \sum_{k,\boldsymbol{r}} -(-1)^k \binom{n_j}{r_c-k} 2^{-n_c} \prod_{b\neq c}^B \binom{a_b}{r_b} 2^{-n_b} \delta\left(G - \sum g_b r_b\right) \delta\left(R - \sum r_b\right) \tag{7.11.3}$$

$$= \frac{n_c}{\Omega[R,G]} \sum_{k,\boldsymbol{r}} -(-1)^k \prod_b^B \binom{n_b}{r_b} 2^{-r_b} \delta\left(G - kg_c - \sum g_b r_b\right) \delta\left(R - k - \sum r_b\right), \tag{7.11.4}$$

where we relabelled the r_c term in the last line. Simplifying the solution we find

$$\langle r_j|G\rangle = n_c \sum_{k=1}^{\infty} \frac{-(-1)^k \Omega[R-k, G-kg_c]}{\Omega[R,G]}. \tag{7.11.5}$$

Similarly we can derive,

$$\boxed{\begin{aligned} \langle r_c|G\rangle &= n_c \sum_{k=1}^{\infty} \frac{-(-1)^k \Omega[R-k, G-kg_c]}{\Omega[R,G]} \\ \langle r_c, r_d|G\rangle &= n_c n_d \sum_{j=1,k=1}^{\infty} \frac{(-1)^{j+k} \Omega\left[R-j-k, G-g_c j - g_d k\right]}{\Omega\left[R,G\right]} \quad c \neq d \\ \langle r_c(r_c-1)|G\rangle &= n_c(n_c-1) \sum_{k=1}^{\infty} (k-1)(-1)^k \frac{\Omega\left[R-k, G-kg_c\right]}{\Omega\left[R,G\right]}. \end{aligned}} \tag{7.11.6}$$

7.12 Saddlepoint approximation for finite populations

Repeating our saddlepoint method of approximation on the finite population partition function

$$-\frac{\partial}{\partial \mu} \log Z[\mu,\lambda]\Big|_{\mu=\mu^*} = \sum_b \frac{\rho_b}{1+e^{\mu^* + g_b \lambda^*}} = \frac{R}{V} \tag{7.12.1}$$

and the other saddle point equation,

$$-\frac{\partial}{\partial \lambda} \log Z[\mu,\lambda]\Big|_{\lambda=\lambda^*} = \sum_b \frac{g_b \rho_b}{1+e^{\mu^* + g_b \lambda^*}} = \frac{G}{V} = \gamma, \tag{7.12.2}$$

the variances are,

$$\begin{aligned}\sigma_{\mu,\mu} &= \frac{\partial^2}{\partial\mu^2}\log Z\left[\mu,\lambda\right]\Big|_{\mu=\mu^*,\lambda=\lambda^*} = \sum_b \frac{\rho_b e^{\mu+\lambda g_b}}{\left(1+e^{\mu+\lambda g_b}\right)^2} > 0\\ \sigma_{\lambda,\lambda} &= \frac{\partial^2}{\partial\mu^2}\log Z[\mu,\lambda]\Big|_{\mu=\mu^*,\lambda=\lambda^*} = \sum_b \frac{\rho_b g_b^2 e^{\mu+\lambda g_b}}{\left(1+e^{\mu+\lambda g_b}\right)^2} > 0\end{aligned} \tag{7.12.3}$$

and the covariance between the two is

$$\sigma_{\mu,\lambda} = \frac{d^2}{d\mu d\lambda}\log Z\left[\mu,\lambda\right]\Big|_{\mu=\mu^*,\lambda=\lambda^*} = \sum_b \frac{\rho_b g_b e^{\mu+\lambda g_b}}{\left(1+e^{\mu+\lambda g_b}\right)^2}. \tag{7.12.4}$$

The approximation is then,

$$\begin{aligned}&\Omega\left[R,G\right]\\ &= \frac{Z\left[\mu^*,\lambda^*\right]^V e^{\mu^* R+\lambda^* G}}{-4\pi^2}\\ &\times \int_{-i\infty}^{i\infty}\int_{-i\infty}^{i\infty} \exp\left[V\frac{\sigma_{\mu,\mu}(\mu-\mu^*)^2+2\sigma_{\mu,\lambda}(\mu-\mu^*)(\lambda-\lambda^*)+\sigma_{\lambda,\lambda}(\lambda-\lambda^*)^2}{2}\right]d\mu d\lambda\\ &= \frac{Z\left[\mu^*,\lambda^*\right]^V e^{\mu^* R+\lambda^* G}}{2\pi V\sqrt{\sigma_{\mu,\mu}\sigma_{\lambda,\lambda}-\sigma_{\mu,\lambda}^2}}.\end{aligned} \tag{7.12.5}$$

Applying the approximation to our predictive formulas (7.11.6) we find,

$$\begin{aligned}\langle r_c|G\rangle &= \frac{V\rho_c}{1+e^{\mu^*+g_c\lambda^*}}\\ \langle r_c, r_d|G\rangle &= \frac{V\rho_c}{1+e^{\mu^*+g_c\lambda^*}}\frac{V\rho_d}{1+e^{\mu^*+g_d\lambda^*}}\\ \langle r_c(r_c-1)|G\rangle &= \frac{\rho_c\left(V\rho_c-1\right)}{\left(e^{\mu^*+g_c\lambda^*}+1\right)^2}.\end{aligned} \tag{7.12.6}$$

The prediction for the next trial would be,

$$R\rho_c' = \langle r_c|G\rangle = \frac{R\rho_c}{1+e^{\mu^*+g_j\lambda^*}}, \tag{7.12.7}$$

which from equation (7.12.2) will sum to one and from equation (7.12.1) have the correct mean.

7.13 Fermi-Dirac entropy

Again we seek an entropy function to maximize. Taking our generalised binomial distribution with $\theta = \frac{1}{2}$ and adding a Lagrange multiplier,

$$p(R,G,\boldsymbol{r}|\boldsymbol{n},\mathcal{F}) = \prod_b \binom{n_b}{r_b} 2^{-n_b} e^{-\mu r_b-\lambda g_b r_b}. \tag{7.13.1}$$

Searching for the set of predicted counts $\boldsymbol{r}$ that maximizes the probability is equivalent to taking the logarithm and examining where the difference in counts, $\nabla_c f(r_c) = f(r_c) - f(r_c - 1)$, in counts is zero,

$$\left[\log \frac{1 + n_c - r_c}{r_c} - \mu - g_c\lambda\right] \delta r_c = 0. \tag{7.13.2}$$

Defining the ratios,

$$\rho_c = \frac{n_c}{V} \qquad \rho'_c = \frac{r_c}{V} \qquad \nu = \frac{R}{V}, \tag{7.13.3}$$

the bulk limit becomes

$$\sum_b \left[\log\left[\frac{\rho_b}{\nu\rho'_b} - 1\right] - \mu - g_b\lambda\right] \delta\rho_b = 0. \tag{7.13.4}$$

Integrating in ρ'_b we find,

$$H\left[\boldsymbol{\rho}'||\boldsymbol{\rho}\right] = \sum_b \left[-\rho'_b \log\left(\frac{\rho_b - \nu\rho'_b}{\nu\rho'_b}\right) + \frac{\rho_b}{\nu}\log\left(\frac{\rho_b - \nu\rho'_b}{\rho_b}\right) - \mu\rho'_b - \lambda g_b\rho'_b\right], \tag{7.13.5}$$

which we could also have found by applying Stirling's expansion directly to the generalised binomial distribution. Thus our new entropy function to construct models with finite populations,

$$H\left[\boldsymbol{\rho}', \boldsymbol{\rho}\right] = \sum_b \left[-\rho'_b \log\left(\frac{\rho_b - \nu\rho'_b}{\nu\rho'_b}\right) + \frac{\rho_b}{\nu}\log\left(\frac{\rho_b - \nu\rho'_b}{\rho_b}\right)\right] \tag{7.13.6}$$

which we will call the **Fermi-Dirac Divergence**.

7.14 Examples

Our first example is the Planck Oscillator again: $\rho_b = 1/B$, $g_b = b$ and $b = 0, 1, 2, 3, \ldots$ so that the partition function is,

$$\log Z[\mu, \lambda] \propto \sum_{b=0}^{\infty} \log\left(\tfrac{1}{2} + \tfrac{1}{2}e^{-\mu - b\lambda}\right) \tag{7.14.1}$$

and approximating it by solving the saddlepoint equations,

$$\begin{aligned} -\frac{\partial}{\partial\mu}\log Z[\mu,\lambda]\Big|_{\mu=\mu^*} &= \sum_{b=0}^{\infty} \frac{1}{1 + e^{\mu^* + b\lambda^*}} \approx \frac{\log\left[1 + e^{-\mu}\right]}{\lambda} = \nu \\ -\frac{\partial}{\partial\lambda}\log Z[\mu,\lambda]\Big|_{\lambda=\lambda^*} &= \sum_{b=0}^{\infty} \frac{b}{1 + e^{\mu^* + b\lambda^*}} \approx -\frac{\mathrm{Li}_2\left[-e^{-\mu}\right]}{\lambda^2} = \gamma, \end{aligned} \tag{7.14.2}$$

where the sum was again approximated with the corresponding integral.

The partition function for the fermi oscillator is: $g_f = 0$ and $g_s = 1$, with $\rho_f = \frac{1}{2}$ and $\rho_s = \frac{1}{2}$,

$$Z[\mu, \lambda]^V = \left(\tfrac{1}{2} + \tfrac{1}{2}e^{-\mu}\right)^{V/2} \left(\tfrac{1}{2} + \tfrac{1}{2}e^{-\mu-\lambda}\right)^{V/2} \tag{7.14.3}$$

$$= 2^{-V}\left(1 + e^{-\mu}\right)^{V/2} \left(1 + e^{-\mu-\lambda}\right)^{V/2} \tag{7.14.4}$$

and the corresponding structure function is

$$\Omega[R,G] = \binom{V/2}{G}\binom{V/2}{R-G}2^{-V} \qquad G \leq R \leq G + V/2 \quad 0 \leq G \leq V/2. \tag{7.14.5}$$

Using (7.11.6) the predictions are,

$$\rho_s' = \frac{G}{R} \qquad\qquad \rho_f' = \frac{(R-G)}{R}. \tag{7.14.6}$$

Redoing the Kangaroo example for the third time we have,

$$\begin{aligned} &Z[\mu,\lambda_1,\lambda_2]^V \\ &= \left(1/2 + 1/2 e^{-\mu}\right)^{V/4}\left(1/2 + 1/2 e^{-\mu-\lambda_1}\right)^{V/4}\left(1/2 + 1/2 e^{-\mu-\lambda_2}\right)^{V/4}\left(1/2 + 1/2 e^{-\mu-\lambda_1-\lambda_2}\right)^{V/4} \end{aligned} \tag{7.14.7}$$

and the structure function,

$$\Omega\left[R,G_1,G_2\right] = 2^{-V}\sum_{g=0}^{G_1}\binom{V/4}{g}\binom{V/4}{G_1-g}\binom{V/4}{G_2-G_1+g}\binom{V/4}{R-g-G_2}. \tag{7.14.8}$$

The structure function is simply the sum of all the possible configurations that are allowed by our constraints. Focusing on the most probable value of g gives us,

$$\begin{aligned} 0 &= \nabla \log \binom{V/4}{g}\binom{V/4}{G_1-g}\binom{V/4}{G_2-G_1+g}\binom{V/4}{R-g-G_2} \\ &= \log\left(\frac{1-g+V/4}{g}\right)\left(\frac{1-g+G_1}{g-G_1+V/4}\right)\left(\frac{1-g+G_1+G_2+V/4}{g-G_1+G_2}\right)\left(\frac{1-g+G_2+R}{g+G_2-R+V}\right) \end{aligned} \tag{7.14.9}$$

and asymptotically with $x = \frac{g}{V\nu}$,$\gamma_1 = \frac{G_1}{V\nu}$, $\gamma_2 = \frac{G_2}{V\nu}$ and $\nu = 1/2$,

$$\log\left[\frac{(x-\gamma_1)(-1+x+\gamma_2)(-1+4x\nu)(-1+4x\nu-4\gamma_1\nu+4\gamma_2\nu)}{x(x-\gamma_1+\gamma_2)(1+4x\nu-4\gamma_1\nu)(1+4\nu(-1+x+\gamma_2))}\right] \tag{7.14.10}$$

Substituting in $\gamma_1 = \frac{1}{3}$, $\gamma_2 = \frac{1}{3}$ and $\nu = \frac{1}{2}$,

$$\log\left[\frac{(-\frac{2}{3}+x)(-\frac{1}{3}+x)(-1+2x)^2}{(1+2(-\frac{2}{3}+x))x^2(\frac{1}{3}+2x)}\right] \tag{7.14.11}$$

and solving gives

$$x = 1/4. \tag{7.14.12}$$

We plot the Fermi-Dirac divergence in Figure (7.2). In this example we assumed that we examined half of the kangaroos out of a population that spread uniformly between the outcomes. Interestingly we find a simple analytical answer which might indicate that we can solve this example analytically.

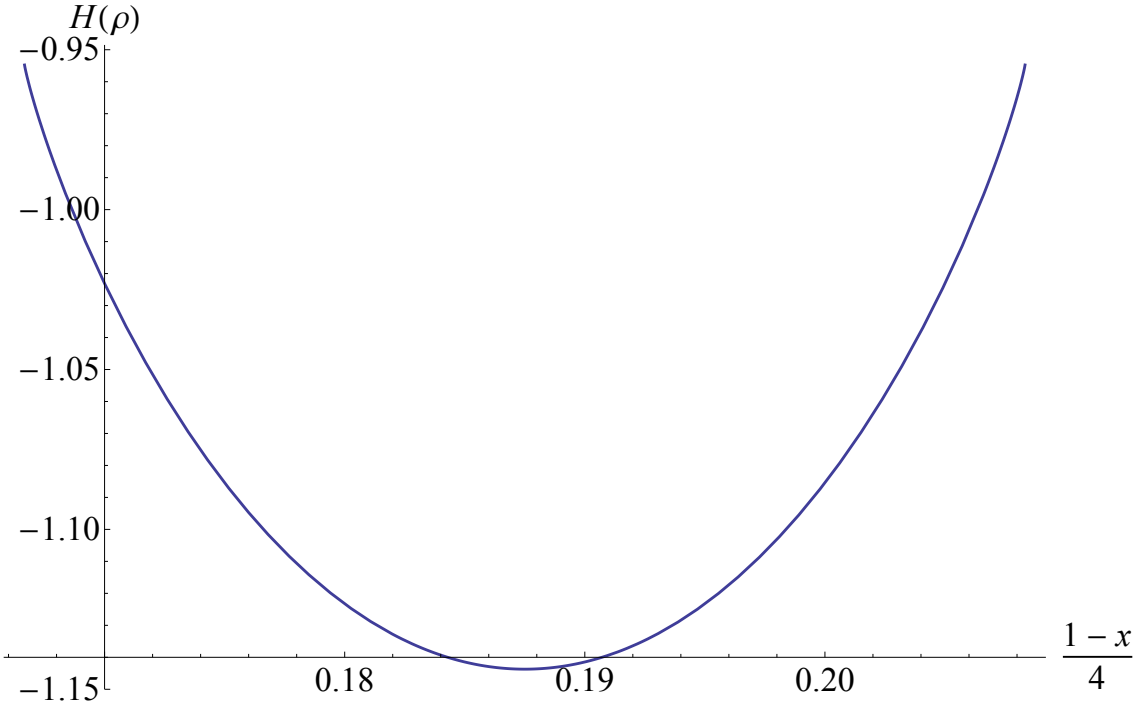


Figure 7.2: Fermi-Dirac entropy of the Kangaroo example as a function of the correlation $(1-x)/4$

Chapter 8

Summary and conclusions

We end by providing a topics-based overview.

1. As outlined in the Introduction, the topic of this dissertation was inspired by the discovery of the conceptual and mathematical superiority of the Bayesian approach to statistical probability and thereby to statistical physics. Reconceptualising probability as originating not as a ratio of counts but as a state of knowledge or information is a fundamental shift that has required us to question, rethink and redefine many ideas and quantities familiar to physics classrooms and laboratories.

2. With Chapter 2 preparing the way by summarising some of the relevant mathematical tools and techniques, the framework for knowledge-based probability is outlined in Chapter 3. The Bayesian concepts of prior, likelihood, evidence and prediction and the corresponding mathematics together provide the foundation which pervades the rest of the work.

3. The shift from a counts-based to a knowledge-based view of probability has not been completed but on the contrary only started. One of the fundamental rethinks concerns the status of the concept of *independence*. The Bayesian view is that the ubiquitous usage of what in physics is called *independence* does not necessarily relate to, or rely on, the physical independence of the quantities under scrutiny but is rather about *logical independence*, whereby knowledge of one particular answer gives no clue about the answer to another. Logical independence is equivalent to the impossibility of learning.

   This dissertation is about a secondary shift within this larger context, namely from *logical independence* to *exchangeability*, an attempt to work out the consequences of replacing the one with the other. *Logical independence* is defined as the factorisation of joint probability, while *exchangeability* is defined by the invariance of probability under permutation of its arguments. Clearly, exchangeability includes logical independence as a special case but is more general.

The central idea developed in this dissertation is that some seemingly arcane cases of exchangeability beyond logical independence have in fact properties that, while mathematically challenging, yield not only some relations and results that seem surprisingly familiar in form (but not in context), but also a bunch of new relations and properties which are new to us and to the field.

4. To set the scene and define our terms for the independence-exchangeability debate, the first part of Chapter 4 focuses on conventional logical independence-based probability, but based not on relative frequency but on information and knowledge. It sets out the foundations and stages of constructing probability based on determining the outcome space and Laplace's Principle of Indifference. We show in some detail how to project from the primordial outcome space onto partitions (the equivalent of binnings in experimental physics) and the corresponding change to counts or occupation numbers. This is done in several ways, including the entirely different approach of waiting-time and stopping-rule problems. Along the way, it becomes clear that both the concepts of *indistinguishability* and *occupation numbers* are associated not with independence per se but rather with the wider class of exchangeability.

   In Section 4.11 it is shown that the Principle of Maximum Entropy is really a special case of the Principle of Minimum Relative Entropy (also termed the Kullback-Leibler divergence) and that the latter could therefore apply in cases where Maximum Entropy would not.

   We have also shown that the usual grand canonical ensemble requires a prior for the number of trials (the equivalent of particle number in gases) which to our knowledge has not been mentioned in the literature before.

5. The simplest example of a logically non-independent but exchangeable sequence is the urn, since the drawing of a ball without replacement changes all the probabilities. Starting with the resulting hypergeometric distribution, we provide in Chapter 5 two views or approaches to probabilities for exchangeable sequences.

   The first is based on the Pólya urn generalisation and conjugate priors. In the second, we rederive two representation theorems for exchangeable sequences commonly known as de Finetti's theorem and the Heath-Sudderth representation. These gives us a general framework to assign probabilities and construct models which can be viewed as an equivalent formulation of Bayes' Theorem.

   The statement that the Laplace-De-Finetti theorem is applicable if we do not have information on the population sizes seems to be a new insight into an old theorem.

6. In Section 5.6, we apply known arguments of stability directly to priors and hence show how various transformation laws can be used to construct stable priors. For example, while Jaynes used invariance of a mean to argue that the likelihood should be a Cauchy distribution, we consider this train of thought to be much more applicable to priors. Normally, such invariance arguments are used to construct forward

probabilities. We believe they are much better suited to the construction of prior probabilities. Some other priors derived in this section such as the stable-scale are, to our knowledge, unknown in the literature.

7. Chapter 7 attempts to follow the steps and logic of the original conventional logical-independence-based statistical physics, but now based on the exchangeability relations of Chapter 5. Among many other relations, we derive a generalised entropy corresponding to the Laplace rule of succession and also an entropy corresponding to finite sampling. This gives us two additional methods of constructing models and analysing correlations. The *Bose-Einstein divergence* and *Fermi-Dirac divergence* are the corresponding generalisations of the Kullback-Leibler divergence and therefore also of the Principle of Minimum Relative Entropy of Chapter 4. We therefore claim that the Principle of Minimum Relative Entropy (Kullback-Leibler divergence) is not the most general formulation and that Bose-Einstein type statistics follows from exchangeablity.

8. New analytical results on posteriors for a correlation coefficient under different circumstances have been relegated to Appendix A because of their high degree of mathematical complexity. We there also derived three different entropies or methods to construct models for them.

9. Examples of various insights and relations are given throughout the dissertation, e.g. certain statistics combined with fixed "energy" levels. In some cases these models are not independent of the prior distribution and the examples lead to a whole set of general families of distributions.

   Many of these little examples are mere hints at what can be done with the formalism and structure developed and are far from complete. At the moment, they are on the level of mathematical proof of concept; physical applications will follow wherever equivalents of these mathematical quantities appear in physics.

   The application in Section 6.2 of Bayes Factors to high energy physics data obtained from the L3 Collaboration is the one exception where real data is analysed. In this dissertation, we have improved on the approach of our conference proceedings in de Kock *et al.* (2011*a*) by introducing Cauchy priors. This attempt to automate the use of Bayes Factors by assigning Cauchy priors to each parameter differs from what is found in the literature and we believe it is an improvement.

   Both the proceedings and the analysis in this section are about understanding the workings of Bayes Factors, priors and the relation to conventional minimum-χ^2 fitting, showing how both χ^2 and Maximum Likelihood also naturally fit into the Bayesian framework, while missing the contributions of the determinants and the automatic Bayes "Occam's penalty" for introducing extra parameters.

10. Logical independence or, as it is usually understood, physical or statistical independence has proven a hugely successful premise for physics. Whether the expansion

to exchangeability will do the same remains to be seen. But there are enough interesting examples and there is enough mathematical analogy in exchangeability to warrant the hope that the physical applications will follow. For example, interpretation of the "sampling without replacement" methodology within the urn analogy in terms of physics and physical experiments is as yet unclear. We do not doubt, however, that interpretations and applications will be found in time.

In conclusion, a simple extension from logical independence to exchangeability has proven to have many consequences. The success of logical independence has been put into a new context and its limitations highlighted. The extensions based on exchangeability are fundamental, wide-ranging and exciting.

Appendices

Appendix A

Hypergeometric functions

We shall find the *hypergeometric function* useful specifically the *Gaussian hypergeometric function*, which is defined as

$$ {}_2F_1\left[\begin{matrix}a,b\\c\end{matrix}\middle|\,x\right]=\sum_{n=0}^{\infty}\frac{a^{\overline{n}}b^{\overline{n}}}{c^{\overline{n}}}\frac{x^n}{n!}=1+\frac{ab}{c}x+\frac{a(a+1)b(b+1)}{c(c+1)2!}x^2+\ldots \tag{A.0.1}$$

see Exton (1976) and similarly,

$$ {}_1F_1\left[\begin{matrix}a\\c\end{matrix}\middle|\,x\right]=\sum_{n=0}^{\infty}\frac{a^{\overline{n}}}{c^{\overline{n}}}\frac{x^n}{n!}\qquad\qquad {}_0F_1\left[\begin{matrix}-\\c\end{matrix}\middle|\,x\right]=\sum_{n=0}^{\infty}\frac{1}{c^{\overline{n}}}\frac{x^n}{n!}. \tag{A.0.2}$$

The subscripts refer to the number of terms in the numerators and denominator and obviously ${}_2F_1\left[a,b;c,x\right]={}_2F_1\left[b,a;c,x\right]$. The functions is not defined for negative values of c and the sum terminates for non-positive integer values of a. When the sum is infinite it converges if $|x|<1$ and diverges if $|x|>1$. It can be represented by an Eulerian integral,

$$ {}_2F_1\left[\begin{matrix}a,b\\c\end{matrix}\middle|\,x\right]=\frac{(c-1)!}{(a-1)!(c-a-1)!}\int_0^1 u^{a-1}(1-u)^{c-a-1}(1-xu)^{-b}du, \tag{A.0.3}$$

where $c>a>0$. Furthermore, the integral A.0.3 remains unchanged in form under the transformations,

$$ u=1-v,\qquad u=v/(1-x-vx)\qquad \text{and}\quad u=(1-u)/(1-vx) \tag{A.0.4}$$

and we have the Euler transformations,

$$\begin{aligned} {}_2F_1\left[\begin{matrix}a,b\\c\end{matrix}\middle|\,x\right]&=(1-x)^{-a}{}_2F_1\left[\begin{matrix}a,c-b\\c\end{matrix}\middle|\,\frac{x}{x-1}\right]\\ &=(1-x)^{-b}{}_2F_1\left[\begin{matrix}c-a,b\\c\end{matrix}\middle|\,\frac{x}{x-1}\right]\\ &=(1-x)^{c-a-b}{}_2F_1\left[\begin{matrix}c-a,c-b\\c\end{matrix}\middle|\,x\right]. \end{aligned} \tag{A.0.5}$$

The incomplete beta function can also be expressed as a Gauss function ($0 < z < 1$),

$$\int_0^z u^{a-1}(1-u)^{c-a-1}du = \frac{z^a}{a}{}_2F_1\left[\begin{matrix}a, 1-c+a \\ 1+a\end{matrix}\middle| z\right]. \tag{A.0.6}$$

The hypergeometric function is also a Laplace transform,

$${}_2F_1\left[\begin{matrix}a, b \\ c\end{matrix}\middle| \frac{k}{s}\right] = \frac{s^b}{(b-1)!}\int_0^\infty e^{-su}u^{b-1}{}_1F_1\left[\begin{matrix}a \\ c\end{matrix}\middle| ku\right]du. \tag{A.0.7}$$

There are many such simplified forms; for example from the negative binomial theorem it follows that,

$${}_2F_1\left[\begin{matrix}a, b \\ b\end{matrix}\middle| x\right] = (1-x)^{-a} = {}_1F_0\left[\begin{matrix}a \\ -\end{matrix}\middle| x\right] = \sum_{n=0}^\infty a^{\overline{n}}\frac{x^n}{n!}. \tag{A.0.8}$$

${}_1F_1$ is called the *confluent hypergeometric function* and follows from

$$\lim_{|b|\to\infty}{}_2F_1\left[\begin{matrix}a, b \\ c\end{matrix}\middle| \frac{x}{b}\right] = {}_1F_1\left[\begin{matrix}a \\ c\end{matrix}\middle| x\right] \tag{A.0.9}$$

If $a = c$ this series simplifies even further and we have ${}_0F_0[x] = e^x$. Again the following Eulerian integral representation is useful,

$${}_1F_1\left[\begin{matrix}a \\ c\end{matrix}\middle| x\right] = \frac{(c-1)!}{(a-1)!(c-a-1)!}\int_0^1 u^{a-1}(1-u)^{c-a-1}e^{xu}du. \tag{A.0.10}$$

The other confluent form was introduced by Tricomi (1947),

$$U\left[\begin{matrix}a \\ a-c+1\end{matrix}\middle| x\right] = x^{-a}{}_2F_0\left[\begin{matrix}a \\ b\end{matrix}\middle| -\frac{1}{x}\right], \tag{A.0.11}$$

which is represented by the integral,

$$U\left[\begin{matrix}a \\ c\end{matrix}\middle| x\right] = \frac{1}{\Gamma[a]}\int_0^\infty e^{-xt}t^{a-1}(1+t)^{c-a-1}dt \tag{A.0.12}$$

and is related to the error function and incomplete gamma functions,

$$\mathrm{Erfc}(x) = \int_x^\infty e^{-t^2}dt = {}^1\!/_2\Gamma\left({}^1\!/_2, x^2\right) = e^{-x^2}U\left[\begin{matrix}{}^1\!/_2 \\ {}^1\!/_2\end{matrix}\middle| x^2\right]. \tag{A.0.13}$$

A.1 Multivariate hypergeometric functions

We can generalise the hypergeometric function to two variables by considering the product of two hypergeometric functions,

$${}_2F_1\left[\begin{matrix}a, b \\ c\end{matrix}\middle| x\right]{}_2F_1\left[\begin{matrix}a', b' \\ c'\end{matrix}\middle| y\right] = \sum_{j_1,j_2}^\infty \frac{a^{\overline{j_1}}(a')^{\overline{j_2}}b^{\overline{j_1}}(b')^{\overline{j_2}}}{c^{\overline{j_1}}(c')^{\overline{j_2}}}\frac{x^{j_1}y^{j_2}}{j_1!j_2!}, \tag{A.1.1}$$

and then replacing each pair of products $a^{\overline{j_1}}(a')^{\overline{j_2}}$ by the composite $a^{\overline{j_1+j_2}}$ and $c^{\overline{j_1}}(c')^{\overline{j_2}}$ by the composite $c^{\overline{j_1+j_2}}$ to obtain a non-factorisable bivariate hypergeometric function. Appell (1880) was the first author to treat this in a systematic way and defined four functions that bear his name. We are only interested in the first function,

$$F_1\left[\begin{matrix}a,b,b'\\c\end{matrix}\middle|\,x,y\right]=\sum_{j_1,j_2}^{\infty}\frac{a^{\overline{j_1+j_2}}b^{\overline{j_1}}(b')^{\overline{j_2}}}{c^{\overline{j_1+j_2}}}\frac{x^{j_1}y^{j_2}}{j_1!j_2!}, \tag{A.1.2}$$

because of its single integral representation due to Picard (1880), which we use in the next section,

$$\begin{aligned}&\frac{(a-1)!}{(c-a-1)!(c-1)!}F_1\left[\begin{matrix}a,b,b'\\c\end{matrix}\middle|\,x,y\right]\\&=\int_0^1 u^{a-1}(1-u)^{c-a-1}(1-ux)^{-b}(1-uy)^{-b}du,\end{aligned} \tag{A.1.3}$$

where $a>0$ and $c-a>0$. Again we can express a double incomplete beta integral with an Appell function $(0<z_1<z_2<1)$,

$$\begin{aligned}\int_{z_1}^{z_2}u^a(1-u)^b&=(z_2-z_1)z_1^a(1-z_1)^bF_1\left[\begin{matrix}1,-a,-b\\2\end{matrix}\middle|\,1-\frac{z_2}{z_1},1-\frac{1-z_2}{1-z_1}\right]\\&=(z_2-z_1)z_2^a(1-z_2)^bF_1\left[\begin{matrix}1,-a,-b\\2\end{matrix}\middle|\,1-\frac{z_1}{z_2},1-\frac{1-z_1}{1-z_2}\right].\end{aligned} \tag{A.1.4}$$

There is also corresponding limiting forms for the Appell functions, which was first discussed by Humbert (1921), for example,

$$\lim_{\epsilon\to 0}F_1\left[\begin{matrix}\frac{1}{\epsilon},b,b'\\c\end{matrix}\middle|\,\epsilon x,y\right]=\Phi_2\left[\begin{matrix}b,b'\\c\end{matrix}\middle|\,x,y\right]=\sum_{k_1,k_2}\frac{b^{\overline{k_1}}(b')^{\overline{k_2}}}{c^{\overline{k_1+k_2}}}\frac{x^{k_1}y^{k_2}}{k_1!k_2!}, \tag{A.1.5}$$

where Φ_2 is the second Humbert function. Functions of more than two variables where investigated by Lauricella (1893), who defined multiple hypergeometric functions, of which we only need the generalisation of the Appell F_1 function,

$$F_D^{(n)}\left[\begin{matrix}a;b_1,\ldots,b_n\\c\end{matrix}\middle|\,x_1,\ldots,x_n\right]=\sum_{\boldsymbol{k}}\frac{a^{\overline{k_1+\cdots+k_n}}b_1^{\overline{k_1}}\ldots b_n^{\overline{k_n}}}{c^{\overline{k_1+\cdots+k_n}}}\frac{x_1^{k_1}\ldots x_n^{k_n}}{k_1!\ldots k_n!} \tag{A.1.6}$$

and the confluent form

$$\Phi_2^{(n)}\left[\begin{matrix}b_1,\ldots,b_n\\c\end{matrix}\middle|\,x_1,\ldots,x_n\right]=\sum_{\boldsymbol{k}}\frac{b_1^{\overline{k_1}}\ldots b_n^{\overline{k_n}}}{c^{\overline{k_1+\cdots+k_n}}}\frac{x_1^{k_1}\ldots x_n^{k_n}}{k_1!\ldots k_n!}. \tag{A.1.7}$$

The Eulerian integral is,

$$\begin{aligned}&\frac{\Gamma[a]\Gamma[c-a]}{\Gamma[c]}F_D^{(n)}\left[\begin{matrix}a;b_1,\ldots,b_n\\c\end{matrix}\middle|\,x_1,\ldots,x_n\right]\\&=\int_0^1 u^{a-1}(1-u)^{c-a-1}(1-ux_1)^{-b_1}\ldots(1-ux_n)^{-b_n}du,\end{aligned} \tag{A.1.8}$$

where $\Re[a]$ and $\Re[c-a]$ are positive. If we apply one of the $2n+1$ transformations

$$\begin{aligned} u = 1 - v \quad u = v/([1-x_1]+vx_1), \ldots, u = v/([1-x_n]+vx_n), \\ u = (1-v)/(1-vx_1), \ldots, u = (1-v)/(1-vx_n) \end{aligned} \tag{A.1.9}$$

then the following $(2n+1)$ forms of the function $F_D^{(n)}$ arise:

$$\begin{aligned} & F_D^{(n)}\left[\begin{matrix} a; b_1, \ldots, b_n \\ c \end{matrix}\middle| x_1, \ldots, x_n\right] \\ & = (1-x_1)^{-b_1} \ldots (1-x_n)^{-b_n} F_D^{(n)}\left[\begin{matrix} c-a; b_1, \ldots, b_n \\ c \end{matrix}\middle| \frac{x_1}{x_1-1}, \ldots, \frac{x_n}{x_n-1}\right] \\ & = (1-x_1)^{-a} F_D^{(n)}\left[\begin{matrix} a; c-b_1-\cdots-b_n, b_2, \ldots, b_n \\ c \end{matrix}\middle| \frac{x_1}{x_1-1}, \frac{x_1-x_2}{x_1-1}, \ldots, \frac{x_1-x_n}{x_1-1}\right] \\ & = (1-x_1)^{c-a-b_1}(1-x_2)^{-b_2} \ldots (1-x_n)^{-b_n} \\ & \times F_D^{(n)}\left[\begin{matrix} c-a; c-b_1-\cdots-b_n, b_2, \ldots, b_n \\ c \end{matrix}\middle| x_1, \frac{x_2-x_1}{x_2-1}, \ldots, \frac{x_n-x_1}{x_n-1}\right]. \end{aligned} \tag{A.1.10}$$

The simplification,

$$F_D^{(n)}\left[\begin{matrix} a; b_1, \ldots, b_n \\ c \end{matrix}\middle| x, \ldots, x\right] = {}_2F_1\left[\begin{matrix} a, b_1+\cdots+b_n \\ c \end{matrix}\middle| x\right], \tag{A.1.11}$$

also follows from the Eulerian integral. The incomplete integral is $(0 < z < 1)$,

$$\int_0^z u^{a-1}(1-u)^{c-a-1}(1-ux_1)^{-b_1} \ldots (1-ux_n)^{-b_n} du \tag{A.1.12}$$

$$= z^{a-1} F_D^{(n+1)}\left[\begin{matrix} a; b_1, \ldots, b_n, 1+a-c \\ 1+a \end{matrix}\middle| x_1, \ldots, x_n, z\right] \tag{A.1.13}$$

and the double incomplete integral $(0 < z_1 < z_2 < 1)$,

$$\begin{aligned} & \int_{z_1}^{z_2} (x_1 u)^{b_1} \prod_{k=2}^{n} (1-x_b u)^{b_k} \, du \\ & = (z_2-z_1)(x_1 z_1)^{b_1} \prod_{k=2}^{n} (1-x_k z_1)^{b_k} \\ & \times F_D^{(n)}\left[\begin{matrix} 1; -b_1, \ldots, -b_n \\ 2 \end{matrix}\middle| 1-\frac{z_2}{z_1}, 1-\frac{1-x_2 z_2}{1-x_2 z_1} \ldots, 1-\frac{1-x_n z_2}{1-x_n z_1}\right] \\ & = (z_2-z_1)(x_1 z_2)^{b_1} \prod_{k=2}^{n} (1-x_k z_2)^{b_k} \\ & \times F_D^{(n)}\left[\begin{matrix} 1; -b_1, \ldots, -b_n \\ 2 \end{matrix}\middle| 1-\frac{z_1}{z_2}, 1-\frac{1-x_2 z_1}{1-x_2 z_2} \ldots, 1-\frac{1-x_n z_1}{1-x_n z_2}\right]. \end{aligned} \tag{A.1.14}$$

The Lauricella $F_D^{(n)}$ function also has interesting multiple Laplace integral representations,

$$\Gamma[b_1]\dots\Gamma[b_n]F_D^{(n)}\left[\begin{matrix}a;b_1,\dots,b_n\\c\end{matrix}\middle|x_1,\dots,x_n\right] \tag{A.1.15}$$

$$=\int_0^\infty\dots\int_0^\infty e^{-t_1-\dots-t_n}t_1^{b_1-1}\dots t_n^{b_n-1}{}_1F_1\left[\begin{matrix}a\\c\end{matrix}\middle|x_1t_1+\dots+x_nt_n\right]dt_1\dots dt_n, \tag{A.1.16}$$

where $\Re[b_1],\dots,\Re[b_n]>0$, and the single integral representation

$$\Gamma[a]F_D^{(n)}\left[\begin{matrix}a;b_1,\dots,b_n\\c\end{matrix}\middle|x_1,\dots,x_n\right] \tag{A.1.17}$$

$$=\int_0^\infty e^{-t}t^{a-1}\Phi_2^{(n)}\left[\begin{matrix}b_1,\dots,b_n\\c\end{matrix}\middle|x_1t,\dots,x_nt\right]dt, \tag{A.1.18}$$

where $\Re[a]>0$. A Taylor expansion takes the form,

$$\begin{aligned}&F_D^{(n)}\left[\begin{matrix}a;b_1,\dots,b_n\\c\end{matrix}\middle|x_1+y_1,\dots,x_n+y_n\right]\\&=\sum_{j_1,\dots,j_n}\frac{\overline{a_1^{j_1+\dots+j_n}}\overline{b_1^{j_1}}\dots\overline{b_n^{j_n}}}{\overline{c^{j_1+\dots+j_n}}j_1!\dots j_n!}F_D^{(n)}\left[\begin{matrix}a;b_1,\dots,b_n\\c\end{matrix}\middle|x_1,\dots,x_n\right].\end{aligned} \tag{A.1.19}$$

A.2 Applications

To illustrate the use of multivariate hypergeometric functions we examine the 2x2 contingency table of Section 6.3.1 again, but keep one of the the marginals fixed and known. Our 2x2 table was,

	$\mathcal{B}$	$\overline{\mathcal{B}}$
$\mathcal{L}$	n_1	n_2
$\overline{\mathcal{L}}$	n_3	n_4

and say we have some additional information that half of the counts will have the property $\mathcal{L}$. What is the posterior for the correlation? Previously in eq. (6.3.13) with both marginals free we found that our generating function for the joint probability distribution was,

$$\begin{aligned}\Phi\left[\mu,ic\right]&=\int p(\boldsymbol{\rho}|\boldsymbol{n},\mathcal{H})e^{\mu(\rho_1+\rho_2+\rho_3+\rho_4)-ic(\rho_1+\rho_4-\rho_2-\rho_3)}d\boldsymbol{\rho}\\&=(N+1)!\left(\mu-\frac{ic}{2}\right)^{-n_1-1/2}\left(\mu+\frac{ic}{2}\right)^{-n_2-1/2}\left(\mu+\frac{ic}{2}\right)^{-n_3-1/2}\left(\mu-\frac{ic}{2}\right)^{-n_4-1/2}.\end{aligned} \tag{A.2.1}$$

If we now add a fixed marginal ($\rho_A=\rho_1+\rho_2\to\lambda_A$) we find for $\frac{\Phi[\mu,\lambda_A,ic]}{(N+1)!}$,

$$\left(\mu+\lambda_A-\frac{ic}{2}\right)^{-n_1-1/2}\left(\mu+\lambda_A+\frac{ic}{2}\right)^{-n_2-1/2}\left(\mu+\frac{ic}{2}\right)^{-n_3-1/2}\left(\mu-\frac{ic}{2}\right)^{-n_4-1/2}. \tag{A.2.2}$$

To invert this generating function we need the second confluent Humbert function (A.1.5),

$$\Phi_2^{(n)}\left[\begin{matrix} b_1,\ldots,b_n \\ c \end{matrix}\middle| x_1,\ldots,x_n\right] = \sum_{\boldsymbol{k}} \frac{b_1^{\overline{k_1}}\ldots b_n^{\overline{k_n}}}{c^{\overline{k_1+\cdots+k_n}}}\frac{x_1^{k_1}\ldots x_n^{k_n}}{k_1!\ldots k_n!}, \tag{A.2.3}$$

so that we can write,

$$\begin{aligned}&\Phi\left[\mu,\lambda_A,ic\right]\\ &= \int e^{\left(\frac{ic}{2}-\mu\right)\rho}\rho^{N+1}\Phi_2^{(3)}\left[\begin{matrix} n_1+{}^1\!/_2, n_2+{}^1\!/_2, n_3+{}^1\!/_2 \\ N+2 \end{matrix}\middle| -\lambda_A\rho, (-ic-\lambda_A)\rho, -ic\rho\right] d\rho,\end{aligned} \tag{A.2.4}$$

which follows from the Euler integral. Setting $\rho = 1$, we have,

$$\begin{aligned}\Phi\left[\lambda_A,ic\right] &= e^{\frac{ic}{2}}\Phi_2^{(3)}\left[\begin{matrix} n_1+{}^1\!/_2, n_2+{}^1\!/_2, n_3+{}^1\!/_2 \\ N+2 \end{matrix}\middle| -\lambda_A, (-ic-\lambda_A), -ic\right]\\ &= e^{\frac{ic}{2}}\sum_{j_2,j_3}\frac{(n_2+{}^1\!/_2)^{\overline{j_2}}(n_3+{}^1\!/_2)^{\overline{j_3}}}{(N+2)^{\overline{j_2+j_3}}}\frac{(-ic)^{j_2+j_3}}{j_2!j_3!}{}_1F_1\left[\begin{matrix} 1+n_1+n_2+j_2 \\ N+2+j_2+j_3 \end{matrix}\middle| \lambda_A\right],\end{aligned} \tag{A.2.5}$$

where the second line is a Taylor expansion around $c = 0$. We recognise our beta distribution generating function again, which we then use to invert the $\lambda_A \to \rho_A$ transform,

$$\begin{aligned}\Phi\left[p_A,ic\right] &= (N+1)!\frac{\rho_A^{n_1+n_2}}{(n_1+n_2)!}\frac{(1-\rho_A)^{n_3+n_4}}{(n_3+n_4)!}\\ &\times e^{\frac{ic}{2}}{}_1F_1\left[\begin{matrix} n_2+{}^1\!/_2 \\ 1+n_1+n_2 \end{matrix}\middle| -ic\rho_A\right]{}_1F_1\left[\begin{matrix} n_3+{}^1\!/_2 \\ 1+n_3+n_4 \end{matrix}\middle| -ic(1-\rho_A)\right].\end{aligned} \tag{A.2.6}$$

From this we can read off that ρ_A is distributed like a beta distribution,

$$p(\rho_A|\boldsymbol{n},\mathcal{H}) = (N+1)!\frac{\rho_A^{n_1+n_2}}{(n_1+n_2)!}\frac{(1-\rho_A)^{n_3+n_4}}{(n_3+n_4)!}, \tag{A.2.7}$$

which is good to know and a simple consequence of our Dirichlet posterior. We assumed that the value of ρ_A is known and thus we can condition on it using Bayes theorem which results in,

$$\Phi\left[ic \mid p_A\right] = e^{\frac{ic}{2}}{}_1F_1\left[\begin{matrix} {}^1\!/_2+n_2 \\ 1+n_1+n_2 \end{matrix}\middle| -ic\rho_A\right]{}_1F_1\left[\begin{matrix} {}^1\!/_2+n_3 \\ 1+n_3+n_4 \end{matrix}\middle| -ic(1-\rho_A)\right], \tag{A.2.8}$$

So that our correlation parameter is distributed like the convolution of two beta distributions. The convolution of two beta distributions represents a mathematical challenge because the special function representation of the convolution is a piece-wise Appell function which is not the most transparent construction to use, see also Pham-Gia and Turkkan (1998). We want to solve,

$$\begin{aligned}&\left[\frac{(n_1+n_2)!}{\rho_A^{n_1+n_2}}\frac{\left(\frac{\rho_A}{2}+\phi\right)^{n_1-{}^1\!/_2}}{\Gamma[n_1+{}^1\!/_2]}\frac{\left(\frac{\rho_A}{2}-\rho\right)^{n_2-{}^1\!/_2}}{\Gamma[n_2+{}^1\!/_2]}\right]\\ &\otimes\left[\frac{(n_3+n_4)!}{(1-p_A)^{n_3+n_4}}\frac{\left(\frac{1-\rho_A}{2}-\phi\right)^{n_3-{}^1\!/_2}}{\Gamma[n_3+{}^1\!/_2]}\frac{\left(\frac{1-\rho_A}{2}+\phi\right)^{n_4-{}^1\!/_2}}{\Gamma[n_4+{}^1\!/_2]}\right].\end{aligned} \tag{A.2.9}$$

Defining a prefactor,

$$\frac{(n_1+n_2)!(n_3+n_4)!}{\Gamma[n_1+1/2]\Gamma[n_2+1/2]\Gamma[n_3+1/2]\Gamma[n_4+1/2]\rho_A^{n_1+n_2}(1-\rho_A)^{n_3+n_4}}, \tag{A.2.10}$$

the rest of the formulas will be unnormalised. For $-\frac{1}{2} < \phi < -\left|\rho_A - \frac{1}{2}\right|$, we have $-\frac{1-\rho_A}{2} + \phi < -\frac{\rho_A}{2} < \frac{1-\rho_A}{2} + \phi < \frac{\rho_A}{2}$, so that the integral starts at the beginning of the first beta distribution $a = -\frac{1}{2}\rho_A$ and stops at the beginning of the second beta distribution $b = \frac{1-\rho_A}{2} + \phi$. With these limits, the convolution is

$$\begin{aligned}&\int_a^b \left(\frac{\rho_A}{2}+p\right)^{n_1-1/2}\left(\frac{\rho_A}{2}-p\right)^{n_2-1/2}\\ &\times\left(\frac{1-\rho_A}{2}-\phi+p\right)^{n_3-1/2}\left(\frac{1-\rho_A}{2}+\phi-p\right)^{n_4-1/2}dp,\end{aligned} \tag{A.2.11}$$

and making the transformation $v = \frac{b-p}{b-a}$

$$\begin{aligned}&(b-a)^{n_1+n_4}(-a-b)^{n_2-1/2}(2b-2\phi)^{n_3-1/2}\\ &\times\int_0^1 (1-v)^{n_1-1/2}\,v^{n_4-1/2}\left[1-\frac{b-a}{a+b}v\right]^{n_2-1/2}\left[1-\frac{b-a}{2b-2\phi}v\right]^{n_3-1/2}dv\end{aligned} \tag{A.2.12}$$

Using Picard's result Eq. (A.1.3), we find an Appell function,

$$\begin{aligned}&\frac{\Gamma[n_1+1/2]\Gamma[n_4+1/2]}{\Gamma[n_1+n_4+1]}\left(\tfrac{1}{2}+\phi\right)^{n_1+n_4}\left(\rho_A-\tfrac{1}{2}-\phi\right)^{n_2-1/2}(1-\rho_A)^{n_3-1/2}\\ &\times F_1\left[\begin{matrix}n_4+1/2,-n_2+1/2,-n_3+1/2\\ n_1+n_4+1\end{matrix}\middle|\,-\frac{\frac{1}{2}+\phi}{\rho_A-\frac{1}{2}-\phi},\frac{\frac{1}{2}+\phi}{1-\rho_A}\right].\end{aligned} \tag{A.2.13}$$

Similarly for $\left|\frac{1}{2}-\rho_A\right| < \phi < \frac{1}{2}$ the integral starts at the end of the second beta distribution $a = \phi - \frac{1-\rho_A}{2}$ and ends at the end of the first beta distribution $b = \frac{\rho_A}{2}$, using the transformation $v = \frac{\rho-a}{b-a}$,

$$\begin{aligned}&\left(\tfrac{1}{2}-\phi\right)^{N-1}\int_0^1\left(\frac{\rho_A}{\frac{1}{2}-\phi}-v\right)^{n_1-1/2}v^{n_2-1/2}(1-v)^{n_3-1/2}\left(\frac{\phi+\frac{1}{2}-\rho_A}{\frac{1}{2}-\phi}+v\right)^{n_4-1/2}dp\\ &=\frac{\Gamma[n_2+1/2]\Gamma[n_3+1/2]\left(\frac{1}{2}-\phi\right)^{n_2+n_3}\left(\phi+\frac{1}{2}-\rho_A\right)^{n_4-1/2}\rho_A^{n_1-1/2}}{\Gamma[n_2+n_3+1]}\\ &\times F_1\left[\begin{matrix}n_2+1/2,-n_1+1/2,-n_4+1/2\\ n_2+n_3+1\end{matrix}\middle|\,\frac{\frac{1}{2}-\phi}{\rho_A},-\frac{\frac{1}{2}-\phi}{\phi+\frac{1}{2}-\rho_A}\right].\end{aligned} \tag{A.2.14}$$

Finally, for $-\left|\frac{1}{2}-\rho_A\right| < \phi < \left|\frac{1}{2}-\rho_A\right|$ and if $\rho_A < \frac{1}{2}$, $a = -\frac{\rho_A}{2}$ and $b = \frac{\rho_A}{2}$. Using $v = \frac{p-a}{b-a}$,

$$\begin{aligned}
&\rho_A^{N-1} \int_0^1 v^{n_1-1/2}\,(1-v)^{n_2-1/2} \left(\frac{\frac{1}{2}-\phi}{\rho_A} - (1-v)\right)^{n_3-1/2} \left(\frac{\frac{1}{2}+\phi}{\rho_A} - v\right)^{n_4-1/2} dp \\
&= \frac{\Gamma[n_1+1/2]\Gamma[n_2+1/2]\rho_A^{n_1+n_2}\left(\frac{1}{2}-\phi-\rho_A\right)^{n_3-1/2}\left(\frac{1}{2}+\phi\right)^{n_4-1/2}}{\Gamma[n_1+n_2+1]} \qquad \text{(A.2.15)} \\
&\times F_1\left[\begin{matrix} n_1+1/2, -n_3+1/2, -n_4+1/2 \\ n_1+n_2+1 \end{matrix}\middle| -\frac{\rho_A}{\frac{1}{2}-\phi-\rho_A}, \frac{\rho_A}{\frac{1}{2}+\phi}\right].
\end{aligned}$$

If $\rho_A > \frac{1}{2}$, $a = \phi - \frac{1-\rho_A}{2}$ and $b = \phi + \frac{1-\rho_A}{2}$ and we use $v = \frac{p-a}{b-a}$,

$$\begin{aligned}
&(1-\rho_A)^{N-1} \int_0^1 \left(\frac{\phi-\frac{1}{2}+\rho_A}{1-\rho_A} + v\right)^{n_1-1/2} \left(\frac{\frac{1}{2}-\phi}{1-\rho_A} - v\right)^{n_2-1/2} v^{n_3-1/2}\,(1-v)^{n_4-1/2}\,dp \\
&= \frac{\Gamma[n_3+1/2]\Gamma[n_4+1/2]\,(1-\rho_A)^{n_3+n_4}\left(\phi-\frac{1}{2}+\rho_A\right)^{n_1-1/2}\left(\frac{1}{2}-\phi\right)^{n_2-1/2}}{\Gamma[n_3+n_4+1]} \\
&\times F_1\left[\begin{matrix} n_3+1/2, -n_1+1/2, -n_2+1/2 \\ n_3+n_4+1 \end{matrix}\middle| -\frac{1-\rho_A}{\phi-\frac{1}{2}+\rho_A}, \frac{1-\rho_A}{\frac{1}{2}-\phi}\right].
\end{aligned} \qquad \text{(A.2.16)}$$

This completes the example of section 6.3.1 namely it gives the posterior of a correlation variable with the marginal probabilities fixed analytically.

Appendix B

Cox's theorems

Following Cox (1961), we will first consider some desiderata that we require for a calculus of probable inference. Commentary on the derivation can be found in Jaynes (2003), Paris (1994), Dupré and Tipler (2009), van Horn (2003) and Halpern (1999), while the functional equation theory can be found in Aczél (1966) and Aczél (2003).

B.1 Cox's properties

A probable inference, as in common usage, is a partial agreement on the evidence. Everyone agrees more fully on some inferences than others. Hence it is natural to suppose that plausibility can be ordered, which leads us to the first Cox property,

> If $\mathcal{A}$ is believed more than $\mathcal{B}$, and $\mathcal{B}$ is believed more than $\mathcal{C}$, then $\mathcal{A}$ is believed more than $\mathcal{C}$.

For the second property, consider the plausibility of an assertion made by a famous author that Noah's ark can still be seen on a clear day, resting where it was left by the receding waters of the Flood, on the top of Mount Ararat. For this statement to be plausible it must be based on the memory of the author and not his imagination. Then assuming it was made from memory, to be plausible his memory must still be trustworthy after many years. Finally, assuming his recount his truthful and his memory sound it must be plausible that he or those he depended on could be sure that they had truly seen Noah's Ark. This shows that any assertion can be broken in a chain of propositions and the second Cox property is,

> The plausibility on given evidence that both inferences are true is determined by their separate plausibilities, one on the given evidence and the other on this evidence with the additional assumption that the first inference was true.

The third property is simple: if an argument makes a certain inference more plausible then it makes the contradictory inference less plausible, thus

> `The plausibility of an inference on given evidence determines the plausibility of its contradictory on the same evidence.`

The three properties together we will call Cox's properties and we will demand them from our calculus of probable inference.

B.2 The Algebra of propositions

Ordinary algebra is the algebra of quantities. Boolean algebra on the other hand is the calculus of propositions which we will denote by calligraphy letters $\mathcal{A}, \mathcal{B}, \ldots$. $\overline{\mathcal{A}}$ is read as `NOT` $\mathcal{A}$ and is the denial of it. Obviously, the denial of $\overline{\mathcal{A}}$ is to affirm $\mathcal{A}$,

$$\overline{\overline{\mathcal{A}}} = \mathcal{A}. \tag{B.2.1}$$

The proposition $\mathcal{A}$ `AND` $\mathcal{B}$ is called the *conjunction* of $\mathcal{A}$ and $\mathcal{B}$ and as in normal speech the ordering is unimportant,

$$(\mathcal{A}, \mathcal{B}) = (\mathcal{B}, \mathcal{A}) \tag{B.2.2}$$

Similarly the expression $(\mathcal{A}, \mathcal{A})$ implies that we have stated proposition $\mathcal{A}$ twice and not that it has occurred twice thus,

$$(\mathcal{A}, \mathcal{A}) = \mathcal{A}. \tag{B.2.3}$$

Logically the proposition $((\mathcal{A}, \mathcal{B}), \mathcal{C})$ is the same as $(\mathcal{A}, (\mathcal{B}, \mathcal{C}))$ so we can omit the parentheses,

$$((\mathcal{A}, \mathcal{B}), \mathcal{C}) = (\mathcal{A}, (\mathcal{B}, \mathcal{C})) = (\mathcal{A}, \mathcal{B}, \mathcal{C}). \tag{B.2.4}$$

The proposition $\mathcal{A}$ `OR` $\mathcal{B}$ is called the disjunction. Combining it with the conjunction leads to

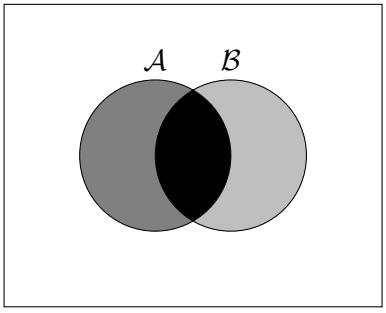


Figure B.1: Venn diagram for $\mathcal{A}$ and $\mathcal{B}$.

De Morgan's law, which we will argue from a Venn diagram: All the results on which proposition $\mathcal{A}$ is true is marked black and gray, all the results on which proposition $\mathcal{B}$ is true is shaded black and light gray and all results on which both propositions are true is shaded black. Results for which both propositions are false are marked white. Then the conjunction of $\overline{\mathcal{A}}$ and $\overline{\mathcal{B}}$ is the white and light gray area combined with the white and gray area which is the white and both gray areas. This is equivalent to all the non-black areas which is $\overline{\mathcal{A}, \mathcal{B}}$ and gives De Morgan's law,

$$\overline{(\mathcal{A}, \mathcal{B})} = \overline{\mathcal{A}} + \overline{\mathcal{B}}. \tag{B.2.5}$$

More complicated structures can also arise; for example: when we assert $(\mathcal{A}+\mathcal{B},\mathcal{C})$ either $\mathcal{A}$ or $\mathcal{B}$ is true, but $\mathcal{C}$ is true in any case, which is the same as saying either $(\mathcal{A},\mathcal{C})$ or $(\mathcal{B},\mathcal{C})$ is true,

$$(\mathcal{A}+\mathcal{B},\mathcal{C}) = (\mathcal{A},\mathcal{C}) + (\mathcal{B},\mathcal{C}). \tag{B.2.6}$$

We can also form invariants in Boolean algebra, namely the truism $\mathcal{A}+\overline{\mathcal{A}}$ and the absurdity $(\mathcal{A},\overline{\mathcal{A}})$, which are mutually contradictory. The invariants are the same no matter the content of the proposition $\mathcal{A}$. In summary we have the collection of formulas,

$$\begin{aligned}
\overline{\overline{\mathcal{A}}} &= \mathcal{A}, & & \\
(\mathcal{A},\mathcal{A}) &= \mathcal{A} & \mathcal{A}+\mathcal{A} &= \mathcal{A}, \\
(\mathcal{A},\mathcal{B}) &= (\mathcal{B},\mathcal{A}) & \mathcal{A}+\mathcal{B} &= \mathcal{B}+\mathcal{A}, \\
\overline{(\mathcal{A},\mathcal{B})} &= \overline{\mathcal{A}}+\overline{\mathcal{B}} & \overline{(\mathcal{A}+\mathcal{B})} &= (\overline{\mathcal{A}},\overline{\mathcal{B}}) \\
((\mathcal{A},\mathcal{B}),\mathcal{C}) &= (\mathcal{A},(\mathcal{B},\mathcal{C})) & ((\mathcal{A}+\mathcal{B})+\mathcal{C}) &= (\mathcal{A}+(\mathcal{B}+\mathcal{C})) \\
(\mathcal{A}+\mathcal{B},\mathcal{C}) &= (\mathcal{A},\mathcal{C})+(\mathcal{B},\mathcal{C}) & (\mathcal{A},\mathcal{B})+\mathcal{C} &= (\mathcal{A}+\mathcal{C},\mathcal{B}+\mathcal{C}) \\
(\mathcal{A}+\mathcal{B},\mathcal{B}) &= \mathcal{B} & (\mathcal{A},\mathcal{B})+\mathcal{B} &= \mathcal{B} \\
(\mathcal{A}+\overline{\mathcal{A}},\mathcal{B}) &= \mathcal{B} & (\mathcal{A},\overline{\mathcal{A}})+\mathcal{B} &= \mathcal{B} \\
(\mathcal{A}+\overline{\mathcal{A}},\mathcal{B}) &= \mathcal{B} & (\mathcal{A},\overline{\mathcal{A}})+\mathcal{B} &= \mathcal{B} \\
\mathcal{A}+\overline{\mathcal{A}}+\mathcal{B} &= \mathcal{A}+\overline{\mathcal{A}} & (\mathcal{A},\overline{\mathcal{A}},\mathcal{B}) &= (\mathcal{A},\overline{\mathcal{A}}),
\end{aligned} \tag{B.2.7}$$

which form the basic equations for Boolean algebra.

B.3 Desiderata

Cox's properties leads to certain desiderata or things that we desire from our axioms. From the first property we require **Transitivity** and **Universal Comparability** and from this it follows that plausibility can be represented by rational numbers, for which we will write,

> $\pi(\mathcal{A}|\mathcal{X}) \equiv$ `A rational number assigned to the logical proposition` $\mathcal{A}$ `given the evidence` $\mathcal{X}$`.`

To see that this follows, we note that transitivity asserts that if $A \geq B$ and $B \geq C$, $A \geq C$ follows and universal comparability asserts that we can compare any two propositions i.e. we believe one of them is more certain or we are equally certain of them. Next note that any finite set of propositions $\{A_1,\ldots,A_n\}$ has a specific ordering that can be represented by rational numbers. And if we add a new proposition A_{n+1}, transitivity and universal comparability ensures that the proposition will have a unique place in the ordering. The new proposition can still be represented by a rational number because we can always find a rational number between any two rational numbers, thus always preserving our representation. The rational number representation has massive computational benefits

as well and it would be foolish to deny ourself this advantage. Also as a convention we shall assume greater certainty will be represented with a greater number and lesser certainty with a lesser number, we can also reverse this order but we do not gain anything by doing this, so we make the natural choice. Hence we formulate

> `Desideratum I: Degrees of belief in logical propositions are represented by rational numbers. As a convention greater certainty will correspond to a larger number.`

For the second Cox property we have to remember that it is the ordering of the propositions that is important, not the actual numerical value; thus there is no natural scale. The consequence of using the rational numbers is that our learning rule must be homogeneous i.e. multiplying all the numbers with a fixed positive number does not change the ordering or our conclusions. The learning rule must also be commutative because Boolean algebra is commutative. Hence follows

> `Desideratum II: We desire a homogeneous, commutative learning rule that takes` $\pi(\mathcal{A}|\mathcal{X})$ `and` $\pi(\mathcal{A}|\mathcal{B},\mathcal{X})$ `to give` $\pi(\mathcal{A},\mathcal{B}|\mathcal{X})$.

The third desiderata follows directly from the third cox property,

> `Desideratum III: We desire a function that maps plausibilities onto themselves that corresponds to the logical NOT operation.`

B.4 Axioms

Cox's properties are intuitive principles that we seek in a calculus of plausible reasoning, which translates into three mathematical desiderata for the axioms of our calculus. Now we will show that the desiderata are strong enough to specify our axioms.

Let us investigate the functional equation for the learning rule,

$$F[\pi(\mathcal{A}|\mathcal{X}),\pi(\mathcal{B}|\mathcal{A},\mathcal{X})]=\pi(\mathcal{A},\mathcal{B}|\mathcal{X}) \tag{B.4.1}$$

and change the background information slightly $\mathcal{X}\to\mathcal{X}'$,

$$F[\pi(\mathcal{A}|\mathcal{X}'),\pi(\mathcal{A}|\mathcal{B},\mathcal{X}')]=\pi(\mathcal{A},\mathcal{B}|\mathcal{X}'). \tag{B.4.2}$$

Since our rational number representation is scaled by a fixed positive number a, of course our homogeneity property, yields

$$F[a\pi(\mathcal{A}|\mathcal{X}),a\pi(\mathcal{A}|\mathcal{B},\mathcal{X})]=a^kF[\pi(\mathcal{A}|\mathcal{X}),\pi(\mathcal{A}|\mathcal{B},\mathcal{X})]. \tag{B.4.3}$$

Writing $x=\pi(\mathcal{A}|\mathcal{X})$ and $y=\pi(\mathcal{A}|\mathcal{B},\mathcal{X})$ we have,

$$F[ax,ay]=a^kF[x,y] \qquad\qquad a>0,k>0. \tag{B.4.4}$$

According to $(B.4.4)$,

$$F[x,y] = F\left[x\cdot 1, x\cdot\frac{y}{x}\right] = x^k F\left[1,\frac{y}{x}\right] \tag{B.4.5}$$

and the functions,

$$F[x,y] = x^k f\left(\frac{y}{x}\right) \qquad\qquad x\neq 0 \tag{B.4.6}$$

do in fact satisfy $(B.4.4)$. Since this approach does not make sense for $x\neq 0$, we consider

$$F[0,y] = F[0\cdot y, 1\cdot y] = y^k F[0,1] \qquad\qquad y\neq 0, \tag{B.4.7}$$

and (with $x=y=0$ and $k\neq 0$)

$$F[0,0] = 0. \tag{B.4.8}$$

Combining the equations we have for $k\neq 0$,

$$F[x,y] = \begin{cases} x^k f\left(\frac{y}{x}\right) & (x\neq 0) \\ y^k c & (x=0, y\neq 0) \\ 0 & (x=y=0), \end{cases} \tag{B.4.9}$$

where c is an arbitrary constant and f is an arbitrary function of one variable. Clearly zero is an invariant of our homogeneous learning rule and the other invariant we label e. These two invariants must correspond to the invariants of Boolean algebra namely the truism and the absurd. As a convention we will choose zero to be the absurd and e as the truism thus we have from Boolean algebra for $x\neq 0$,

$$\begin{aligned} F[x,e] &= F[e,x] = x \\ &= x^k f\left[\frac{e}{x}\right] = e^k f\left[\frac{x}{e}\right], \end{aligned} \tag{B.4.10}$$

implying that

$$f[x] = \left(\frac{x}{e}\right)^{k-1} \quad \text{and} \quad f[x] = \frac{x}{e^{k-1}}, \tag{B.4.11}$$

which can only be true if $k=2$. Hence follows

`Axiom I: The only commutative, homogeneous learning rule is multiplication on the interval [0,e],`

$$\pi(\mathcal{A}|\mathcal{X})\pi(\mathcal{B}|\mathcal{A},\mathcal{X}) = \pi(\mathcal{A},\mathcal{B}|\mathcal{X}). \tag{B.4.12}$$

`As a convention we assign one to the propositions that are truisms and zero to propositions we know are absurd,`

$$\pi(\mathcal{X}|\mathcal{X}) = 1 \qquad\qquad \pi(\overline{\mathcal{X}}|\mathcal{X}) = 0. \tag{B.4.13}$$

The third Cox's property seeks a function that negates itself, thus

$$S\left[S\left[\pi(\mathcal{A}|\mathcal{X})\right]\right] = \pi(\mathcal{A}|\mathcal{X}). \tag{B.4.14}$$

and

$$S\left[\pi(\mathcal{A}|\mathcal{X})\right] = \pi(\overline{\mathcal{A}}|\mathcal{X}). \tag{B.4.15}$$

from which we have $S(0) = 1$ and $S(1) = 0$ as well. S is also strictly decreasing. Suppose that it is not: take $\gamma < \delta$, but insist that $S(\lambda) = S(\delta)$. Applying S again we have, $\lambda = \delta$ which is a contradiction. It also follows that S is continuous because it is strictly decreasing and maps the whole interval $(0, 1)$ onto $(0, 1)$, implying there are no gaps and no gap discontinuities. Since S is continuous and $S(0) > 0, S(1) < 1$ we can pick $0 < \nu < 1$ to give $S(\nu) = \nu$. Our second convention is that for probabilities we have,

$$S\left[{}^{1}\!/_{2}\right] = {}^{1}\!/_{2}, \tag{B.4.16}$$

thus

$$p(\mathcal{A}|\mathcal{X}) = \pi(\mathcal{A}|\mathcal{X})^{k}, \tag{B.4.17}$$

where k is chosen to ensure (B.4.16). Notice that it preserves multiplication as our learning rule. The difference between plausibilities and probabilities is that probabilities has ${}^{1}\!/_{2}$ as a fixed point for the negation operation S.

Consider the following $y = p(Y|Z)$ and $\frac{x}{y} = p(X|Y, Z)$,

$$\begin{aligned} yS\left(\frac{x}{y}\right) &= p(Y|Z)S[p(X|Y,Z)] = p(Y|Z)p(\overline{X}|Y,Z) \\ &= p(\overline{X}, Y|Z) \end{aligned} \tag{B.4.18}$$

whilst

$$\begin{aligned} S[y] &= p(\overline{Y}|Z) = p((\overline{X} + \overline{Y}, X + \overline{Y})|Z), \\ S[x] &= p(\overline{X} + \overline{Y}|Z). \end{aligned} \tag{B.4.19}$$

Examining,

$$\begin{aligned} S[x]S\left[\frac{S[y]}{S[x]}\right] &= S[x]S[p(\overline{X} + Y|\overline{X} + \overline{Y}, Z)] \\ &= p(X, \overline{Y}, \overline{X} + \overline{Y}|Z) \\ &= p(X, \overline{Y}|Z) \end{aligned} \tag{B.4.20}$$

and combining we have,

$$yS\left[\frac{x}{y}\right] = S[x]S\left[\frac{S[y]}{S[x]}\right] \qquad 0 < x \leq y \leq 1, \tag{B.4.21}$$

which is our functional equation for S.

We construct a operation $\circ$ which is commutative,

$$\begin{aligned} u \circ v &= S\left[S[u]S\left[\frac{v}{S[u]}\right]\right] \\ &= S\left[S[v]S\left[\frac{u}{S[v]}\right]\right] \\ &= v \circ u \end{aligned} \tag{B.4.22}$$

by using our functional (B.4.21) and where $0 < v \leq 1$ and $0 < u \leq S[v]$. Applying our functional (B.4.21) to this construction,

$$\begin{aligned} (u \circ v)S\left[\frac{u}{u \circ v}\right] &= S[u]S\left[\frac{S[u \circ v]}{S[u]}\right] \\ &= S[u]\frac{v}{S[u]} \\ &= v. \end{aligned} \tag{B.4.23}$$

So in some sense the $\circ$ operator is the inverse of our functional equation (B.4.21). Reversing the order of the operations we see

$$\begin{aligned} \left[uS\left[\frac{v}{u}\right]\right] \circ v &= S\left[S[v]S\left[\frac{uS\left[\frac{v}{u}\right]}{S[v]}\right]\right] \\ &= S\left[S[v]\frac{S[u]}{S[v]}\right] \\ &= u, \end{aligned} \tag{B.4.24}$$

the same behaviour. On the left hand side of our associativity equation, we apply our functional (B.4.21),

$$u \circ (v \circ w) = S\left[S[u]S\left[\frac{S\left[S[v]S\left[\frac{w}{S[v]}\right]\right]}{S[u]}\right]\right] \quad = S\left[S[v]S\left[\frac{w}{S[v]}\right]S\left[\frac{u}{S[v]S\left[\frac{w}{S[v]}\right]}\right]\right], \tag{B.4.25}$$

and comparing it to the right hand side of our associativity equation,

$$\begin{aligned} (u \circ v) \circ w &= S\left[S[v]S\left(\frac{u}{S[v]}\right)S\left[\frac{w}{S[v]S\left(\frac{u}{S[v]}\right)}\right]\right] \\ &= S\left[S[v]S\left(\frac{w}{S[v]}\right)S\left[\frac{u}{S[v]S\left(\frac{w}{S[v]}\right)}\right]\right], \end{aligned} \tag{B.4.26}$$

we see they are equal and that the operator $\circ$ is also associative. Expanding au using B.4.23

$$\begin{aligned} au &= a(u \circ v)S\left[\frac{v}{u \circ v}\right] \\ &= a(u \circ v)S\left[\frac{av}{a(u \circ v)}\right], \end{aligned} \tag{B.4.27}$$

and then applying the operation $\circ av$, we have

$$\begin{aligned} au \circ av &= \left[a(u \circ v) S\left[\frac{av}{a(u \circ v)}\right]\right] \circ av \\ &= a(u \circ v), \end{aligned} \tag{B.4.28}$$

according to (B.4.24) and $\circ$ is also distributive. Knowing $S[1/2] = 1/2$ we can show $1/2 \circ 1/2 = 1$. Using induction and the distributive property, we have for $n > 0$,

$$\frac{1}{2^n} \circ \frac{1}{2^n} = \frac{1}{2^{n-1}}. \tag{B.4.29}$$

Let $\circ^m \left(\frac{1}{2^n}\right)$ stand for $\frac{1}{2^n} \circ \frac{1}{2^n} \circ \cdots \circ \frac{1}{2^n}$, m times. Since $\circ^m \left(\frac{1}{2^n}\right) = 1$ if $m = 2^n$ by induction this notation is well defined. Suppose $\circ^m \left(\frac{1}{2^n}\right) < \frac{m}{2^n}$, choose a number between them $\frac{1}{2^{p/q}}$,

$$\circ^m \left(\frac{1}{2^n}\right) < \frac{1}{2^{p/q}} < \frac{m}{2^n}, \tag{B.4.30}$$

and take the qth power,

$$\left(\circ^{m^q}\right) \frac{1}{2^{nq}} < \frac{1}{2^p} < \frac{m^q}{2^{qn}}. \tag{B.4.31}$$

The last part of the inequality gives $m^q > 2^{nq-p}$, resulting in,

$$\left(\circ^{m^q}\right) \frac{1}{2^{nq}} > \left(\circ^{2^{nq-p}}\right) \left(\frac{1}{2^{nq-p}}\right) = \frac{2^{nq-p}}{2^{nq}} = \frac{1}{2^p}. \tag{B.4.32}$$

Contradicting ourselves, repeating the derivation for $<$ follows the same manner, thus we have,

$$\circ^m \left(\frac{1}{2^n}\right) = \frac{m}{2^n}. \tag{B.4.33}$$

Finally choosing $u = \frac{m_1}{2^n}$ and $v = \frac{m_2}{2^n}$,

$$\begin{aligned} (u \circ v) &= \frac{m_1}{2^n} \circ \frac{m_2}{2^n} \\ &= \circ^{m_1} \frac{1}{2^n} \circ \circ^{m_2} \frac{1}{2^n} = \circ^{m_1+m_2} \frac{1}{2^n} \\ &= \frac{m_1 + m_2}{2^n} = u + v. \end{aligned} \tag{B.4.34}$$

Using the negation identity we have,

$$(u + v) S\left(\frac{u}{u + v}\right) = v = u + v - u, \tag{B.4.35}$$

and solving for S,

$$S(x) = 1 - x. \tag{B.4.36}$$

`Axiom II: All operators that correspond to the logical NOT operation can be transformed to the probability base in which the operator takes the form,`

$$p(\overline{A}|X) = 1 - p(A|X) \tag{B.4.37}$$

Bibliography

Achard, P., Adriani, O., Aguilar-Benitez, M., Alcaraz, J., Alemanni, G., Allaby, J., Aloisio, A., Alviggi, M., Anderhub, H., Andreev, V. *et al.* (2011). Test of the τ-Model of Bose-Einstein Correlations and Reconstruction of the Source Function in Hadronic Z-boson Decay at LEP. *The European Physical Journal C*, vol. 71, no. 5, pp. 1–25.

Aczél, J. (1966). *Lectures on Functional Equations and Their Applications*. Academic Press.

Aczél, J. (2003). The Associativity Equation Re-Revisited. In: Erikson, G. and Zhai, Y. (eds.), *Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 23rd International Workshop*. American Institute of Physics.

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, vol. 19, no. 6.

Appell, P. (1880). Sur les séries hypergéométriques de deux variables et sur des équations différentielles linéaires aux dérivées partielles. *Comptes renduc hebdomadaires des séances de l'Académie des sciences*, vol. 90, pp. 296–298, 731–735.

Band, W. (1955). *An Introduction to Quantum Statistics*. D Van Nostrand.

Bayes, T. (1763). An essay toward solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, vol. 53, pp. 370–418.

Bender, C.M. and Orszag, S.A. (1999). *Advanced Mathematical Methods for Scientists and Engineers I: Asymptotic Methods and Perturbation Theory*. Advanced Mathematical Methods for Scientists and Engineers. Springer.

Bernoulli, J. (1713). *Ars Conjectandi*. Thurnisiorum.

Bleistein, N. and Handelsman, R.A. (1986). *Asymptotic Expansions of Integrals*. Dover edn. Dover Publications.

Boltzmann, L. (1974). *Theoretical Physics and Philosophical Problems: Selected Writings*. Vienna Circle collection. Kluwer Academic. Editor B McGuinness.

Constantini, D. (1987). Symmetry and the indistinguishability of classical particles. *Physics Letters A*, vol. 123, no. 9, pp. 433–436.

Cover, T.M. and Thomas, J.A. (2012). *Elements of Information Theory*. Wiley.

Cox, R.T. (1946). Probability, Frequency and Reasonable Expectation. *American Journal of Physics*, vol. 14, p. 1.

Cox, R.T. (1961). *The Algebra of Probable Inference*. The Johns Hopkin Press.

d'Agostini, G. (2003). *Bayesian Reasoning in Data Analysis: A Critical Introduction*. World Scientific.

Daniels, H.E. (1954). Saddlepoint approximations in statistics. *Annals of Mathematical Statistics*, vol. 25, p. 631.

Darwin, C.G. and Fowler, R.H. (1922*a*). LXXI. On the partition of energy.Part II. Statistical principles and thermodynamics. *Philosophical Magazine Series 6*, vol. 44, no. 263, pp. 823–842.

Darwin, C.G. and Fowler, R.H. (1922*b*). XLIV. On the partition of energy. *Philosophical Magazine Series 6*, vol. 44, no. 261, pp. 450–479.

De Finetti, B. (1974). *Theory of Probability: a critical introductury treatment*. illustrated, reprint edn. Wiley.

de Kock, M.B. (2009). *Gaussian and non-Gaussian-based Gram-Charlier and Edgeworth expansions for correlations of identical particles in HBT interferometry*. Master's thesis, Stellenbosch: University of Stellenbosch.

de Kock, M.B., Eggers, H.C. and Csörgő, T. (2011*a*). From χ^2 to Bayesian model comparison and L'evy expansions of Bose-Einstein correlations in e^+e^- reactions. *Proceedings of Science*, vol. PoS (WPCF2011) 033.

de Kock, M.B., Eggers, H.C. and Schmiegel, J. (2011*b*). Determining source cumulants in femtoscopy with Gram-Charlier and Edgeworth series. *Modern Physics Letters A*, vol. 26, pp. 1771–1782.

de Kock, M.B., Eggers, H.C. and Schmiegel, J. (2011*c*). Edgeworth versus Gram-Charlier Series: x-cumulant and probability density tests. *Physics of Particles and Nuclei Letters*, vol. 8, pp. 1023–1027.

de Moivre, A. (1718). *The Doctrine of Chances or, A Method of Calculating the Probability of Events in Play*. W. Pearson.

Diaconis, P. (1977). Finite Forms of De Finetti's theorem on Exchangeability. *Synthese*, vol. 37, pp. 271–281.

Diaconis, P. and Ylvisaker, D. (1979). Conjugate Priors for Exponential Families. *The Annals of Statistics*, vol. 7, no. 2, pp. 269–281.

Dupré, M.J. and Tipler, F.J. (2009). New axioms for rigorous bayesian probability. *Bayesian Analysis*, vol. 4, no. 3, pp. 599–606.

Edgeworth, F.Y. (1884). Philosophy of chance. *Mind*, vol. 9, pp. 222–235.

Exton, H. (1976). *Multiple Hypergeometric Functions and Applications*. Ellis Horwood.

Feller, W. (1974). *Introduction to Probability Theory and Its Applications*. No. Vol 2 in Wiley Mathematical Statistics Series. Wiley.

Fisher, R.A. (1922*a*). On the interpretation of χ^2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, vol. 85, no. 1, pp. 87–94.

Fisher, R.A. (1922*b*). On the Mathematical Foundations of Theroretical Statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 222, pp. 309–368.

Graham, R.L., Knuth, D.E. and Patashnik, O. (1994). *Concrete Mathematics: A Foundation for Computer Science*. 2nd edn. Addison-Wesley Publishing.

Grandy, W.T. (1987). *Foundations of Statistical Mechanics: Volume I: Equilibrium Theory*. Astrophysics and Space Science Library. Springer.

Greenwood, M. and Yule, G.U. (1920). An Inquiry into the Nature of Frequency Distributions Representative of Multiple Happenings with Particular Reference to the Occurrence of Multiple Attacks of Disease or of Repeated Accidents. *Journal of the Royal Statistical Society*, vol. 83, no. 2, pp. pp. 255–279.

Gull, S.F. (1989). Bayesian Data Analysis: Straight Line Fitting. In: Skilling, J. (ed.), *Maximum Entropy and Bayesian Methods*. Kluwer Academic.

Gull, S.F. and Skilling, J. (1984). Maximum entropy method in image processing. *Communications, Radar and Signal Processing, IEE Proceedings F*, vol. 131, pp. 646–659.

Hald, A. (1960). The compound hypergeometric distribution and a system of single sampling inspection plans based on prior distributions and costs. *Technometrics*, vol. 2, no. 3, pp. 275–340.

Haldane, J.B.S. (1931). A note on inverse probability. *Mathematical Proceedings of Cambridge Philosophical Society*, vol. 28, pp. 55–61.

Halpern, J.Y. (1999). A Counterexample to the Theorems of Cox and Fine. *Journal of Artificial Intelligence Research*, vol. 10, pp. 67–85.

Heath, D. and Sudderth, W. (1976). De Finetti's Theorem on Exchangeable Variables. *The American Statistician*, vol. 30, no. 4.

Humbert, P. (1921). *The Confluent Hypergeometric Functions of Two Variables*. Robert Grant & Son.

James, F.E. (2006). *Statistical Methods in Experimental Physics*. 2nd edn. World Scientific.

Jaynes, E.T. (1968). Prior probabilities. *IEEE Transactions On Systems Science and Cybernetics*, vol. 4, no. 3, pp. 227–241.

Jaynes, E.T. (1982). Some Applications and Extensions of the De Finetti Representation Theorem. In: Goel, P.K. and Zellner, A. (eds.), *Bayesian Inference and Decision Techniques with Applications: Essays in Honor of Bruno de Finetti*. North-Holland Publishers.

Jaynes, E.T. (1986). Monkeys, Kangaroos, and N. In: Justice, J.H. (ed.), *Maximum Entropy and Bayesian Methos in Applied Statistics*. Cambridge University Press.

Jaynes, E.T. (1990). Straight Line Fitting - A Bayesian Solution. In: Grandy, W.T. and Schick, L. (eds.), *Tenth Annual MAXENT Worshop Proceedings*. Kluwer Academic.

Jaynes, E.T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press.

Jeffreys, H. (1961). *Theory of Probability*. 3rd edn. Claredon Press.

Jeffreys, H. and Jeffreys, B.S. (1950). *Methods of Mathematical Physics*. Cambridge University Press.

Johnson, N.L., Kemp, A.W. and Kotz, S. (2005). *Univariate Discrete Distributions*. 3rd edn. John Wiley & Sons.

Johnson, W.E. (1932). Probability: The Deductive and Inductive Problems. *Mind*, vol. 41, no. 164, pp. 409–423. New Series.

Kass, R.E. and Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 773–795.

Kendall, M.G. and Moran, P.A.P. (1963). *Geometrical Probability*. Charles Griffin AND Company.

Keynes, J.M. (1921). *A Treatise on Probability*. Macmillan.

Khinchin, A.I.A. (1949). *Mathematical Foundations of Statistical Mechanics*. Dover Publications.

Khinchin, A.Y. (1960). *Mathematical Foundations of Quantum Statistics*. Graylock Press.

Laplace, P.S. (1774). Memoir on the probability of the causes of events. *Statistical Science*, vol. 1, no. 3, pp. pp. 364–378.

Laplace, P.S. (1812). *Theéorie Analytique des Probabilités*, vol. 2. 3rd edn. Courcier Imprimeur.

Lauricella, G. (1893). Sulle funzioni ipergeometriche a piu variabili. *Rendiconti Del Circolo Matematico Di Palermo*, vol. 7, pp. 111–158.

Lindley, D.V. (2006). *Understanding Uncertainty*. John Wiley & Sons.

MacKay, D.J.C. (1998). Choice of Basis for Laplace Approximation. *Machine Learning*, vol. 33, no. 1.

Neyman, J. (1937). Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. *Philosophical Transaction of the Royal Society A*, vol. 236, pp. 333–380. Mathematical, Physical, Engineering Sciences.

O'Hagan, A. and Forster, J. (2004). *Kendall's Advanced Theory of Statistics*, vol. 2B. 2nd edn. Arnold. Bayesian Inference.

Paris, J.B. (1994). *The uncertain reasoner's companion. A mathematical perspective*. No. 39 in Cambridge tracts in theoretical computer Science. Cambridge University Press.

Pearson, K. (1907). On the influence of past experience on future expectation. *Philosophical Magazine*, vol. 6, no. 13.

Perks, W. (1947). Some observations on inverse probability including a new indifference rule. *Journal Institute of Actuaries*, vol. 73, pp. 285–334.

Pham-Gia, T. and Turkkan, N. (1998). Distribution of the linear combination of two general beta variables and applications. *Communications in Statistics - Theory and Methods*, vol. 27, no. 7, pp. 1851–1869.

Picard, E. (1880). Sur une extension aux fonctions de deux variables du problème de Riemann relatif aux fonctions hypergéométriques. *Comptes renduc hebdomadaires des séances de l'Académie des Sciences*, vol. 90, pp. 1267–1269.

Poincaré, H. (1896). Sur les intégrales irrégulières des équations linéaires. *Acta Mathematica*, vol. 8, pp. 295–344.

Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P. (1992). *Numerical Recipes in C: The Art of Scientific Computing*. 2nd edn. Cambridge University Press.

Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Graduate School of Business Administration, Harvard University.

Schrödinger, E. (1952). *Statistical Thermodynamics*. Dover edn. Dover Publications.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464.

Stigler, S.M. (1982). Thomas Bayes's Bayesian Inference. *Journal of the Royal Statistical Society Series A(General)*, vol. 145, no. 2, pp. 250–258.

Stigler, S.M. (1986). Laplace's 1774 memoir on inverse probability. *Statistical Science*, vol. 1, no. 3, pp. 359–363.

Stuart, A. and Ord, J.K. (1994). *Kendall's Advanced Theory of Statistics*, vol. 1. Sixth edn. Arnold. Distribution Theory.

Tierney, L., Kass, R.E. and Kadane, J.B. (1986). Fully exponential laplace approximations to posterior expectations and variances. *Journal of American Statistical Association*, vol. 84, pp. 710–716.

Tricomi, F. (1947). Sulle funzioni ipergeometriche confluenti. *Annali di Matematica Pura ed Applicata*, vol. 26, no. 1, pp. 141–175.

van Horn, K.S. (2003). Constructing a logic pf plausible inference: a guide to Cox's theorem. *International Journal of Approximate Reasoning*, vol. 34, pp. 3–24.

von Mises, R. (1939). *Probability, Statistics and Truth*. William Hodge and Company. Translated by J Neyman, D Sholl and E Rabinowitsch.

Watson, G.N. (1948). *Theory of Bessel Functions*. Cambridge University Press.

Whittaker, E.T. and Watson, G.N. (1927). *A Course of Modern Analysis*. 4th edn. Cambridge University Press. Reprinted 1990.

Zabell, S.L. (1989). The Rule of Succession. *Erkenntnis*, vol. 31, pp. 283–321.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. Wiley Classics Library. John Wiley & Sons.