THE HERFINDAHL-HIRSCHMAN INDEX AS AN OFFICIAL STATISTIC OF BUSINESS CONCENTRATION: CHALLENGES AND SOLUTIONS

by

George Georgiev Djolov

Dissertation presented for the degree of Doctor of Philosophy in Business Management and Administration in the Faculty of Economic and Management Sciences at Stellenbosch University

Promoter: Professor Eon van der Merwe Smit

December 2012

Declaration

By submitting this dissertation, I declare that the entirety of the work contained therein is my own, original work, that I am the authorship owner thereof (unless to the extent explicitly otherwise stated) and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

George G Djolov Date: 23 October 2012

Abstract

This dissertation examines the measurement of business concentration by the Herfindahl-Hirschman Index (HHI). In the course of the examination, a modification to this method of measurement of business concentration is proposed, in terms of which the accuracy of the conventional depiction of the HHI can be enhanced by a formulation involving the Gini index. Computational advantages in the use of this new method are identified, which reveal the Gini-based HHI to be an effective substitute for its regular counterpart. It is found that theoretically and in practice, the proposed new method has strengths that favour its usage. The practical advantages of employing this method are considered with a view to encouraging the measurement of business concentration using the Gini-based index of the HHI.

Opsomming

Hierdie verhandeling ondersoek die meting van sakekonsentrasie deur middel van die Herfindahl-Hirschman-indeks (HHI). ‘n Wysiging aan hierdie metode word voorgestel, deur middel waarvan die akkuraatheid van die konvensionele voorstelling van die HHI verhoog word, deur ‘n formulering wat die Gini-indeks betrek. Die berekeningsvoordele van hierdie nuwe metode word geïdentifiseer en dit word aangetoon dat die Gini-gebaseerde HHI ’n doeltreffende plaasvervanger vir sy meer bekende teenvoeter is. Daar word bevind dat die voorgestelde nuwe metode teoretiese en praktiese sterkpunte het wat die gebruik daarvan ondersteun. Die praktiese voordele van die voorgestelde metode word oorweeg met die oog op die aanmoediging van die gebruik van die Gini-gebaseerde HHI-indeks as maatstaf van sakekonsentrasie.

Acknowledgments

The writing of a PhD dissertation is a complicated and time-consuming endeavour, which cannot be accomplished without assistance and understanding from others.

I would like to start by thanking my promoter, Professor Eon Smit, for many challenging and intellectually-enriching discussions, for making sure that I stayed focused and on track, for making me appreciate the importance of substantiated and carefully-developed arguments, and ultimately for demonstrating the serious responsibility faced when presenting a new academic contribution in a field where one must “stand on the shoulders of giants.”

I would also like to thank Mr Joe de Beer, the Deputy Director-General of Economic Statistics at Stats SA, and Mr Gerhardt Bouwer, the Head of the National Accounts Division at Stats SA, both of whom gave me the support necessary to complete this work, while working full time.

I would like to thank the University of Stellenbosch Business School for nominating me to attend, in the concluding stages of my work, the 2011 Summer Academy of the European Doctoral Programmes Association in Management and Business Administration, held in Soreze, France. Both this programme, and the feedback I was fortunate to receive by the Examination Panel during my Oral Defence, were instrumental in shaping and refining a substantial portion of the final arguments in this dissertation.

Last, but not least, I would like to thank my wife, Hilary, and our daughter, Embeth, for their assistance and support, as well for putting up with many hours of absence on my part while I went about researching and writing the dissertation.

Dedication

In loving memory of my mother, Anna, and to the women in my life:

my wife, Hilary, and our daughter, Embeth.

CONTENTS

List of figures and tables

1. INTRODUCTION

> "What is our aim in writing…? Do we want it to impress by its length and convoluted style, or to be read, understood, and remembered? Technical writing is plagued by the belief that it is judged by its length – the longer, the better. …As a reaction, there are frequent calls to be brief and to write clearly".

With these famous remarks, Ehrenberg (1982: 326) left a lasting impression on the statistics profession in terms of what is expected of effective academic contributions in the field of statistics. In a succession of articles, Ehrenberg (1977: 278-287, 293-297; 1981: 68-70; 1982: 326-328) shaped what are nowadays accepted standards for statistical writing.

Among these are the requirements that statistical contributions should achieve substance over volume, as well as brevity and clarity; that they should facilitate understanding rather than indulging in contrived theorising; and achieve mastery over technical jargon, methods and notation, without resorting to scientism; that knowledge should be integrated and language clear and simple; and that analytic techniques should simplify the analysis of data, while enhancing the accuracy and usefulness of the results.

Ehrenberg twice cautioned (1977: 293; 1981: 70), and so, subsequently, did Marron (1999: 70-74), that if these standards were ignored, the outcome would always be an ineffective statistical contribution. Similarly, Taguchi and Clausing (1990a: 229) assert that the objective of any statistical enquiry is *understanding* – while Tukey (1980: 23-24; 1986: 74-75) goes further to say that this requirement for understanding is the reason statistical enquiries should be exploratory in nature. Tukey's description of how statistical enquiries function in practice is underlined by George Box (1988: 14), who observed:

> "…following the leadership of … Tukey … there is now more willingness to view the statistical investigator as a detective involved in an iterative and adaptive procedure in which deduction and induction alternate."

They certainly do, as will be demonstrated here. The present enquiry is a statistical contribution in the field of business statistics, which, as Sharma (2010: 3) observes, belongs to the branch of applied statistics. It is a field that uses mathematical and statistical theory to formulate and solve problems in other disciplines such as economics, business administration, public administration, medicine, education, and psychology. Our enquiry touches mainly on the first three. It is concerned with investigating whether the Herfindahl-Hirschman index of business concentration, is a statistically relevant index.

Introduced by Herfindahl (1955: 96, 98) and Hirschman (1964: 761), the Herfindahl-Hirschman Index (hereafter referred to as the HHI) is widely regarded as an invaluable tool for measuring business concentration. However, its mathematical and statistical accuracy has been called into question, and its stature as an official statistic remains ambiguous.

According to Rhoades (1993: 188), the HHI has become the most popular measure of business concentration, whether employed by economists, business executives or competition regulators. And there is no sign that, in this milieu, its popularity is on the wane. A notable editorial in *Business Week* (1998: 112) advised company executives that, "It would be nice if you could just watch your Herfindahls". In the same vein, *The Economist* (1998: 62) effused that that the calculation of the HHI is a simple matter because, "The Herfindahl's great virtue is its simplicity". Even the venerable *New Palgrave Dictionary of Economics and the Law (*2002 edition*)* considers the HHI to be the most practically relevant measure of business concentration. The United States Department of Justice and its Federal Trade Commission have now twice reaffirmed it as their index of choice for determining the business concentration of markets (1997 [1992]: 15-17; 2010: 19). The European Union has followed suit (European Union, 2004: 7). So too has South Africa (Competition Commission, 2009: 16, 18).

But popularity holds no sway with scientific enquiry. In a devastating, but little known, critique, Benoit Mandelbrot (1997: 215-216) dismissed the HHI saying:

> "This index has no independent motivation, and … it is odd that it should ever be mentioned in the literature, even solely to be criticised because it is an example of inconsiderate injection of a sample of second moment in a context where … the existence of expectation is controversial. …According to reports, Herfindahl's index is taken seriously in some publications. This is hard to believe."

Mandelbrot's criticism should not be overlooked – as it seems thus far to have been. Yet, the late Nobel Laureate Paul Samuelson remarked once – as cited in Hudson (2011: 9) – that:

> "On the scroll of great non-economists who have advanced economics by quantum leaps, next to John von Neumann we read the name of Benoit Mandelbrot."

An examination of the HHI in light of Mandelbrot's critique and specifically relating to its statistical accuracy, has the potential to change the way the HHI is used and understood. Despite its official status as a popular business concentration measure in many countries, the statistical agencies of those countries themselves make no particular effort to publish the HHI index regularly, and on the odd occasion that they do publish it, do not disclose the statistical accuracy of the official numbers. One example is offered by the United States Census Bureau (Census Bureau, 2006: 67-131). Another is Statistics South Africa (Stats SA, 1999: 230-277). But these are not isolated examples. Generally, the statistical agencies of any number of OECD countries publish the HHI a decade apart without disclosing the precision of the estimates (OECD, 2006: 33-36). Neither is this a new development. In the 1970s, Du Plessis (1979: 303, 308) observed that official information on business concentration in countries such as Australia, Germany, France, South Africa, the United Kingdom and the United States is of a piecemeal and limited character, typified by insufficient and incomparable data over time.

Thus, while the present enquiry as to whether the HHI is a statistically-relevant index may be said to be compelled by Mandelbrot's critique, it is equally motivated by the reality that the HHI is weak in stature as an official statistic. This enquiry investigates, therefore, whether we can satisfy the sticklers for accuracy and scientific precision, while retaining a popular and useful measure of business concentration. Without addressing Mandelbrot's critique, the HHI cannot be held to be a reliable measure of business concentration.

This is not the first time an enquiry of this kind has been suggested. Adelman (1969: 101) suggested decades ago that there was a need for a statistical test of significance for the HHI, while Reekie (1989: 199-216) has drawn attention to the fact that "preconceptions and prejudices" rather than objective measures may deliver a self-fulfilling diagnosis of business concentration where "experience and reason" would tend to demur.

Arriving at an answer requires an exploratory process of understanding why the statistical relevance of the HHI is under attack, and how this may be resolved. The framework for this exploration is summarised as follows.

Chapter 2 reviews the statistical properties of the HHI, finding that the index is indeed a statistical decision-making tool for business concentration. This is on account of its representation by the coefficient of variation, according to which its sampling distribution is approximated by the familiar Chi-square distribution. This approximation was discovered by McKay and is usually referred to as McKay's approximation. The subordination of the HHI to the Chi-square distribution highlights that on the basis of the observed HHI levels we can use the HHI to conduct hypotheses tests on business concentration, which aim to verify objectively without preconceptions or prejudices whether markets are concentrated. The conclusions from such a statistical test can provide economists, business executives, and regulators alike with a check on their conclusions about these levels, as well as force them to interrogate their analysis should the test contradict them. Provided that perfect data conditions are fulfilled, we will see that the application of the Chi-square distribution to the HHI gives an accurate method for determining the accuracy of HHI estimates, in

terms of allowing for the construction of confidence intervals. Such intervals make it possible to test the veracity of HHI estimates.

By illustrating the existence of confidence intervals for the HHI we can at least in part begin to make a case for why the HHI should be taken seriously, or indeed be believable. Such an illustration has a number of applications in a number of practical areas concerning the disciplines to which this enquiry is directed:

a) In business analysis, economic analysis, as well as investigations by competition regulators, it should be possible to report and discuss the HHI in the context of its probable range, as well as its expected value from that range. This in turn should give sense of how the observed HHI value compares in relation to its possible values as communicated by the data. Such a reorientation of analysis should provide for an improvement to the current practice where the HHI is descriptively handled as a single number without any sense of its real magnitude.
b) Confidence intervals for the HHI based on the Chi-square distribution should enable competition regulators to determine with confidence the statistical significance of testable hypotheses about their HHI thresholds. These are thresholds concerning the degree of monopolisation in markets. For business executives and economists, such intervals would also open up the prospect of using the index to test hypotheses about the nature of competition and the forms it can take.
c) Confidence intervals for the HHI based on the Chi-square distribution will equip statistical agencies with a method by which they can derive and disclose the accuracy of published numbers. It helps to be reminded that whether a statistic gains the status of an official statistic is dependent on whether it can be accurately measured (Elvers and Rosn, 1997: 622-626). If this cannot be done, its reliability or validity is compromised, which can lead to erosion of public trust in the numbers as well. The confidence intervals for the HHI make it possible to determine its accuracy, which undoubtedly would be an improvement to the current practice where its official estimates are disclosed without their margin of error.

In Chapter 3, it is established that the statistical nature of the HHI stands or falls by the data conditions it is subjected to. In its original depiction, it is an index that always presumes the existence of perfect data conditions, even though it is known that such conditions rarely exist in reality. This changes the appropriateness of the applicable measure of relative variability, which now becomes the Gini index. This will be demonstrated by two important inequalities of mathematical statistics known respectively as the De Vergottini and Glasser inequalities. In conjunction, the two show that, in the limit, as the number of observations increases, the Gini index and the coefficient of variation are equal. The same is also demonstrated by the Glasser inequality alone, in addition to also showing that in the event of fewer observations, it is only the Gini index that keeps its accuracy as a measure of relative variability. Another way by which this is typically expressed is to say that whenever the sampling distribution of the data is skewed, the measurement of relative variability by the Gini index is more accurate than that by the coefficient of variation; and whenever the skewness fades the two are equally accurate. Mandelbrot's criticism comes from these well-known results. Its insight is to show that the way Herfindahl and Hirschman have formulated the HHI ignores them.

The illustration of the Glasser inequality is handled by showing its limiting solution, which is the asymptotic equality between the Gini index and the coefficient of variation. This is done by showing a simplified proof for this equality, which is not matched in its simplicity to any of the historical or existing proofs comprehensively catalogued by Piesch (2005: 266-269, 275-282, 284). The proof is an extension to an incomplete derivation discussed by Milanovic (1997: 45-46). The resultant derivation will also explain Sawilowsky's (2006: 627-628) seemingly strange Monte Carlo results, which show that as the number of observations increases, the maximum value of the Gini index is 33%. Sawilowsky treated this as an unexplained empirical finding. It is not. It is in fact the exact solution to the De Vergottini and Glasser inequalities, which numerically corroborates that the Gini index and the coefficient of variation are asymptotically equal.

We will come to see that, in its original depiction, the HHI is certainly "inconsiderately injected" with the coefficient of variation. Indeed the coefficient is an example of a second moment of the

sampling distribution of the data, which on account of the Glasser inequality, is known to yield biased estimates of relative variability. This is in the sense that the measure systematically overstates the relative variability of the data except when the number of observations is large. *It does not help then to know what the expected value or range of the coefficient of variation is.* Considering that reality is pervaded by imperfect data conditions, their existence becomes controversial, essentially because *under these conditions we know that the coefficient will cease to be an accurate measure of relative variability.* It will also become apparent that because of the Glasser inequality, to rectify the situation we need only replace the Gini index for the coefficient of variation. As this outcome of the Glasser inequality carries through to any other measure that holds the coefficient of variation in its formulation, it will be proposed that the HHI should be reformulated to include the Gini index instead. This resultant expression or the necessity for it, is so far unknown or unrecognised.

En route to this reformulation, two other surprising and welcome results emerge. The first, is an explanation for a well-known result due to Kamat (1953: 452; 1961: 170, 172-174) and Ramasubban (1956: 120-121; 1959: 223), which showed empirically that the Chi-square distribution approximates the sampling distribution of the Gini index. They, however, assumed that this was just an empirical regularity. This enquiry establishes that their finding is not an empirical regularity. It is a special result of the Glasser inequality in terms of which the Gini index and the coefficient of variation are asymptotically equal, meaning that by default they also share the same sampling distribution. Secondly, they were unable to derive confidence intervals for the Gini index from the Chi-square distribution. Because of the aforementioned equality now established, the confidence intervals for the coefficient of variation based on the Chi-square distribution can also be said to extend to the Gini index. Therefore, confidence intervals for the Gini index are gained from the Chi-square distribution, which was formerly unknown. This result promises to yield a practical improvement in the analysis of data for the Gini index, which – as for the HHI – is carried out descriptively, by recourse to a single number, without consideration for its probable range or magnitude.

It will be shown that the replacement of the coefficient of variation with the Gini index is not just cosmetic, but strengthens the measure by substituting a robust measure of relative variability, for one that is not robust but relatively weak. There are many definitions of “robustness” as a statistical concept, but a neat and handy one provided by Morgenthaler (2007: 272, 277-278) is that robustness is the property of an estimator to retain its accuracy when the ideal data conditions for which it is designed begin to disappear or no longer exist.

The Gini index is equally comfortable under ideal data conditions and imperfect data conditions. This is the essence of the Glasser inequality. *Every time we opt to measure relative variability by the Gini index we avoid the potential pitfall of overestimation by the coefficient of variation*. By extension, the same holds for the HHI when reformulated in terms of the Gini index. The result is a robust measure of business concentration.

Due to the improvement in the accuracy of HHI by the inclusion of the Gini index, it becomes imperative to find an estimation technique by which to maintain this accuracy. Chapter 4 deals with this comprehensively, but not exhaustively, opening the door for future research in this area. Yitzhaki (1998: 24) has identified that there are more than a dozen estimation techniques available for the Gini index – and thus for the Gini-based HHI too. Cataloguing all of them would provide a unified picture of all the different methods that can be used to estimate the HHI, and whether these are of benefit. For now, to concentrate attention on the fact that the HHI gains accuracy from the Gini index, only two robust estimation techniques are considered. One of these, based on the Gini index, is the exceptionally popular Lerman and Yitzhaki method, extended by Ogwang. It uses the ranks and values of the observations, but the authors emphasise the mathematical benefits of easy computation at the expense of showing how and why the technique is derived. Therefore, a simple derivation of the technique is provided to demonstrate that it is a direct interpretation of the classical definition of the Gini index as twice the area of the Lorenz curve. Hopefully, tagging on a simple proof such as this will contribute to a sustained use of the technique in future.

In the present case the Lerman and Yitzhaki method is plugged-in into the HHI to produce a robust estimator for the index based on the ranks and values of the observations. A second robust estimator is presented in terms of the range of the data, for cases where there is lack of confidence in the accuracy of market share data for intermediary firms. This may occur because information tends to be most available at the extremes – for the biggest and smallest firms – while there may be lack of data for those in-between. It is a direct extension of that by Glasser, linking the Gini index with the range based on the Chebyshev theorem in terms of which the range is the quadrupled standard deviation of the data. Essentially, Glasser re-expresses the asymptotic equality between the Gini index and the coefficient of variation using this result, in order to connect the Gini index with the range. The contribution here is to extend this to the HHI.

The conventional method for the estimation of the HHI in terms of summing the squared market shares will also be discussed, but only to highlight that it has all the defects suggested by Mandelbrot. As will be seen, this is because it proceeds directly from the coefficient of variation, and suffers from the drawback that under imperfect data conditions it does not yield an accurate measure of business concentration from the HHI. In ideal conditions, the estimation of business concentration by the HHI is accurate. But just as accurate as it would be if the Gini-based index were used instead.

The point to remember is that keeping the Gini index as a permanent fixture of the HHI has the advantage of making it a reliable measure of business concentration irrespective of the data conditions encountered. It is also important to note that the new HHI estimation techniques illustrated are substantially different from the conventional estimation technique or any of its variations. Due to the Glasser inequality, conventional techniques leave the HHI as a potentially-biased measure of business concentration, susceptible to overstating levels of concentration – because neither the conventional technique, nor its variations, answer Mandelbrot's criticism. This means that in these cases its relevance as a reliable statistic measuring business concentration is diminished, because the coefficient of variation always operates in the background. It would then certainly be "hard to believe" that the HHI should be taken seriously. But the Glasser inequality

also suggests that the necessary step to resolve this is by reformulating the HHI in terms of the Gini index.

At this stage, it can be stated with confidence that the reformulation of the HHI in terms of the Gini index, as well as the subordination of its sampling distribution to the Chi-square distribution – whether reached by the coefficient of variation or the Gini index – is an extension of, or a special case of, the Glasser inequality. This seems to diminish the need for simulation studies that seek to find what the distribution of the HHI is. To be sure, such studies can be conducted, but what they will give is secondary findings because they would not tell us anything new that we do not already know from the asymptotic equality between the Gini index and the coefficient of variation. We will see this in Chapter 5, which for completeness, will deal empirically with the confirmation of the Glasser inequality when imperfect data conditions prevail.

Chapter 6 examines how the reformulation of the HHI by the Gini index can be used in practice. In terms of the data-analytic tool or technique that is to be used to show this, a prominent recommendation is that by Everitt and Dunn (1982: 45):

> "to choose the simplest … from those applicable to one's data, since this will generally ease the, at times, difficult task of interpretation of final results."

In terms of simplicity of implementation, and the simplified interpretation of results, Wu (1992: 140) suggests that Taguchi's transnumeration technique seems unrivalled. While there might be other mechanical methods there is nothing to suggest that transnumeration is inappropriate or less powerful. Transnumeration is a data-analytic technique involving statistical story-telling to facilitate numerical literacy and understanding of the practical uses of a statistic. Applied to the HHI, it demonstrates that:

a) The HHI is a statistical decision-making tool. It is an index with many possibilities of relevance to decision-makers in different contexts whenever they have to deal with the issue of business concentration or its estimation.

b) The HHI is intimately connected to the Gini index. It is an index for which expectations exist, and their associated values can be obtained by referral to the Chi-square distribution, which approximates the sampling distribution of the HHI.

c) The index is subordinate to the balance of probabilities to the extent that its statistical significance can be verified in any practical situation it is applied to without the need for doubt about the credibility or plausibility of the numbers.

d) In terms of measurement, the HHI will be mis-reported if it is only reported by itself without its confidence limits. There is now a way to determine the accuracy of the estimates.

e) Estimation with confidence intervals supports and influences decision-making through hypothesis testing. As a decision-making tool, the HHI makes it possible to study business concentration directly in terms of tests and hypotheses related to the Chi-square distribution.

f) Ultimately, because the HHI follows the Chi-square distribution, it must be regarded and treated as a statistic and a test procedure all in one. For this alone it more than adequately qualifies as an official statistic of business concentration.

g) Most important of all: it is an index that can test the statistical significance of the context it is subjected to. In terms of this, the HHI can be accurately estimated together with confidence limits, and this is quickly and effectively achieved when the HHI is treated as a robust measure of business concentration. This in effect creates a situation in terms of which only the Gini representation of the HHI should be used for the measurement of business concentration. This should not be hard to accept, because as will be seen, the HHI is just a variant of the Gini index, or to put it differently, the Gini index is just another version of the HHI.

The aim of the concluding chapter – Chapter 7 – is to integrate the preceding chapters and capture them in summary, so as to reinforce the findings above, and in particular the last point. And to lead

ultimately to the conclusion that while Mandelbrot's criticism of the HHI is legitimate and warranted, it does not necessitate writing off the HHI as a useful measure of business concentration. It does, however, mean re-formulating it as an expression of the Gini index.

On a point of clarity, regarding mathematical demonstrations, in order to keep these simple, it is assumed throughout that every population member is sampled. In other words, this means that the sample size (n) is equal to the population membership (N), or $n = N$. As Glasser (1962a: 628) explains, the meaning of this assumption is that samples of any size are being considered, as the population number varies.

2. REVIEW OF THE HHI

2.1 OVERVIEW

Rhoades (1993: 188) observed that the HHI is regarded by economists, competition regulators and business executives as the best-known and most widely used measure of business concentration. According to the United States Department of Justice and its Federal Trade Commission, markets can be divided into those that are un-concentrated; moderately concentrated; and highly concentrated, based on the HHI. In 1992, these agencies published the following thresholds for this breakdown (1992 [1997]: 15-17):

a) Un-concentrated markets have HHI levels below 10%;
b) Moderately-concentrated markets have HHI levels between 10% and 18%; and
c) Highly-concentrated markets have HHI levels above 18%.

In the course of publishing the revised thresholds in 2010, the agencies clarified (2010: 19) that the thresholds are "based on their experience". Commenting on the new thresholds the agencies stipulated that (2010: 19):

a) Un-concentrated markets, which do not affect competition adversely, now have HHI levels below 15%;
b) Moderately-concentrated markets, which can potentially affect competition adversely, now have HHI levels between 15% and 25%; and
c) Highly-concentrated markets, in which competition has the potential to be stifled outright, now have HHI levels above 25%.

These thresholds are not just a regulatory practice confined to the United States. For instance the South African competition authorities advise that they assess business concentration by the HHI thresholds set by the United States Department of Justice and its Federal Trade Commission

(Competition Commission, 2000: 24; 2009: 16, 18). Likewise, the competition authorities of the European Union make use of the same thresholds (European Union, 2004: 7).

The popularity of this measure among regulators holds true in the marketplace too: In an editorial in *Business Week* (1998: 112), company executives were advised that: “It would be nice if you could just watch your Herfindahls”. A similar editorial in *The Economist* (1998: 62) asserts that: “The Herfindahl’s great virtue is its simplicity” implying that the computation and interpretation of the Herfindahl measure of concentration is straightforward.

Indeed, a number of standard textbooks and works in economics and business administration, such as those of Acar and Sankaran (1999: 970, 975); Andreosso and Jacobson (2005: 98-99); Cabral (2000: 155); Carlton and Perloff (2000: 247-250); Smith and Du Plessis (1996: 3-10); Kelly (1981: 51-52, 55); Leach (1997: 15-18); Fourie and Smith (2001: 31-39); Fedderke and Szalontai (2009: 242-245); Fedderke and Naumann (2011: 2920-2922); and Salop and O’Brien (2000: 597-598, 610-611), use the HHI without any reference to its statistical character. This is understandable: so far a scant amount of work exists on this subject matter. But there is no reason why this should be so.

2.2 FORMULATION

According to Herfindahl (1955: 96), the HHI is the ratio between the index of heterogeneity and the number of observations, i.e. the number of firms in a market.

$$HHI = \frac{c^2 + 1}{n} \qquad 1.1$$

where c is the coefficient of variation of the observed values and n the number of firms.

This HHI definition appears among others in the works of Rosenbluth (1955: 62); Hart (1975: 425, 427-429); Adelman (1969: 100-101); Reekie (1989: 47); and Church and Ware (2000: 429). The coefficient of variation of observed values refers to the values of firms’ market shares. The market share of a firm may either refer to the proportion it holds of total industry output, or sales, or

production capacity. In the course of the present enquiry the terms *market* and *industry* are used interchangeably, as they refer to the same concept, i.e. a group of firms engaged in the same or similar kinds of production activity (OECD, 2008: 413).

The numerator of the HHI is usually denoted by R, and according to Bronk (1979: 669); Hürlimann (1995: 263; 1998: 128); and Hwang and Lin (2000: 134-135, 144), this numerator is known in statistics as the index of heterogeneity:

$$R = c^2 + 1 \qquad 1.2$$

As Gibrat found some time ago (1931 [1957]: 53):

> "…we know empirically that in most cases, particularly in the field of economics, distributions are not symmetrical but skew. This is immediately obvious from the fact that mean, median and mode do not coincide."

Gibrat (1931 [1957]: 57-58) found that this description also covers the distribution of firms' market shares. Subsequently Lawrence (1988: 231-233, 241-242, 251) provided a detailed survey of additional studies that corroborate Gibrat's finding. More recently, Axtel (2001: 1819-1820), as well as Gaffeo *et al*. (2003: 119-121) also found that generally firms' market shares have a positively skewed distribution, which is unimodal, i.e. with a single peak, that may also include extreme observations depending on how large the market share gap is between the top and bottom firms.

Generally, for any unimodal distribution, irrespective of its shape, the index of heterogeneity ranges between 1 and 2, or:

$$1 \leq R \leq 2 \qquad 1.3$$

Thus when there is no heterogeneity in the variation of the data, the heterogeneity index is 1 as the coefficient of variation is 0. Conversely when there is complete heterogeneity in the variation of the

data, the heterogeneity index is 2 as the coefficient of variation then is 1. The main ingredient of the heterogeneity index is the coefficient of variation, which is sometimes also called the relative standard deviation and its square is also called the relative variance (Hürlimann, 1998: 128). The last three terms are used interchangeably in the present enquiry. Bronk (1979: 668-669) also provides a proof for a number of other well-known results concerning the coefficient of variation, which by implication also extend to the heterogeneity index:

a) When the coefficient of variation of the data is equal to zero its sampling distribution is uniform;
b) When the coefficient of variation of the data is equal to one its sampling distribution is exponential;
c) When the coefficient of variation of the data is anywhere between these extremes, as well as when it approaches them, its sampling distribution is indeterminate in the sense that it can take any positively skewed unimodal form.

These findings have also been reported by Hürlimann (1995: 263) and independently reproved by Hwang and Lin (2000: 135-144, 144). Consequently the range of the coefficient of variation, and by implication that of the heterogeneity index, does not only denote abstract values. More importantly it gives signals about the shape of the data's distribution.

The same findings apply to the HHI too, given that its main ingredient, as for the heterogeneity index, is the coefficient of variation. However they apply for a slightly different range of values to those of the coefficient of variation, or the heterogeneity index. The HHI has a minimum value that comes progressively close to zero as the number of observations (i.e. firms) increases, and a maximum value of 1 when there is only a single firm. To see this, recall the De Vergottini inequality for the coefficient of variation as reproduced by Piesch (2005: 284), which is also discussed in the first appendix to this chapter. This one-sided non-strict inequality stipulates that the maximum value of the product of the reciprocal of the square root of 3 and the coefficient of variation of the ranks of the data (c_i), is the coefficient of variation of its values, or:

$$\frac{1}{\sqrt{3}} c_i \leq c \qquad 1.4$$

The fact that inequalities are being dealt with so early on into the enquiry should not be surprising. As Bellman (1954: 21) remarked:

> "It has been said that mathematics is the science of tautology, which is to say that mathematicians spend their time proving that equal quantities are equal. This statement is wrong on two counts: In the first place, mathematics is not a science, it is an art; in the second place, it is fundamentally the study of inequalities rather than equalities."

While a strict equality, or what is commonly called just an equality, might be a rarity in practice, asymptotic equalities arising out of one-sided non-strict inequalities are treated in the same way as equalities. This is the reality of mathematics in practical situations, and has been called by Tukey (1986: 74) the "ultimate oversimplification". The present enquiry will offer several examples of this. We can for instance rewrite expression 1.4 to show that the ratio between the coefficients of variation from ranks and values never exceeds 1, which implies that both sides of the expression are asymptotically equal:

$$\frac{\frac{1}{\sqrt{3}} c_i}{c} \leq 1 \Rightarrow \frac{1}{\sqrt{3}} c_i = c \qquad 1.5$$

Remember that any two measures are asymptotically equal if in the limit the ratio between them approaches 1 as the number of observations increases. Then the equality expression between them is said to be an asymptotic formula in terms of which the measures are asymptotically equal, even if the one measure does not actually equal the other measure for all observed values. In short, the one measure essentially behaves like the other measure as the number of observations becomes larger. Thus, expression 1.5 is an asymptotic formula, representing an asymptotic equality in terms of which the product of the reciprocal of the square root of 3 and the coefficient of

variation for the ranks of the observations is equal to the coefficient of variation of the data if obtained from its values.

Formally, the mathematical convention for the computation of an asymptotic formula is to treat it in the same way as an equality. As Goldreich and Wigderson (2008: 578) explain:

> "Most of the time, we are interested not so much in the full … function.... …And usually we do not look for an exact formula…: for most purposes it is enough to have a good upper bound."

Simply put, an asymptotic equality is by mathematical convention treated like an equality. This practice is an extension to that of treating all one-sided non-strict inequalities as equalities because asymptotic equalities arise from such inequalities, just as in the present case. We know that all such inequalities have an equality analogue by virtue of giving an exact solution to some minimum or maximum value. This is why as Lange (1959: 157-158; 1963: 490-491) points out, it is an established mathematical convention to treat any one-sided non-strict inequality in the same way as an equality. This convention remains in force. For instance the 1992 edition of the *Academic Press Dictionary of Science and Technology* reports that any asymptotic formula should be expressed with a strict equality notation. Similarly, the 2009 edition of the *Oxford Concise Dictionary of Mathematics* indicates that any asymptotic formula is to be expressed as an equality. The reasoning behind this convention is straightforward. It reveals that the one measure is equal to the other measure up to a constant that can never exceed 1. In the words of Goldreich and Wigderson, this is a good upper bound because it permanently fixes the values of the measures to be the same as the number of observations increases. The fact that there will be some number of observations for which this limiting case is not reached, meaning that for those cases the ratio between the two measures does not completely converge to 1, is not sufficient grounds to argue that the equality between them is absent. This is why the two sides of expression 1.5 are ultimately equated.

In response to expression 1.5, some further simplifications follow by reworking the coefficient of variation for the ranks of the data. In respect of the ranks, the coefficient of variation can be expressed as follows:

$$c_i = \frac{\sigma_i}{\mu_i} \Rightarrow c^2 = \frac{\sigma_i^2}{\mu_i^2} \qquad 1.6$$

where the standard deviation of the ranks is σ_i, and their mean is μ_i.

Generally we know that the variance of the ranks is:

$$\sigma_i^2 = \frac{1}{12}\left(n^2 - 1\right) = \frac{1}{12}(n-1)(n+1) \qquad 1.7$$

While, in turn we also know that the mean of the ranks is:

$$\mu_i = \frac{n+1}{2} \Rightarrow \mu_i^2 = \frac{(n+1)(n+1)}{4} \qquad 1.8$$

Then by substitution of expressions 1.7 and 1.8 into expression 1.6, we get:

$$c_i^2 = \frac{\frac{1}{12}\left(n^2 - 1\right)}{\left(\frac{(n+1)}{2}\right)^2} = \frac{\frac{1}{12}(n-1)(n+1)}{\frac{(n+1)(n+1)}{4}} = \frac{4}{12} \bullet \left(\frac{n-1}{n+1}\right)$$
$$= \frac{1}{3} \bullet \left(\frac{n-1}{n+1}\right) \qquad 1.9$$

By taking the square root of expression 1.9, we obtain:

$$c_i = \frac{1}{\sqrt{3}} \bullet \sqrt{\left(\frac{n-1}{n+1}\right)} \qquad 1.10$$

In turn expression 1.10 can be substituted into expression 1.5 giving the following relationship between the coefficient of variation from the values of the observations and that of their respective ranks:

$$c = \frac{1}{\sqrt{3}} \bullet \frac{1}{\sqrt{3}} \sqrt{\left(\frac{n-1}{n+1}\right)} = \frac{1}{3}\sqrt{\left(\frac{n-1}{n+1}\right)} \qquad 1.11$$

Furthermore, the square root term in expression 1.11 approaches 1 with 25- or more observations:

$$\sqrt{\left(\frac{n-1}{n+1}\right)} \to 1, n \geq 25 \qquad 1.12$$

Substituting the limiting value of expression 1.12 into expression 1.11 we reach the well-known solution of the De Vergottini inequality, namely that with a growing number of observations the coefficient of variation of their values is approximately one-third:

$$c \cong \frac{1}{3} \qquad 1.13$$

Piesch (2005: 284) refers to this result as one of the important special cases of the De Vergottini inequality. It shows that it is unlikely that as the number of observations grows the coefficient of variation will reach its theoretical maximum value of 1. Smaller values will have to be contended with. To account for this practical possibility the ranges of the coefficient of variation and the heterogeneity index are sometimes re-expressed as follows (Bronk, 1979: 669; Hürlimann, 1995: 263; Hwang and Lin, 2000: 135):

$$0 \leq c < 1, 1 \leq R < 2 \qquad 1.14$$

From expression 1.14 we can infer that in the case of a single firm ($n = 1$) the coefficient of variation is zero, and the maximum value of the HHI is 1. Concerning the HHI's minimum value, from expression 1.11 it follows that the square of the coefficient of variation for the values of the data is:

$$c = \frac{1}{3}\sqrt{\left(\frac{n-1}{n+1}\right)} \Rightarrow c^2 = \frac{1}{9}\left(\frac{n-1}{n+1}\right) \qquad 1.15$$

By substitution of expression 1.15 into expression 1.1, the minimum value or lower limit of the HHI (HHI_l) is given by:

$$HHI_l = \frac{\frac{1}{9}\left(\frac{n-1}{n+1}\right)+1}{n} = \frac{1}{9n}\left(\frac{n-1}{n+1}\right)+\frac{1}{n} \qquad 1.16$$

The first term of expression 1.16 can be ignored because with 2- or more observations it approaches zero faster than the second term:

$$\frac{1}{9n}\left(\frac{n-1}{n+1}\right) \to 0, n \geq 2 \qquad 1.17$$

In turn, the minimum value of the HHI can be approximated by the reciprocal of the number of observations, i.e. firms. Thus the range of the HHI is:

$$\frac{1}{n} < HHI \leq 1 \qquad 1.18$$

The necessary condition as to how the HHI acquires its minimum value will be further examined in Chapter 3. For now we should keep in mind that the practical existence of the range for the coefficient of variation as per expression 1.14 is undoubted. It is the main reason why as summarised by Hwang and Lin (2000: 1979):

> "Unfortunately, the exact probability distribution of the sample coefficient of variation under most populations is still unknown."

As aforementioned, when the coefficient of variation of the data is anywhere between these extremes its sampling distribution is indeterminate in the sense that it can take any positively skewed unimodal form. By extension the same holds for the heterogeneity index, and in turn the same can be inferred to apply for the HHI when its values happen to be somewhere between its minimum and maximum limits.

2.3 DISTRIBUTION

There is no need to despair that the exact probability distribution of the sample coefficient of variation is unknown. After all, whatever this distribution might be, its shape is positively skewed. Acting on this knowledge, McKay (1932: 697-698) proposed that the sampling distribution of the coefficient of variation can be approximated by the Chi-square distribution with n-1 degrees of freedom. Because this distribution is positively skewed it immediately lends itself as a natural contender for the sampling distribution of the coefficient of variation. On Egon Pearson's advice (1932: 703), Fieller (1932: 699) replicated McKay's proposed approximation, and came to the conclusion that "…the approximation…is…quite adequate for any practical purpose". Pearson (1932: 703) followed up with another independent assessment likewise reaching the same conclusion. Iglewicz and Myers (1970: 167-169) continued re-evaluating McKay's approximation finding what McKay, Fieller, and Pearson before them had already found. In their case they concluded that (Iglewicz and Myers, 1970: 169):

> "…the... approximation…of…McKay's can certainly be recommended on the basis of both accuracy and simplicity."

There have been more studies that have confirmed this finding, notably those by David (1949: 388-390); Iglewicz, Myers and Howe (1968: 581); Umphrey (1983: 630-634); Reh and Scheffler (1996: 451-452); Vangel (1996: 21, 24-25); Forkman and Verrill (2007: 10-11); Forkman (2009: 234); George and Kibria (2012: 1226-1234, 1239); and Gulhar *et al*. (2012: 48-50; 55-58, 61). The gist of these various studies is that, irrespective of the distribution of the data, McKay's approximation is accurate with any number of observations. Sometimes this is technically described by the statement that McKay's approximation is valid provided that the population coefficient of variation does not exceed its limiting value of one-third (Forkman, 2009: 234, 239). By recourse to the De Vergottini inequality, as per expressions 1.11 and 1.13, we can see that this condition is satisfied for any number of observations – and conclude, like Fieller, that the approximation is adequate for any practical purpose.

George and Kibria (2012: 1227-1228), as well as Gulhar *et al*. (2012: 49), describe how the approximation can be computed from McKay's original confidence interval (Λ_1^c), which is given by:

$$\Lambda_1^c = \left(\frac{c}{\sqrt{\left| c^2 \left(\frac{X_u^2}{n} - 1 \right) \right| + \frac{X_u^2}{n-1}}}, \frac{c}{\sqrt{\left| c^2 \left(\frac{X_l^2}{n} - 1 \right) \right| + \frac{X_l^2}{n-1}}} \right) \qquad 1.19$$

where X_l^2 and X_u^2 are respectively the lower and upper critical values from the Chi-square distribution with n-1 degrees of freedom

George and Kibria (2012: 1227-1228), and Gulhar *et al*. (2012: 49), also describe how the approximation can be computed from McKay's modified confidence interval (Λ_2^c), proposed by Vangel (1996: 23-24). This interval is given by:

$$\Lambda_2^c = \left(\frac{c}{\sqrt{\left| c^2 \left(\frac{2 + X_u^2}{n} - 1 \right) \right| + \frac{X_u^2}{n-1}}}, \frac{c}{\sqrt{\left| c^2 \left(\frac{2 + X_l^2}{n} - 1 \right) \right| + \frac{X_l^2}{n-1}}} \right) \qquad 1.20$$

where X_l^2 and X_u^2 are respectively the lower and upper critical values from the Chi-square distribution with n-1 degrees of freedom

On a technical note, the presence of the absolute values in the denominator of both confidence intervals, prevents the possibility of cases where the limits of the intervals do not exist in their absence. A practical illustration of this possibility is provided by Wong and Wu (2002: 74, 80). In practice, there appears to be no difference depending on which of the intervals is applied. For instance, Vangel (1996: 25) finds that the computations from the modified McKay confidence interval differ from the original McKay confidence interval only in the fourth decimal place. So up to the third digit after the decimal their computed values are found to be the same. This finding however should not be taken to imply that we have to choose either interval. For comparative purposes we can work with both, and subsequently decide which one to adopt. Once both intervals

are constructed, we know that the width (W) of either one of them measures the accuracy of the estimate of the expected value (Diaconis and Efron, 1983: 100). The width is the difference between the upper (U) and lower limit (L) for either confidence interval:

$$W = U - L \tag{1.21}$$

We also know that either confidence interval gives the bias (B) of that estimate (Diaconis and Efron, 1983: 100; Boddy and Smith, 2009: 32-33, 53-54). This is the average amount by which the observed value of the coefficient of variation differs from its true value, and for either interval, is given by half its width:

$$B = \frac{1}{2}W \tag{1.22}$$

The bias represents the systematic error – however caused – by which the estimated expected value persistently deviates from the true value. Ideally this systematic error should be 0. However as Mooney and Duval (1993: 33) remind us, in practice, this error does not distort the estimate if its deviations from the true value do not exceed 25% or 0.25 percentage points. Estimates that exceed this bias criterion should be considered to be affected by bias in the sense that the estimated expected value either understates or overstates the true value of the measure. In such cases the practical solution to the problem of systematic error irrespective of its direction is to correct the confidence intervals in terms of their lower and upper limits by increasing the former and decreasing the latter. As we can see from expression 1.22 the corresponding width of a confidence interval with a tolerable bias of 25%, is 50%. Thus in the event of bias exceeding 25% the corrective amount (K) to add to the lower limit of a confidence interval and subtract from its upper limit is:

$$K = \frac{|W - 50\%|}{2} \tag{1.23}$$

This adjustment to the limits decreases the bias in the estimate of the expected value by bringing it in line to the tolerable limit of 25% where it is still small enough to not have any real influence on it.

Following from the calculations of both intervals, the interval with the smallest width is the accurate interval because it is the one that in the present case will give the least bias in the estimation of the coefficient of variation. This incidentally will also determine which interval is to be accepted as the one giving the most accurate range of estimates for the coefficient of variation in practice. To clarify, the middle point of the confidence interval or that of the range it produces gives the expected value (E) of the coefficient of variation. It is the average of the limits:

$$E = \frac{L + U}{2} \qquad 1.24$$

Table 1 in the second appendix to this chapter gives a condensed table of Chi-square values covering the conventional significance levels of 1%, 5%, and 10%, which in descending order correspond respectively to confidence intervals with coverage probabilities of 99%, 95%, and 90%. The table is created from the Chi-square distribution table published in Ott (1993), which also covers the 99.8%, 98%, and 80% confidence intervals.

Using Table 1 we can devise and test significance tests for the coefficient of variation in the same way as we do for the standard deviation. In particular, if we have in mind some specific value (y) for the coefficient of variation and we wish to know if this value is different, lower or higher we will define our null hypothesis as assuming that this value exists ($H_o : c = y$) and compare it to anyone of the possibilities from the alternative hypothesis in terms of which this value may be lower ($H_A : c < y$), higher ($H_A : c > y$), or different ($H_A : c \neq y$). We will reject the null hypothesis if the range of the calculated confidence interval for the coefficient of variation does not contain the value and will adopt the alternative hypothesis being tested. Because the confidence interval is bidirectional we only test the null hypothesis against the bidirectional alternative hypothesis since the assessment here simultaneously reveals whether the tested value is lower or higher.

McKay's confidence intervals for the coefficient of variation enable us not only to derive the expected range of its values but also thereby to justifiably talk about its expected value. More

importantly by a mere rewrite of expression 1.1, they extend to the HHI too. To see this, let's rearrange expression 1.1:

$$c = \sqrt{nHHI - 1} \qquad 1.25$$

Then by substitution of expression 1.25 into expressions 1.19 we get McKay's original confidence interval for the HHI:

$$\Lambda_1^{HHI} = \left(\frac{\sqrt{nHHI - 1}}{\sqrt{(nHHI - 1) \bullet \left| \left(\frac{X_u^2}{n} - 1 \right) \right| + \frac{X_u^2}{n - 1}}}, \frac{\sqrt{nHHI - 1}}{\sqrt{(nHHI - 1) \bullet \left| \left(\frac{X_l^2}{n} - 1 \right) \right| + \frac{X_l^2}{n - 1}}} \right) \qquad 1.26$$

In turn by substitution of expression 1.25 into expression 1.20 we get McKay's modified confidence interval for the HHI:

$$\Lambda_2^{HHI} = \left(\frac{\sqrt{nHHI - 1}}{\sqrt{(nHHI - 1) \bullet \left| \left(\frac{2 + X_u^2}{n} - 1 \right) \right| + \frac{X_u^2}{n - 1}}}, \frac{\sqrt{nHHI - 1}}{\sqrt{(nHHI - 1) \bullet \left| \left(\frac{2 + X_l^2}{n} - 1 \right) \right| + \frac{X_l^2}{n - 1}}} \right) \qquad 1.27$$

The implications of the last two expressions ought to be readily apparent. *Firstly*, for the first time we now learn that due to the coefficient of variation the approximate sampling distribution of the HHI is the Chi-square distribution. Because of this finding our knowledge about the distribution of the HHI should improve to the extent that we can no longer ignore the existence of an expectation for the HHI or its expected range, much less ignore having to provide their values when working with the index.

So far, there is no evidence from economics, business administration, or competition regulation that shows the HHI as a statistic with self-contained confidence limits. Instead, in these areas, the index is only descriptively discussed. For instance a number of prominent works, by among others Herfindahl (1955: 96); Rosenbluth (1955: 62); Reekie (1989: 47); Smith and Du Plessis (1996: 3-

10); Leach (1997: 15-18); Ancar and Sankaran (1999: 970, 975); Church and Ware (2000: 429); Cabral (2000: 155); Carlton and Perloff (2000: 247-250); Fourie and Smith (2001: 31-39); Theron (2001: 638-645); Doyle (2005: 205); and Andreosso and Jacobson (2005: 98-99), make use of the HHI without reporting its confidence limits or its significance. The list is of course much longer.

What is important to keep in mind is that the aforementioned listing is not in any way an attempt to single out anyone of these studies. The intention is only to highlight the point that up to now the reporting of confidence intervals for the HHI has not been practiced. Instead the customary practice is to report the HHI as a single number. This is clearly misleading, because in that case we have no real knowledge of its magnitude, and McKay's confidence intervals for the HHI make this clear enough. From these intervals we can comfortably derive the expected range of the HHI as well as its expected value. In addition we can use McKay's confidence intervals for the HHI to determine the accuracy with which the HHI is estimated. For instance, if exceptionally high accuracy is wanted from the estimation of the HHI, based on McKay's confidence intervals we can derive 99% confidence limits for the HHI. The corresponding Chi-square values for deriving such an interval are published in Table 1 in the second appendix to this chapter.

Secondly, by introducing McKay's confidence intervals for the HHI we have achieved a rejoinder with Adelman (1969: 101). In particular Adelman (1969: 101) pleaded that:

> "We need a test of significance for H … to see whether differences over a time, or differences among industries at any one time, may be attributed to chance, or whether something more abides".

Until now this plea has remained unanswered. Expressions 1.26 and 1.27 change this. In a literal and figurative sense, McKay's confidence intervals make it possible to do hypothesis testing of business concentration by using the HHI directly. This is in principle the idea that Adelman had in mind. To evaluate the potential of this, let us refer to how the 2006 edition of the *Collins Dictionary of Economics* describes concentration measures:

> "Concentration measures are widely used in economic analysis and for purposes of applying Competition Policy to indicate the degree of competition or monopolisation present in a market."

This description of what a concentration index measures coincides with the HHI regulatory thresholds discussed at the outset of the present chapter. In both instances the degree of competition in a market is inversely related to the degree of monopolisation. Judging from the regulatory thresholds, regulators regard the degree of concentration in a market as stifling or eliminating competition whenever the HHI exceeds 25%. Lower HHI values are deemed to indicate that the degree of monopolisation in a market is either less damaging or not harmful to the extent of competition in that market. From these threshold values we can formulate the following testable hypothesis for the HHI:

Null hypothesis, Ho: $HHI = 25\%$, vs.

Alternative hypotheses, Ha: $HHI \neq 25\%$

It should be noted that there is nothing special about testing for the HHI at the 25% level. In the present case this is done merely for *illustrative* purposes. However, whatever threshold level is chosen, it should be kept in mind that it will adhere to the same process of statistical testing as for the illustrative case here. Returning to this case, and applying to the 25% level of the HHI the above-mentioned economic terminology from the *Collins Dictionary of Economics*, suggests that the HHI hypothesis for this level can be expressed in the following qualitative terms:

Ho: The degree of monopolisation in a market is borderline harmful to its degree of competition, vs.

Ha: The degree of monopolisation in a market may or may not be borderline harmful to its degree of competition

Since a confidence interval is used to test this null hypothesis if the evidence favours the alternative hypothesis we would be in a position to answer whether the HHI level is higher or lower than 25%. The confidence interval and its associated hypothesis test on the degree of monopolisation will help competition regulators decide with a degree of certainty whether this degree is harmful or not. There is no prescription for which confidence level should be adopted. In this regard either of the Chi-square values reported in Table 1 can be used for a 99%, 95%, or 90% confidence interval. Here these are deliberately ordered in a declining order of confidence in order to remind that the strength of the decision will weaken the smaller the confidence interval becomes in its coverage probability. For instance if the judgement by a competition regulator concerning the degree of monopolisation in a market is meant to be communicated as having been made with utmost care then the HHI hypothesis test should be done with a 99% confidence interval. Whatever the outcome from such a test we would know with certainty that the chance of being wrong from the resultant hypothesis decision it leads to will be only 1%, or conversely the chance of being correct about such decision will be 99%. On the other hand if the gravity of the situation under examination is not serious then nothing stops using either a 90% or a 95% confidence interval, provided it is acknowledged that with a 90% confidence interval the degree of confidence will be lower than with a 95% confidence interval.

Whichever of the conventional levels of statistical significance are chosen, it is important to keep in mind that there is *nothing magical* to them. They are merely statistical conventions about the role of chance we are prepared to give in the analysis of data, and the conclusions that flow from it. Whether as a result of them we end up with a 99% confidence level (in the case of the 1 significance level), a 95% confidence level (in the case of the 5% significance level), or a 90% confidence level (in the case of the 10% significance level), will not answer whether logically the finding from an analysis makes any sense, or whether it is practically significant by being useful in the real world. To answer the first question is always dependent on context knowledge as relates to the subject matter to which the statistical enquiry is applied. To answer the second question is always a matter of professional judgment, which need not even be informed by statistical analysis. Statistical analysis does not prescribe decisions. It informs decision-making. In the current

situation, the possibility of hypothesis testing with the HHI appears helpful in economic analysis. For instance the 2006 edition of the *Collins Dictionary of Economics* also notes that:

> "The significance of market concentration for market analysis lies in its effect on the nature and intensity of competition. Structurally, as the level of seller concentration in a market progressively increases, "competition between the many" becomes "competition between the few" until, at the extreme, the market is totally monopolised by a single supplier. In terms of market conduct, as supply becomes concentrated in fewer and fewer hands (oligopoly), suppliers may seek to avoid mutually ruinous price competition and channel their main marketing efforts into sales promotion and product innovation, activities that are more profitable and effective way of establishing competitive advantage over rivals."

So McKay's confidence intervals for the HHI also open up the possibility of hypothesis testing with the HHI in economic theory. It is clear that such an HHI test can shed light on the shifting direction of competition as well as on the form competition can take. The just-quoted economic terminology from the *Collins Dictionary of Economics* implies that in the first case the HHI test will probe whether there is movement from "competition between the many" to "competition between the few", and the hypothesis test will be:

> Ho: Market structure is characterised by "competition between the many", vs.
> Ha: Market structure is characterised by "competition between the few"

In the second case, the same economic terminology implies that the HHI test will probe whether the form competition takes is between price rivalry and non-pricing activities such as marketing efforts into sales promotion and product innovation, or anything that does not deal with price in general. The hypothesis test for this case will be:

> Ho: Market conduct is characterised by price competition, vs.
> Ha: Market conduct is characterised by non-price rivalry

The last two hypotheses are also relevant to business executives when they have to plan how to keep up their presence in the market they operate in, or when they have to plan for market entry and need to know how to build competitive advantage over rivals. If there is a specific HHI level that attaches to these hypotheses in the same way as exists for regulators from the thresholds they set for the HHI, then the hypotheses can be also equivalently formulated in quantitative terms. This however is not essential because for the qualitative formulations whether the null hypothesis is confirmed or not, depends on whether the calculated HHI value falls or does not fall in the calculated confidence intervals for the HHI. In the first case the null hypothesis will be confirmed. In the second case it will be rejected.

Thirdly, by enabling hypothesis testing for the HHI, McKay's confidence intervals provide an objective statistical standard for determining business concentration. The relevance of this should not be discounted nor belittled. For instance Reekie (1989: 199-216) provides a detailed account of practical instances where competition regulators in many jurisdictions have made such assessments on the basis of preconceptions and prejudices that cannot be reconciled either with experience or reason. Adelman's call for a statistical test of the HHI, made in 1969, also suggests that this may be something that is ongoing. By contrast if a statistical test for the HHI did exist it will tend to force-down decision-making that is absent from objective considerations. This is because the aim of a statistical test is to encourage decision-making that is divorced from preconceptions, while also giving – through the confidence interval – a range of propositions based on different concentration levels from which to assess the magnitude of observed business concentration. McKay's confidence intervals make this possible for the HHI.

2.4 SUMMARY

To summarise, in the present chapter, we have found the HHI to be a statistical decision-making tool for business concentration. This is on account of its representation by the coefficient of variation due to which its sampling distribution approximately follows the Chi-square distribution. By recourse to the Chi-square distribution we can then conduct hypotheses tests on business concentration that aim to verify objectively without preconceptions or prejudices whether markets

are concentrated on the basis of their observed HHI levels. The conclusions from such a statistical test will provide economists, business executives, and regulators alike with a check on their conclusions about these levels as well as force them to interrogate their analysis should the test contradict them. In addition McKay's approximation for the HHI gives an accurate method for determining the accuracy of HHI estimates by referral to the Chi-square distribution. This makes it possible to attest the veracity of any HHI estimate. It should be beyond any doubt that with the HHI confidence intervals any HHI number must be taken seriously and must be regarded as believable *provided* we are prepared to disclose its accuracy as well as the level of confidence we have in it. Possible applications of this finding suggest themselves in a number of practical areas:

a) In business analysis, economic analysis, as well as investigations by competition regulators, the HHI can now be reported and discussed in the context of its probable range, as well as its expected value from that range. This in turn gives sense of the comparative benchmark (or benchmarks) against which the observed HHI value can be compared to. Such a reorientation of analysis would be an improvement to the current practice where the HHI is descriptively handled as a single number without any sense of its real magnitude.
b) McKay's confidence intervals for the HHI now enable competition regulators to determine with confidence the statistical significance of testable hypotheses about their HHI thresholds on the degree of monopolisation in markets. For business executives and economists, these intervals enable the use of the index to test hypotheses about the nature of competition and the forms it can take. The relevance of this point is eloquently captured by Carlton and Israel (2010: 3). After a historical review of how the HHI thresholds are arrived at, Carlton and Israel (2010: 3) concluded that:

"…regardless of the precise cut-off levels used, it would be a mistake … to infer from the fact that there are new HHI thresholds in the 2010 Guidelines that there has been any new research to justify giving special credence to these new thresholds. Indeed, we know of no body of economic research that provides either an econometric or a theoretical basis for the

HHI thresholds in the 2010 Guidelines or for that matter in previous versions of the Guidelines."

Clearly, in the present case, statistical theorising fills a vacuum. By subjecting anyone of the threshold levels to a statistical test makes it possible to find out if they actually make sense in the context of economic theory and research. Systematic testing will thus make it possible to establish whether there is a basis for these levels, or any other for that matter.

c) The possibility of statistical agencies publishing the HHI as an official statistic with the simultaneous disclosure of the accuracy levels of its estimates is now established. Ultimately, whether a statistic gains the status of an official statistic is dependent on whether it can be accurately measured (Elvers and Rosn, 1997: 622-626). If this is not achieved its reliability or validity is compromised apart from the erosion of public trust in the numbers. McKay's confidence intervals make it possible to determine what this accuracy is with reference to the HHI. This is an improvement to the current practice where the official estimates of the HHI are published without disclosure of their confidence limits. This can be witnessed in the officially published HHI numbers by various statistical agencies of any number of OECD countries (OECD, 2006: 33-36). The same lack of disclosure is also evident in South Africa (Stats SA, 1999: 230-277). However without them it is simply impossible to know anything about their margin of error, thereby requiring acceptance of the numbers at face value. From McKay's confidence intervals we know now that this is avoidable.

2.5 APPENDIX: DE VERGOTTINI'S INEQUALITY

This appendix derives the De Vergottini inequality referred to in the body of Chapter 2. Piesch (2005: 284) reproduced it as a well-known result without deriving it. Glasser (1961a: 177) also worked with it, without giving a proof. Likewise De Vergottini (1950: 452) himself, only gave it as an end result. Instead he focused on drawing attention that it has the interesting property of leading to an asymptotic equality between the coefficient of variation and the Gini index, which has an exact solution of one-third. Fuller information on this will be presented in Chapter 3.

In light of the absence of an explicit proof, the derivation of the De Vergottini inequality offered here is new. It is based on its solution in the limit, as the number of observations increases.

Recall that the coefficient of variation for the ranks of the data is the ratio of the standard deviation of the ranks to the mean rank, or:

$$c_i = \frac{\sigma_i}{\mu_i} \qquad \text{1A.1}$$

In turn we know that the standard deviation, or the variance of the ranks, is given by:

$$\sigma_i = \sqrt{\frac{1}{12}\left(n^2 - 1\right)} \Rightarrow \sigma_i^2 = \frac{1}{12}\left(n^2 - 1\right) \qquad \text{1A.2}$$

We also know that the mean rank, or its square, is given by:

$$\mu_i = \frac{n+1}{2} \Rightarrow \mu_i^2 = \frac{(n+1)(n+1)}{4} \qquad \text{1A.3}$$

Substituting expressions 1A.2 and 1A.3 into the squared coefficient of variation for the ranks of the data, leads to:

$$c_i^2 = \frac{\sigma_i^2}{\mu_i^2} = \frac{\frac{1}{12}\left(n^2 - 1\right)}{\frac{(n+1)^2}{4}} = \frac{4}{12}\frac{(n-1)(n+1)}{(n+1)^2} = \frac{1}{3}\frac{(n-1)}{(n+1)} \qquad \text{1A.4}$$

Taking the square root of expression 1A.4 results in the following expression for the coefficient of variation of the ranks:

$$c_i = \frac{1}{\sqrt{3}} \bullet \sqrt{\left(\frac{n-1}{n+1}\right)} \qquad 1A.5$$

As the number of observations increases, from 25 observations and above, the last term in expression 1A.5 approaches 1:

$$\sqrt{\left(\frac{n-1}{n+1}\right)} \to 1, n \geq 25 \qquad 1A.6$$

Substituting the limiting value of expression 1A.6 into expression 1A.5 implies that as the number of observations increases the coefficient of variation of the ranks of the data is approximately equal to the reciprocal of the square root of 3, or:

$$c_i \cong \frac{1}{\sqrt{3}} \qquad 1A.7$$

Now, let's turn our attention to finding the coefficient of variation for the values of the data. Apart from deriving this, the traditional way as the ratio of the standard deviation to the mean of the values, the alternative is to obtain it by assuming that a regression is run through the origin between the values and the ranks of the data. The values are the dependant variable and the ranks are the independent variable. We know that for any regression between two variables, the product of the standard deviation of the independent variable – in this case σ_i – and the slope of the regression (β_1) is equal to the product of the correlation between the variables (ρ) and the standard deviation of the dependent variable, which in this case is σ_x. In short:

$$\sigma_i \beta_1 = \rho \sigma_x \qquad 1A.8$$

We also know that the correlation between values and ranks is the Stuart correlation, as well as that in terms of the solution to the Stuart inequality, its maximum value is 1, or:

$$\rho \leq 1 \Rightarrow \rho = 1 \qquad \text{1A.9}$$

Substituting expression 1A.9 into expression 1A.8, results in:

$$\sigma_x = \sigma_i \beta_1 \qquad \text{1A.10}$$

In turn substituting expression 1A.2 into 1A.10 leads to:

$$\sigma_x = \beta_1 \sqrt{\frac{\left(n^2 - 1\right)}{12}} = \beta_1 \sqrt{\frac{\left(n^2 - 1\right)}{4 \bullet 3}} = \frac{1}{2} \beta_1 \frac{\sqrt{\left(n^2 - 1\right)}}{\sqrt{3}} \qquad \text{1A.11}$$

As for the mean of the values, in the context of a regression, this is given by:

$$\mu = \beta_1 \mu_i \qquad \text{1A.12}$$

Substituting expression 1A.3 into 1A.12 leads to:

$$\mu = \beta_1 \left(\frac{n+1}{2} \right) \qquad \text{1A.13}$$

By taking the ratio between expressions 1A.10 and 1A.13, the coefficient of variation for the values of the data is given by:

$$c = \frac{\sigma_x}{\mu} = \frac{\frac{1}{2} \beta_1 \frac{\sqrt{\left(n^2 - 1\right)}}{\sqrt{3}}}{\frac{1}{2} \beta_1 \left(n+1\right)} = \frac{1}{\sqrt{3}} \frac{\left(n-1\right)^{\frac{1}{2}} \left(n+1\right)^{\frac{1}{2}}}{\left(n+1\right)^{1}} = \frac{1}{\sqrt{3}} \frac{\left(n-1\right)^{1/2}}{\left(n+1\right)^{1/2}}$$

$$= \frac{1}{\sqrt{3}} \sqrt{\left(\frac{n-1}{n+1} \right)} \qquad \text{1A.14}$$

Here too, as the number of observations increases, from 25 observations and above, the last term in expression 1A.14 approaches 1:

$$\sqrt{\left(\frac{n-1}{n+1} \right)} \rightarrow 1, n \geq 25 \qquad \text{1A.15}$$

Substituting the limiting value of expression 1A.15 into expression 1A.14, implies that as the number of observations increases, the coefficient of variation for the values of the data is also approximately the reciprocal of the square root of 3, or:

$$c \cong \frac{1}{\sqrt{3}} \qquad \text{1A.16}$$

The ratio between expressions 1A.16 and 1A.7 is clearly one:

$$\frac{c}{c_i} = 1 \qquad \text{1A.17}$$

Apart from Glasser's reminder (1961a: 179), we know very well that whenever the values of the data are exponentially distributed their coefficient of variation is 1, or:

$$1 = c \qquad \text{1A.18}$$

In essence expression 1A.17 tells us that the ratio between the coefficient of variation of the data's values and its ranks is always 1, but *only* when the data is exponentially distributed. This implies that by substituting expression 1A.18 into expression 1A.17 there is a general expression for that ratio, which is dependent on the value of the coefficient of variation as obtained from the data's values, or:

$$\frac{c}{c_i} = c \qquad \text{1A.19}$$

From expression 1A.19 we can see that if the coefficient of variation of the data's values is 1 we are back to expression 1A.17, in terms of which we know that the data is exponentially distributed. We would recall that in practice, for unimodal distributions, the upper limit of the coefficient of variation for the values of data is not strictly equal to 1, or:

$$0 \leq c < 1 \qquad \text{1A.20}$$

Indeed, we can see from expression 1A.16, that for *any other* distribution, i.e. other than the exponential distribution, the typical value for the coefficient of variation of the data's values is the reciprocal of the square root of 3 as their number of observations increases. Then by substitution of expression 1A.16 into expression 1A.19, followed by rearrangement, we obtain the De Vergottini inequality in terms of its limiting solution:

$$\frac{c}{c_i} = \frac{1}{\sqrt{3}} \Rightarrow c = \frac{1}{\sqrt{3}} c_i \qquad \text{1A.21}$$

Expression 1A.21 is that already referred to in expression 1.5 in the body of Chapter 2.

2.6 APPENDIX: SELECTED CRITICAL CHI-SQUARE VALUES

Table 1: Critical Chi-square values for 99%, 95%, and 90% confidence intervals

Degrees of freedom (df=n-1)	*Lower value* $\alpha = 1\%$	*Upper value* $\alpha = 1\%$	*Lower value* $\alpha = 5\%$	*Upper value* $\alpha = 5\%$	*Lower value* $\alpha = 10\%$	*Upper value* $\alpha = 10\%$
1	0.00	7.88	0.00	5.02	0.00	3.84
2	0.01	10.60	0.05	7.38	0.10	5.99
3	0.07	12.84	0.22	9.35	0.35	7.82
4	0.21	14.86	0.44	11.14	0.71	9.49
5	0.41	16.75	0.83	12.83	1.15	11.07
6	0.68	18.55	1.24	14.45	1.64	12.59
7	0.99	20.28	1.69	16.01	2.17	14.07
8	1.34	21.95	2.18	17.53	2.73	15.51
9	1.74	23.59	2.70	19.02	3.33	16.92
10	2.16	25.19	3.25	20.48	3.94	18.31
11	2.60	26.76	3.82	21.92	4.58	19.68
12	3.07	28.30	4.40	23.34	5.23	21.03
13	3.57	29.82	5.01	24.74	5.89	22.36
14	4.08	31.32	5.63	26.12	6.57	23.68
15	4.60	32.80	6.26	27.49	7.26	25.00
16	5.14	34.27	6.91	28.85	7.96	26.30
17	5.69	35.72	7.56	30.19	8.67	27.59
18	6.27	37.16	8.23	31.53	9.39	28.87
19	6.84	38.58	8.91	32.85	10.12	30.14
20	7.43	40.00	9.59	34.17	10.85	31.41
21	8.03	41.40	10.28	35.48	11.59	32.67
22	8.64	42.80	10.98	36.78	12.34	33.92
23	9.26	44.18	11.69	38.08	13.09	35.17
24	9.89	45.56	12.40	39.36	13.85	36.42
25	10.52	46.93	13.12	40.65	14.61	37.65
26	11.16	48.29	13.84	41.92	15.38	38.89
27	11.81	49.65	14.57	43.19	16.15	40.11
28	12.46	50.99	15.31	44.46	16.93	41.34
29	13.12	52.34	16.06	45.72	17.71	42.56
30	13.79	53.67	16.79	46.98	18.49	43.77
40	20.71	66.77	24.43	59.34	26.51	55.76
50	27.99	79.49	32.36	71.42	34.76	67.50
60	35.53	91.95	40.48	83.30	43.19	79.08
70	43.28	104.21	48.76	95.02	51.74	90.53
80	51.17	116.32	57.15	106.63	60.39	101.88
90	59.20	128.30	65.65	118.14	69.13	113.15
100	67.33	140.17	74.22	129.56	77.93	124.34
120	83.85	163.65	91.57	152.21	95.70	146.57
$df \geq 240$	187.32	300.18	198.99	284.80	205.14	277.14

3. EVALUATIVE DISCUSSION OF THE HHI

3.1 REASONS FOR REFORMULATION

In the previous chapter we found that the main ingredient of the HHI is the coefficient of variation, from which it was established that the Chi-square distribution is the sampling distribution of the HHI. We found that this makes it possible to construct confidence limits for the HHI, which in turn can be used to determine its magnitude, formulate hypotheses tests for the HHI, as well as study the accuracy of its estimates. However this umbilical connection with the coefficient of variation, invariably also opens the door for prospective criticism on the HHI similar to those directed at the coefficient of variation. As Bronk (1979: 665) observes:

> "the coefficient of variation, which is defined as the ratio of the standard deviation to the average for a probability density function has been called 'not of great interest', 'misleading' or not 'suitable for advanced work' by various statisticians."

This concern with the coefficient of variation has not changed. Hürlimann (1998: 128) provides the same, if not more detailed account, of why it exists:

> "One must warn against blind application. The measure has been built starting from the variance, and there is almost general agreement that the variance is appropriate for…measurement only for normal (or approximately normal) distributions."

While Hürlimann emphasises the variance, the same can be said of the mean too. As we know the two cease to accurately estimate scale and location respectively when the data carries outliers and has a non-symmetrical distribution. A famous illustration of this is provided by Iglewicz (1983: 404-411). It follows then that the coefficient of variation whose components include the mean and the standard deviation should be expected to suffer loss of accuracy in the measurement of relative variability when the data are characterised by a non-normal distribution and extreme observations. Practically this then includes just about any case. After all in reality the instances of sanitised data

are few and far between. One cannot help but be reminded on this matter by Geary's famous finding that in practice (Geary, 1947: 241):

> "normality is a myth; there never was, and never will be, a normal distribution".

Tomkins (2004: 24, 26) cites five other empirical investigations done between 1971 and 2004 that corroborate Geary's finding. In the present case there is persistent evidence showing a parallel picture. For instance, as mentioned already, Axtel (2001: 1819-1820) and subsequently Gaffeo *et al*. (2003: 119-121) find that the typical distribution of firms' market shares is positively skewed, i.e. non-symmetrical, and is also prone to contain extreme observations depending on how large the market share gap is between the top and bottom firms.

What then is the alternative or accurate measure of relative variability under such conditions?

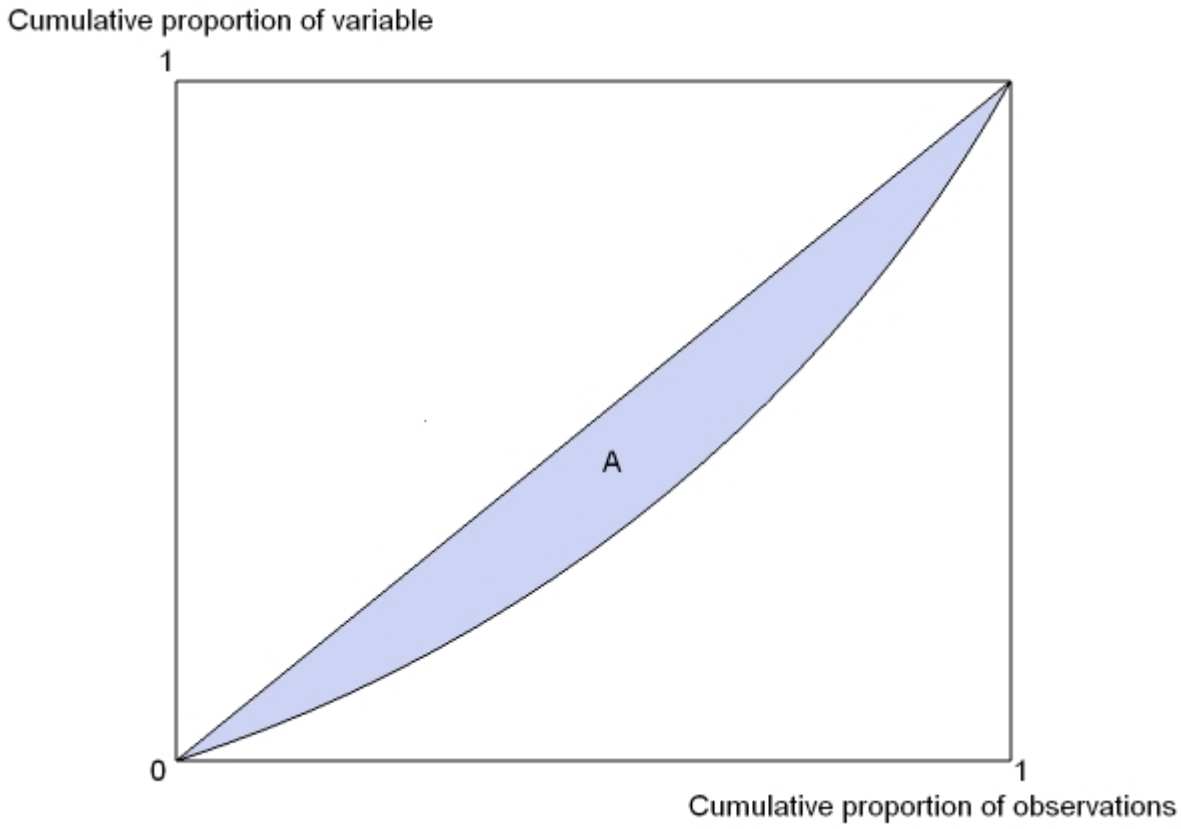
It turns out that it is none other than the well-known Gini index (G). In a certain sense this is not unexpected. As Sharma (2010: 119) reminds us the two best known measures of relative variability are the coefficient of variation and the Gini index. So if we cannot make use of the coefficient of variation we would gravitate towards the Gini index.

The Gini index is twice the Lorenz area (A). Figure 1 gives a graphical illustration of this. The Lorenz area is the ratio of the covariance (cov) of the observation values (x_i) with their ranks (i), to the product of their number (n) and mean (μ), or:

$$G = 2 \bullet A = 2 \bullet \frac{\mathrm{cov}(x_i, i)}{n\mu} \qquad 2.1$$

Various accounts of this formulation for the Gini index appear in Gini (1914 [2005]: 23-27; 1921: 125; 1947: 24; 1965: 94-95), De Vergottini (1950: 453), Kendall and Stuart (1977: 48), Lerman and Yitzhaki (1984: 365), and Olkin and Yitzhaki (1991: 385).

Figure 1: The Lorenz curve



As we see from figure 1, the Lorenz area is an area-within-an-area, because it resides within a rectangle. If it inscribes none of its dimensions its value is zero. If it inscribes half of its dimensions its value is one-half. Consequently the range of the Lorenz area is between zero and one-half:

$$0 \leq \frac{\text{cov}\left(x_i, i\right)}{n\mu} \leq 0.5 \qquad 2.2$$

As a result the Gini index takes on a minimum value of 0 and a maximum value of 1:

$$0 \leq G \leq 1 \qquad 2.3$$

By comparing expressions 2.1 and 2.3 it is easy to see that the Gini index is a measure of relative variability because it compares the covariance of the data relative to its mean. At its minimum value of zero there is no relative variability in the data because its covariance is nil. At its maximum value of one, the relative variation in the data is extreme because its values are completely different from each other when compared to the mean.

Furthermore, we also know from Glasser (1961a: 177, 179-180), that the distributional behaviour of the Gini index coincides with that of the coefficient of variation, such that:

a) When its value is zero its sampling distribution is uniform;
b) When its value is one its sampling distribution is exponential;
c) For any values in-between its sampling distribution is indeterminate in the sense that it too can take on any positively skewed unimodal form.

In practical situations the limits of the Gini index are known not to be so strict, and instead its values tends to lie somewhere between its lowest and highest limits. As a result the exact probability distribution of the sample Gini index is unknown except for knowing that its shape is unimodal and positively-skewed (McDonald, 1981: 168-169). This makes the Chi-square distribution a natural candidate because it fulfils these shape requirements. Gerstenkorn and Gerstenkorn (2003: 470-471) report that Kamat (1953: 452; 1961: 170, 172-174) and Ramasubban (1956: 120-121; 1959: 223) have exploited this to show that, like the coefficient of variation, the sampling distribution of the Gini index is approximated by the Chi-square distribution. In time to come, we will see in this chapter why this is so.

There are a number of other recollections we also need to make. It is definitively known from Gini, circa 1912, as cited in Ceriani and Verme (2011: 19), as well as from Glasser (1961b: 399-400; 1962b: 652-653), that:

a) Whenever the sampling distribution of the data is skewed the measurement of relative variability by the Gini index is more accurate than that by the coefficient of variation;
b) Whenever the skewness fades the two are equally accurate.

As described by Piesch (2005: 264-265) the aforementioned results are mathematically expressed as a one-sided, non-strict inequality, in terms of which the maximum value of the Gini index equals that of the coefficient of variation of the observed values:

$$G \leq c \qquad 2.4$$

Although Gini may have drawn attention to the existence of expression 2.4 earlier, it is actually formally known as the Glasser inequality for the Gini index. It provides an upper bound for the index with the coefficient of variation (Piesch, 2005: 264). Respectively expression 2.4 will also be referred to as the Glasser inequality throughout the text.

By rewriting the Glasser inequality, we can immediately see that we are dealing with an asymptotic equality:

$$\frac{G}{c} \leq 1 \Rightarrow G = c \qquad 2.5$$

Because the ratio between the Gini index and the coefficient of variation converges to 1 as the number of observations increases, both measures of relative variability begin to behave alike as the number of observations grows. This is why the two measures are ultimately equated. The equality is the solution between them for the maximum value the Gini index assumes in the data. This value is none other then that for the coefficient of variation from the values of the observations.

By contrast, when few observations are encountered, as per expression 2.4, it follows that the coefficient of variation systematically exceeds the Gini index when measuring relative variability. In short the difference between them is always positive:

$$c - G \geq 0 \qquad 2.6$$

The practical existence of this inequality has been empirically confirmed by Bendel *et al*. (1989: 395-398). Gini, circa 1912, as cited in Ceriani and Verme (2011: 19), gave the following account of its practical implications:

> "Given that a constant relation exists between different variability indices …, the choice of a specific index to compare variability for different series of infinite observations is irrelevant. But … when the number of observations is limited, which in practice is the most frequent case, … the choice of the index becomes relevant."

Simply stated, with a growing number of observations the Gini index and the coefficient of variation are equal. However the equality between them breaks down when the number of observations decreases. In such cases, as revealed by expression 2.6, the coefficient of variation will be a biased measure of relative variability – yielding consistently inflated estimates – when compared to the Gini index. This in a nutshell is the underlying statistical reason why, as pointed out by Bronk, the relative standard deviation is seen as a measure of trivial practical potential; something that is potentially misleading; or altogether potentially unsuitable for measuring relative variability in practice.

In essence expression 2.6 exposes the otherwise hidden dilemma of having to decide on the robust estimation of relative variability. There are many explanations of what robustness is but a neat and handy one is that provided by Morgenthaler (2007: 272, 277-278), to the effect that it is the property of an estimator to retain its accuracy when the ideal conditions for which it is designed begin to disappear or no longer exist. In practice this is achieved by making such an estimator from the ranks, range, median, median absolute deviation, inter-quartile range, or any such quantity of the data, that can evade outliers and skewed distributions (Conover and Iman, 1981: 124). It is customary to refer to such an estimator as non-parametric to differentiate it from a parametric one, which is not composed of these elements or such quantities. While clearly the coefficient of variation is not a robust estimator of relative variability, we can see clearly from expression 2.1 that by virtue of including the ranks in its formulation the Gini index is.

Generally, as Tomkins (2006: 25) reminds, in practice, a robust estimator should be used when the data conditions are imperfect. This is when:

a) The distribution of the observations is non-symmetrical;

b) The distribution has outliers;

c) The number of observations is few; and lastly

d) There is considerable variation among the observations at hand.

The converses of the aforementioned are known as perfect data conditions. The crux of Geary's finding is that in practice coming across such conditions is the exception rather than the norm. In this case Gini is correct to argue that the choice of the measure that can accurately detect relative variability is relevant. This is at the heart of the Glasser inequality. If we want robustness, as expression 2.6 tells us, we have to choose the Gini index. Following Tomkin's advice we could make this decision based on whether the data conditions are imperfect. But there is no need to. This is because the robust estimator is accurate even if data conditions are perfect (Jacobson, 1970: 268; Croux and Filzmoser, 2007: 282). A diverse and multidisciplinary collection of empirical evidence from 1931 to date assembled and surveyed by Morgenthaler (2007: 272-278) shows this being systematically confirmed. Another collection of evidence from 1970 to date is referred to in Tomkins (2006: 24-25). To this list of studies we may also add that of Stigler (2010: 277-278) confirming the same. The gist from all of these studies is that, as a matter of principle, robust estimators keep their accuracy irrespective of the data conditions they are exposed to. Morgenthaler's explanation of what is robustness makes it easy to understand why. If by design robust estimators are accurate under imperfect data conditions, then just because these conditions disappear it does not mean that their accuracy disappears. It stays. Instead it is the data conditions that change. We too have confirmation of this in our present case by virtue of the asymptotic equality between the Gini index and the coefficient of variation as per expression 2.5. This immediately implies that if we wish to accurately measure business concentration with the HHI, then we have to revise its formulation to replace the coefficient of variation with the Gini index. However, to do that, we need to prove beyond a reasonable doubt that there is an equality between them. Certainly knowing from expression 2.5 that it exists gives us a good starting point. However this is not enough unless we can replicate it. It is this that will show that the replacement is not optional but necessary because by implication it also proves the adverse consequences of

not doing so. It is easy to infer from the asymptotic equality between the measures that if the HHI retains the coefficient of variation it will produce biased, i.e. consistently inflated values of business concentration relative to housing the Gini index instead. Furthermore the HHI will suffer no loss of accuracy if it permanently keeps the Gini index in its formulation because as we know from the asymptotic equality this measure of relative variability is the same as the coefficient of variation when the number of firms (or observations) over which it is measured increases.

3.2 REFORMULATION

Reformulating the HHI is by no means easy. Our problem to accomplish this is succinctly captured by Marron (1999: 68):

> "The mathematical statistics literature has many papers that are unnecessarily difficult to read."

Shalit (1985: 185, 188), Gerstenkorn and Gerstenkorn (2003: 469-471), and more recently Gerber (2007: 133), find that the same problem tends to be characteristic of statistical discussions on the Gini index. They find that the mathematical treatment involved is heavy and inaccessible, as can be seen for instance in the works by Glasser (1961b: 397-400; 1962b: 652-653), Kendall and Stuart (1977: 48), Yitzhaki (1998: 23-24), and Piesch (2005: 268-269, 282, 284). Against this mainstream the approach taken here is to derive a simplified proof. As Schaaf (1951: 22) observed long ago:

> "…the use of a minimum amount of basic algebra might very well simplify many … procedures … and at the same time make them more meaningful."

It turns out that it is possible to use minimum algebra to come up with a simple proof of the asymptotic equality between the Gini index and the coefficient of variation. This will also make it more meaningful to appreciate why these measures of relative variability are substitutes for each other to the extent that the Gini index is the preferable one to use. The clue for this comes from

Milanovic (1997: 45-46) who unfortunately provides an incomplete derivation of this result. Our job here is to complete Milanovic's derivation.

To begin with, we know that the covariance between any two variables is the product of the correlation coefficient between them (ρ) and their respective standard deviations, which in our case are the values of the observations and their ranks:

$$\text{cov}\left(x_i, i\right) = \rho\sigma_x\sigma_i \qquad 2.7$$

By substituting expression 2.7 into expression 2.1, in the same way as De Vergottini (1950: 453), Milanovic (1997: 45) rewrites the latter expression for the Gini index, to:

$$G = \frac{2}{n\mu}\sigma_x\rho\sigma_i \qquad 2.8$$

We already know from the variance of the ranks, as per expression 1.7, that their standard deviation is given:

$$\sigma_i = \sqrt{\frac{1}{12}\left(n^2 - 1\right)} \qquad 2.9$$

Inserting expression 2.9 into expression 2.8, followed by re-arrangement, yields the following expression for the Gini index:

$$G = \frac{2}{n\mu}\sigma_x\rho\frac{\sqrt{\left(n^2-1\right)}}{\sqrt{12}} = \frac{\sqrt{4}}{\sqrt{12}}\frac{\sigma_x}{\mu}\rho\sqrt{\frac{\left(n^2-1\right)}{n^2}} = \frac{1}{\sqrt{3}}c\rho\sqrt{\frac{\left(n^2-1\right)}{n^2}} \qquad 2.10$$

We can see that the square root term for the number of observations approaches 1 as the number of observations increases. This limit is approached quickly from as few as 4 observations. That is:

$$\sqrt{\frac{\left(n^2-1\right)}{n^2}} \to 1, n \geq 4 \qquad 2.11$$

Substituting the limiting value of expression 2.11 into expression 2.10, Milanovic (1997: 46) reduces the latter expression for the Gini index to:

$$G \cong \frac{1}{\sqrt{3}} c\rho \qquad 2.12$$

At this stage Milanovic considers the task completed, and concludes that a simple way to calculate the Gini index has been derived (Milanovic, 1997: 49). This conclusion is close but is not where we need to be. To see this, note that the correlation coefficient refers to the correlation between observation values and their ranks. It is none other than the familiar Stuart correlation, also discussed in the appendix to this chapter. There is an explicit, widely known result for this correlation, known as the Stuart inequality. It stipulates that the maximum value this correlation takes is one (Stuart, 1954: 37, 42):

$$\rho \leq 1 \Rightarrow \rho = 1 \qquad 2.13$$

Recently, Maturi and Elsayigh (2009: 16, 18) have empirically confirmed Stuart's inequality. O'Brien has also done the same previously (1982: 148-151). Kendall – as cited in Jacobson (1970: 267) – gave the following outline of its implications:

> "in virtue of this fairly close relationship between ranks and variates we might expect that if we replace variate values by rank-numbers and then operate on the latter as if they were primary variates we should in many cases draw the same conclusions".

This type of replacement is the essence of the now famous rank transformation advocated by Conover and Iman (1981: 124). It is due to Stuart's inequality that we know that by replacing the values of the observations with their ranks the accuracy of the consequent computations is unaffected, or as Stuart (1954: 37) put it:

> "there is justification for replacing the original observations by their ranks, since the consequent saving in computation entails very little loss of efficiency."

The implication of Stuart's inequality for the present case is that we can replace the correlation coefficient between observation values and ranks by its maximum value, and by doing so simultaneously move to working with the ranks of the data. This in turn implies rewriting expression 2.12 in terms of the ranks of the data:

$$G \cong \frac{1}{\sqrt{3}} c_i \qquad 2.14$$

An empirical confirmation of the approximate equality by expression 2.14 is provided by Glasser (1961a: 177) who also shows that it actually converts to a strict equality as the number of observations increases:

$$G = \frac{1}{\sqrt{3}} c_i \qquad 2.15$$

Piesch (2005: 269) regards expression 2.15 as an extension of the De Vergottini inequality as set out in expression 1.5. It is not hard to see why. The right hand side of expression 2.15 is in fact the De Vergottini inequality. From expression 1.5, we know that the solution for this inequality in terms of its maximum value is the coefficient of variation when obtained from the values of the data. Thus, by substitution of expression 1.5 into expression 2.15, we can rewrite the latter expression into an equality between the Gini index and the coefficient of variation:

$$G = c \qquad 2.16$$

With expression 2.16 we have proved the Glasser inequality for the Gini index, namely that in the limit as the number of observations increases the Gini index and coefficient of variation are equal.

In 2006, Sawilowsky (2006: 627-628) published empirical results of Monte Carlo simulations showing that as the number of observations increases the maximum value of the Gini index is 33% (or 0.33 percentage points). Sawilowsky gave no explanation for this, but we can see that the finding is not accidental. It gives the limiting value of the coefficient of variation for the observations of the data as per expressions 1.11 and 1.13. Because of the equality in expression 2.16 the same

value is then take up by the Gini index. Alternatively, it can also be obtained from the ranks of the data, by substituting expression 1.10 into expression 2.15:

$$G = \frac{1}{\sqrt{3}} \bullet \frac{1}{\sqrt{3}} \sqrt{\left(\frac{n-1}{n+1}\right)} \qquad 2.17$$

The last term in expression 2.17 converges to 1 with 25-or more observations:

$$\sqrt{\left(\frac{n-1}{n+1}\right)} \to 1, n \geq 25 \qquad 2.18$$

Substituting the limiting value of expression 2.18 into expression 2.17 implies that as the number of observations increases the Gini index becomes approximately one-third:

$$G \cong \frac{1}{3} \qquad 2.19$$

Piesch (2005: 284) refers to this result as the other important special case of the De Vergottini inequality. It complements the earlier observation that practically the Gini index is known to take on values that lie somewhere in between its minimum and maximum limits. As the number of observations increases the value of the Gini index tends to one-third. Since the value of expression 2.19 is the same as that of 1.13, it follows that the ratio between them is one, thereby confirming the asymptotic equality between the Gini index and the coefficient of variation. By virtue of replicating this equality, the alternative solution of the Glasser inequality for the Gini index is also established, namely that in the case of fewer observations the coefficient of variation will overstate the actual level of relative variability when compared to the Gini index.

The proof that the Gini index is the same as the coefficient of variation when the number of observations increases implies, firstly, that we can extend McKay's approximation and, secondly, that we can reformulate the HHI. Given that the solution to the Glasser inequality is the equality between the Gini index and the coefficient of variation, it follows immediately that McKay's

confidence intervals automatically apply to the Gini index. From expressions 1.19 and 2.16, the resultant McKay confidence interval for the Gini index is:

$$\Lambda_1^G = \left(\frac{G}{\sqrt{\left| G^2 \left(\frac{X_u^2}{n} - 1 \right) \right| + \frac{X_u^2}{n-1}}}, \frac{G}{\sqrt{\left| G^2 \left(\frac{X_l^2}{n} - 1 \right) \right| + \frac{X_l^2}{n-1}}} \right) \qquad 2.20$$

In turn from expressions 1.20 and 2.16, McKay's modified confidence interval for the Gini index is:

$$\Lambda_2^G = \left(\frac{G}{\sqrt{\left| G^2 \left(\frac{2 + X_u^2}{n} - 1 \right) \right| + \frac{X_u^2}{n-1}}}, \frac{G}{\sqrt{\left| G^2 \left(\frac{2 + X_l^2}{n} - 1 \right) \right| + \frac{X_l^2}{n-1}}} \right) \qquad 2.21$$

Expressions 2.20 and 2.21 are completely new findings. McKay's approximation and its confidence intervals have always been associated with the coefficient of variation. As it turns out they are just as applicable to the Gini index as a by-product of the Glasser inequality. In short McKay's confidence intervals extend to the Gini index. This is an important advancement. While McKay's approximation is known to be accurate, this is valid only for perfect data conditions. This is because, as we would recall, the coefficient of variation is built from the mean and the variance, which are the appropriate measures of location and scale only when these conditions are fulfilled. When this is not so, McKay's approximation will become inaccurate. It will produce inflated estimates of the actual levels of relative variability in the data because the coefficient of variation overstates its measurement in such cases. In essence the present extension of McKay's approximation ensures that it does not loose its accuracy or relevance when the data conditions become imperfect. As a consequence of the Glasser inequality, by replacing the coefficient of variation with the Gini index, McKay's approximation comes to keep its accuracy under such conditions while also changing nothing about its accuracy when the data conditions are perfect.

This extension also helps resolve a conundrum about Kamat's and Ramasubban's findings. Neither Kamat (1953: 452; 1961: 173-174) nor Ramasubban (1956: 120-121; 1959: 223) explained why they found the Chi-square distribution to be a reliable approximation for the sampling distribution of the Gini index, except to emphasise that it is an empirical regularity. Of course from the proof for the Glasser inequality, we know that both the Gini index and the coefficient of variation are directly proportional to each other. By extension it follows that both will have the same approximate sampling distribution. In addition, although Kamat and Ramasubban identified the sampling distribution of the Gini index, they could not provide its confidence intervals. The extension of McKay's approximation bridges that gap, and also highlights a possible improvement in data analysis involving the Gini index. This is because McKay's approximation gives two diagnostic tools, i.e. the confidence intervals as per expressions 2.20 and 2.21, by which to do this analysis. So far the practice in areas such as economics, business administration, competition regulation, and even business statistics is to report and analyse the Gini index descriptively as a single number. The extension of McKay's approximation now makes it possible to report on the range of its confidence limits also, which in its own right facilitates the use of the Gini index for hypothesis testing, as well as making it possible to assess the accuracy of its estimates.

Regarding the reformulation of the HHI, it simply involves replacing the coefficient of variation in expression 1.1 with the Gini index, such that:

$$\mathrm{HHI} = \frac{G^2 + 1}{n} \qquad 2.22$$

On account of the existence of the Glasser inequality, expression 2.22 prevents the HHI from being a potentially-biased measure of business concentration. It is clearly an extension of the Glasser inequality in the sense that any secondary measure of relative variability – like the HHI – which contains the coefficient of variation as its primary component, by association becomes subordinated to the consequences of its solution. In the limit, as the number of observations increases, such a measure will accurately quantify the observed relative variability because it contains the Gini index. It will do the same with fewer observations provided that the Gini index is

not replaced by the coefficient of variation given that in such cases the latter will give systematically higher measurements in relative variability than actually exist.

The reformulation of the HHI by the Gini index preserves its range. To see this we may want to recall that the Gini index is composed in part of the covariance between the ranks of observations and their values. In the case of a single observation ($n = 1$) there is no co-variation between rank and value, and then the Gini index is zero. In that event the HHI reaches its maximum of one. *But the Gini index is also zero when, irrespective of the number of observations, their values are exactly the same, i.e. literally all observations become stacked next to each other. Then there is no co-variation. In that case the HHI reaches its minimum given by the reciprocal number of observations*. To clarify, the number of observations actually refers to the number of firms in a market or industry depicted by their market shares.

More importantly, the reformulation indicates that in the case of the HHI the conventional distinction in economics between absolute and relative measures of concentration is redundant, simply because it does not exist for this index. For example the 2006 edition of the *Collins Dictionary of Economics* notes that:

> "Concentration measures, like…the Herfindahl index, are known as absolute concentration measures since they are concerned with the market shares of a given (absolute) number of firms. By contrast, relative concentration measures are concerned with inequalities in the share of total firms producing for the market. Such irregularities can be recorded in the form of a Lorenz curve."

On the basis of expression 2.22, we can conclude that in the case of the HHI this supposed distinction is contrived because in fact the supposedly different measures of concentration are conceptually inseparable. To validate this let's rewrite expression 2.22 to:

$$G^2 = nHHI - 1 \Rightarrow G = \sqrt{nHHI - 1} \qquad 2.23$$

Expression 2.23 shows that the Gini index is the square root of the product of the number of observations and the HHI index subtracted from 1. In short the HHI reproduces the reflection of the Gini index about the Lorenz curve. In effect it replicates the symmetry of the Lorenz curve. Thus that curve also graphically describes the HHI.

So far there is no awareness that the presented reformulation exists or that it is necessary. For example Fedderke and Szalontai (2009: 242-245), and subsequently Fedderke and Naumann (2011: 2920), reported that they held information on the Gini index and the number of firms, but proceeded to conclude that the HHI cannot be computed from such data. On the contrary it can. In addition we can also obtain confidence intervals for these estimates. By substitution of expression 2.23 into expression 2.20, McKay's confidence interval for the HHI is:

$$\Lambda_1^{HHI} = \left(\frac{\sqrt{nHHI - 1}}{\sqrt{(nHHI - 1) \bullet \left| \left(\frac{X_u^2}{n} - 1 \right) \right| + \frac{X_u^2}{n-1}}}, \frac{\sqrt{nHHI - 1}}{\sqrt{(nHHI - 1) \bullet \left| \left(\frac{X_l^2}{n} - 1 \right) \right| + \frac{X_l^2}{n-1}}} \right) \qquad 2.24$$

By substitution of expression 2.23 into 2.21, McKay's modified confidence interval for the HHI is:

$$\Lambda_2^{HHI} = \left(\frac{\sqrt{nHHI - 1}}{\sqrt{(nHHI - 1) \bullet \left| \left(\frac{2 + X_u^2}{n} - 1 \right) \right| + \frac{X_u^2}{n-1}}}, \frac{\sqrt{nHHI - 1}}{\sqrt{(nHHI - 1) \bullet \left| \left(\frac{2 + X_l^2}{n} - 1 \right) \right| + \frac{X_l^2}{n-1}}} \right) \qquad 2.25$$

We can see that expressions 2.24 and 2.25 are exactly the same as expressions 1.26 and 1.27 respectively. This is not surprising because as shown already, due to the Glasser inequality, McKay's approximation extends to the Gini index also. The only difference is that the current expressions are derived from the Gini index. However this derivation route is important because it shows that if we want the HHI to retain the accuracy it receives from the Gini index, it must in turn be computed from the Gini index. This is the only way by which in practice the HHI will adhere to the Glasser inequality. This will occupy our attention in the upcoming chapter. Before we move to

there, we should note that our confidence intervals for the HHI are two-sided. We would recall from expression 1.18 that the HHI has a closed range with a minimum value as well as a maximum value. Resultantly a one-sided confidence interval for the HHI will presume that the HHI range is not two-sided. Any such interval will then imply that the index has an open-ended range. For example if such an interval is constructed *from* the minimum value it would mean that the HHI is without a maximum value. Conversely, if such an interval is constructed *up to* the maximum value it would mean that the HHI is without a minimum value. As Smithson (2000: 156) points out:

> "One-sided confidence intervals are used mainly when the researcher has no reason to be interested in values in one of the distribution's tails or when such values would have no meaning."

As we have seen in Chapter 2, the minimum and maximum values of the HHI do have contextual meaning for economists, business executives, competition regulators, as well as statisticians. This makes it inconsistent to pay attention to the one limit without the other. On the statistical side there is more. As Cumming and Finch (2001: 533-534, 543-544) remind, generally a confidence interval for a statistic is the *range* of values that contains a specified percentage of the sampling distribution of the statistic. In turn the confidence level represents the expected percentage of times that the confidence interval would contain the population value of the estimated statistic *under repeated random sampling*. By way of an example, consider a 95% confidence interval from anyone of the discussed intervals for the HHI. Such an interval tells us that if random sampling is repeatedly done, 95% of the confidence intervals constructed from each sample will include the population value of the HHI. This value is best captured by the central part of anyone of these confidence intervals rather than by either of their extremes, because the central part is in fact the average estimate for the population value. In short values that are closer to the centre of a confidence interval are *more* plausible than those further away from the centre. Plausibility simply means that the estimates of a confidence interval – from its minimum through to its maximum value – are likely candidates for the population value. The most likely of these is in the centre of the confidence interval. This is the average or expected value. The average value is the expected

value because from among the plausible values revealed by the limits of anyone of the confidence intervals it is the *most* plausible. Here we also need to recall two other things.

Firstly the expected value or average estimate is accurate only if there is no bias, or if bias creep is limited. This is because in a strict statistical sense, accuracy is precision without bias (Grubbs, 1973: 54-56, 66). We can tell precision from the width of the confidence interval, as the width represents the largest error of estimation we are likely to make with the sample size at hand when deriving an estimate for the population value of a statistic. We can also tell what the bias is, because as seen in expression 1.22, it is half the width of the confidence interval. So from a confidence interval we can comfortably establish if the estimate for a population value is accurate. This is why, as Smithson (2000: 185) observes:

> "Contrary to long-standing traditions in … research, confidence intervals should be routinely presented in research reports, perhaps (but not always) in conjunction with significance tests."

Smithson's suggestion that the joint reporting of confidence intervals and significance tests is optional is not surprising since a significance test is also implementable with a confidence interval *provided* such an interval is reported.

Secondly, the estimated expected value arising out of a confidence interval is a hypothetical value in the sense that if accurate it shows what is the *potential* population value from repeated random sampling were such sampling to be done. In any given sample situation, we calculate only one such interval. This interval is an interval, i.e. range estimate for the population value. We have already seen this from our expressions for the HHI confidence intervals, which need a sample estimate, around which the confidence interval is constructed. As a consequence, as Cumming and Finch (2005: 171) emphasise, whatever confidence interval we end up working with, it presupposes a sequence of potential confidence intervals, a proportion of which as depicted by its

confidence level will include the population value. It is in this sense that this level represents the chance that the confidence interval includes the population value, as highlighted earlier on.

Because confidence intervals characterise a range, they are typically two-sided. That said, one-sided confidence intervals beginning at some minimum value or ending at some maximum value can still be constructed *provided* that there are legitimate reasons for this. As indicated earlier on, this means that there is no reason to be interested in either of the extreme values of a statistic, or alternatively when such values have no meaning. In spite of this Bender *et al*. (2005: 238) find that in statistical practice:

> "… one-sided confidence limits are … less frequently used. If confidence intervals are presented, they are almost always two-sided, even in cases in which one-sided … are used."

This is not accidental, and the reasons for this are well known. Cai (2005: 64) summarises them as follows:

> "Although there are some common features, the one-sided interval estimation problem differs significantly from the two-sided problem. In particular, … the one-sided … interval does not perform well … both in terms of coverage probability and expected length."

To recall, the reason why the coverage probability of a one-sided interval performs poorly in relation to that of a two-sided interval is because the choice for a one-sided confidence interval at some stipulated significance level is the same as that for a two-sided confidence interval with double the significance level. So for example, if we start with a 5% significance level, and want a 95% confidence interval for the HHI starting at some minimum value or ending at some maximum value, we are actually constructing a two-sided 90% confidence interval at a 10% significance level either of whose limits are the one-sided limit for the 95% confidence interval. This is because the corresponding chance of occurrence for the range estimate of the two-sided 90% confidence

interval is 90%. Splitting the remaining 10% means that there is an additional 5% chance that the estimate for the population value is below the interval's minimum value, as well as a 5% chance that it is above the interval's maximum value. So by combining the 90% chance with the 5% chance for either of the interval's extremes we derive a one-sided 95% confidence interval. However this does not change the fact that in effect we are operating with a *lower* coverage probability for a two-sided interval, which also has a *higher* error rate for the interval estimates of the HHI. It is this increase in the error rate, i.e. the probability that the interval estimates are wrong, that also diminishes confidence in their precision. As a result of this, the expected length or width of the interval, which as we already noted is a measure of precision, is likewise diminished because the precision of its starting and ending estimates is already lessened. There is thus no surprise in Cai's reminder (2005: 64) that *unless legitimately used*:

> "… one-sided … intervals suffer a pronounced systematic bias in … coverage, although the severity and direction differ."

It is this propensity for bias in one-sided confidence intervals that is avoided by the two-sided confidence interval. This why Bender *et al*. (2005: 238) find that in statistical practice the two-sided confidence interval is the interval estimate of choice.

3.3 SUMMARY

In the present chapter we have effectively provided a rejoinder to Mandelbrot's (1997: 215-216) critique of the HHI. For emphasis, the critique is stated once again:

> "This index has no independent motivation, and … it is odd that it should ever be mentioned in the literature, even solely to be criticised because it is an example of inconsiderate injection of a sample of second moment in a context where … the existence of expectation is controversial. …According to reports, Herfindahl's index is taken seriously in some publications. This is hard to believe."

From the demonstrations in the current chapter we can see that in its original depiction the HHI is certainly inconsiderately injected with the coefficient of variation. Indeed the coefficient is an example of a second moment of the sampling distribution of the data, which on account of the Glasser inequality, is known to yield biased estimates of relative variability. This is in the sense that the measure systematically overstates the relative variability of the data except when the number of observations is large. But in practice perfect data conditions are the exception – not the norm. The present case, involving the distribution of market shares, is not different. Here positively skewed data with or without extreme observations is the norm. In that case it does not help in any way to know that the coefficient of variation has an expected value or an expected range. Their existence becomes controversial because essentially under these conditions the coefficient of variation has already ceased to be an accurate measure of relative variability.

Because of the Glasser inequality, we know that the Gini index is an immediate substitute that can rectify this measurement, because irrespective of the data conditions, it produces accurate estimates of relative variability. It does so when there are only a handful of observations, as well as when their number grows, at which time its measurements are equal to those from the coefficient of variation. This outcome of the Glasser inequality carries through in terms of its consequences across any other secondary measure that incorporates the coefficient of variation in its formulation. As a result of this, the HHI was reformulated to include the Gini index instead. The expression that results has not been previously discovered. En route to this reformulation another surprising and welcome result also emerged. It was found that by proving the Glasser inequality in terms of its solution, which in the limit is the equality between the Gini index and the coefficient of variation, McKay's approximation for the coefficient of variation automatically extends to the Gini index also. This too is a new finding, as historically the approximation has always been associated with the coefficient of variation. It can now also be associated with the Gini index, the sampling distribution of which is already known to be described by the Chi-square distribution as per Kamat's and Ramasubban's findings. This extension is complementary to these findings because it now gives the confidence limits for the Gini index from the Chi-square distribution, which neither Kamat nor Ramasubban were able to provide. While they presented their findings as an empirical regularity,

which as reported by Gerstenkorn and Gerstenkorn (2003: 470) has subsequently come to be considered in that way, the Glasser inequality shows that it is in fact due to the coefficient of variation and the Gini index being asymptotically equal. In response, their approximating sampling distribution is the same. This improves the analysis of data for the Gini index, which can now be carried out together with McKay's confidence intervals as opposed to the currently prevalent practice where its confidence limits are not reported at all.

By association, the subordination of the sampling distribution of the HHI to the Chi-square distribution in terms of McKay's approximation – whether reached by the coefficient of variation or the Gini index – is also an extension of the Glasser inequality. It ultimately diminishes the need for simulation studies that seek to find what this distribution is. To be sure such studies can be done, but they should be treated or handled as secondary findings because they would not tell us anything that we do not already know from the asymptotic equality between the Gini index and the coefficient of variation. For completeness such a demonstration will be provided in Chapter 5. For now we should keep in mind that the consequential improvement in accuracy the HHI receives from the inclusion of the Gini index also highlights the need to find a computational expression by which to maintain this accuracy.

3.4 APPENDIX: STUART'S CORRELATION

The Stuart correlation has been integral to the derivation of the De Vergottini inequality and the asymptotic equality between the Gini index and the coefficient of variation. Thereby it has also had a significant hand in the reformulation of the HHI in terms of the Gini index. Consequentially this appendix discusses this correlation in more detail in terms of the Stuart inequality. Stuart (1954: 39-40) himself provided a very condensed proof of this inequality, summarising it in three steps of integration. Jansen (1992: 364-367) has shown, that by integration, the full proof actually takes thirty steps to complete. O'Brien (1982: 148-151), and Maturi and Elsayigh (2009: 16, 18) were not concerned with a theoretical proof of the inequality. They verified it empirically. By contrast, in the present case, a new theoretically simplified derivation of this inequality is offered, as a way of helping to clarify its existence.

Simply stated, the Stuart correlation is the product of the familiar Pearson correlation between the data's values and its ranks, and the range of the latter's relative values. We will now see how.

David (1968: 574), and Olkin and Yitzhaki (1992: 187), indicate that the Stuart correlation (ρ) is given by:

$$\rho = \sqrt{\frac{3}{4}\left(\frac{n-1}{n+1}\right)} \bullet \frac{4\,\mathrm{cov}\left(x_i, i\right)}{n\sigma_x} \qquad 2A.1$$

In expression 2A.1, the covariance (cov) is between the data's values (x_i) and its ranks (i), while σ_x is the standard deviation of the values.

As reminded by Sharma (2010: 307), we know that the covariance between any two variables is the product of the Pearson correlation (ρ_p) and their respective standard deviations. For the present case this means that:

$$\mathrm{cov}\left(x_i, i\right) = \rho_p \sigma_x \sigma_i \qquad 2A.2$$

Substituting expression 2A.2 into expression 2A.1 leads to:

$$\rho = \sqrt{\frac{3}{4}\left(\frac{n-1}{n+1}\right)} \bullet \frac{4\rho_p\sigma_x\sigma_i}{n\sigma_x} = \sqrt{\frac{3}{4}\left(\frac{n-1}{n+1}\right)} \bullet 4\rho_p\sigma_{i/n} \qquad 2A.3$$

In turn, rearranging expression 2A.3, leads to the following expression for the Stuart correlation:

$$\rho = \rho_p \bullet 4\sigma_{i/n}\sqrt{\frac{3}{4}\left(\frac{n-1}{n+1}\right)} \qquad 2A.4$$

Expression 2A.4 makes two things quite obvious. Firstly, it tells us that the Stuart correlation measures the association between the values of the data and its ranks, over the range of the latter's relative values. Secondly, it helps to recall from Chebyshev's theorem, that the range of any data is its quadrupled standard deviation corrected for over-measurement. In the case of the *relative ranks* for any distribution, we know that this range ($r_{i/n}$) by the Chebyshev theorem is given by:

$$r_{i/n} = 4\sigma_{i/n} \bullet \sqrt{\frac{3}{4}\left(\frac{n-1}{n+1}\right)} \qquad 2A.5$$

Substituting expression 2A.5 into expression 2A.4, leads to the following re-expression for the Stuart correlation:

$$\rho = \rho_p \bullet r_{i/n} \qquad 2A.6$$

In turn expression 2A.6 can be rewritten in terms of the ratio between the correlations:

$$\frac{\rho}{\rho_p} = r_{i/n} \qquad 2A.7$$

It is known that that the distribution of the relative ranks of any data has well-defined bounds, with a minimum value of $1/n$, which approaches zero as the number of observations increases, and a maximum value that is always 1. Given that the range of the relative ranks is also the difference

between these extremes, $1-1/n$, it then follows that with 25-or more observations it begins to approach 1:

$$r_{i/n} \to 1, n \geq 25 \qquad \text{2A.8}$$

Substituting the limiting value for the range of the relative ranks from expression 2A.8 into expression 2A.7 reveals that there is an asymptotic equality between the Stuart correlation and the Pearson correlation:

$$\frac{\rho}{\rho_p} = 1 \Rightarrow \rho = \rho_p \qquad \text{2A.9}$$

However we can see that this asymptotic equality also exists when the Pearson correlation is equal to 1:

$$1 = \rho_p \qquad \text{2A.10}$$

Substituting expression 2A.10 into expression 2A.9, keeps the solution of the ratio unchanged, and shows that with 25-or more observations, the Stuart correlation is also 1:

$$\rho = 1 \qquad \text{2A.11}$$

Thus as the number of observations increases, the Stuart and Pearson correlations are both 1. This is the numerical solution to their asymptotic equality. By comparing expressions 2A.7 and 2A.9, we can see that the asymptotic equality between the correlations is governed by the range of the relative ranks. While from expression 2A.8 we can see that the limiting value of this range is 1, it clearly can also be less than that. Then the ratio between two correlations becomes instantaneously a one-side non-strict inequality:

$$\frac{\rho}{\rho_p} \leq 1 \qquad \text{2A.12}$$

This inequality will be defined with respect to the Stuart correlation, and it will reach its limit of 1, whenever the Pearson correlation is 1. In response, by substituting expression 2A.10 into expression 2A.12, the Stuart inequality is reached:

$$\rho \leq 1 \Rightarrow \rho = 1 \qquad \text{2A.13}$$

4. COMPUTATIONAL METHODS FOR THE HHI

Departing from Chapter 3, the aim of the present chapter is to pursue in more detail the computation of the HHI. As part of this the conventional computation of the HHI by the coefficient of variation will be considered first. This will then be followed by the calculation of the HHI involving the Gini index. Lastly, such a calculation will be considered in terms of the range of the data by drawing on the Chebyshev theorem to make this demonstration. This theorem has been invoked on a number of occasions so far, most recently in the demonstration of the Stuart correlation. For completeness, and without detracting from the aforementioned aim of the present chapter, the appendix to the present chapter will focus on a simple proof of the theorem's main result. This result is typically stated without its proof because as Touhey (1995: 139-140) points out its demonstration is considered too complicated – therefore the simpler proof provided here is a valuable contribution in its own right.

4.1 CALCULATION BY THE COEFFICIENT OF VARIATION

If the values of the data are expressed in proportional terms, then they sum up to one:

$$\sum_{i=1}^{n} x_i = 1 \qquad 3.1$$

There are two well know things about such a series. Firstly its mean is the reciprocal of the number of observations:

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1}{n} \qquad 3.2$$

And secondly its squared coefficient of variation is the product of the number of observations and the sum of its squared values subtracted from one:

$$\frac{\sigma_x^2}{\mu^2} = n\sum_{i=1}^{n} x_i^2 - 1 \qquad 3.3$$

Leveraging on expression 3.3, Hirschman (1964: 761) argued that the HHI should be computed from that expression, since by a slight rewrite it yields expression 1.1. To see this, expression 3.3 is rearranged to give the HHI:

$$\sum_{i=1}^{n} x_i^2 = \frac{\frac{\sigma_x^2}{\mu^2} + 1}{n} = \frac{c^2 + 1}{n} \qquad 3.4$$

Expression 3.4 shows that to compute the HHI we need to sum up the squared values of the data when it is expressed in proportional terms. Given that in practice market shares are already reported in percentages this simply means that the HHI is computed as the sum of squared market shares. As reported by Hay and Morris (1991: 210), Cabral (2000: 155), and Carlton and Perloff (2000: 247) this method is the conventional way to compute the HHI. It is also the only method described for its calculation in the 2002 edition of the *New Palgrave Dictionary of Economics and The Law*. Alternative computational techniques do not stray much from this method. For instance Kelly (1981: 51-52, 55) proposed that the conventional method should be applied on the averaged market shares, as opposed to those obtained by a single collection. Elsewhere, Salop and O'Brien (2000: 597-598, 610-611) have argued that the conventional method should be applied on a revised distribution of market shares that is corrected for any missing information not contained in the initial numbers. They call the HHI applied on this distribution the modified HHI (MHHI). According to Salop and O'Brien (2000: 611):

> "the MHHI is equal to the HHI plus a set of terms reflecting … cross-ownership within the industry."

In effect this is the same as applying the conventional method on the revised market shares, which are the initial market shares corrected to include initially missing information on who owns whom.

The conventional method and its aforementioned derivatives could be used to estimate the HHI and derive its confidence intervals either from expressions 1.26 and 2.24 or expressions 1.27 and

2.25. The trouble with this type of computation would be that it is entirely predicated on the coefficient of variation and by extension it relapses into inaccuracy under imperfect data conditions. Then as we know from the Glasser inequality the coefficient of variation will produce inflated measurements of relative variability in the data, and in turn by extension the same applies for the HHI. Simply stated the HHI will show higher levels of business concentration than it should. Thus if we keep to the conventional method, Mandelbrot's criticism of the HHI applies. On the basis of the Glasser inequality, and particularly its limiting solution that the coefficient of variation and the Gini index are the same with 25-or more observations, it could be argued that there is nothing wrong with computing the HHI by the conventional method. This argumentation would be correct, but it would miss the point that it is precisely because of the equality between the Gini index and the coefficient of variation, that we can just as well use the former to measure relative variability. This is because irrespective of the number of observations, or whatever the data conditions, as seen from the Glasser inequality, the Gini index will not overstate the relative variability of the data. This is why it is necessary to reformulate the HHI in terms of the Gini index, as was done in expression 2.22. By such reformulation the Glasser inequality is satisfied. The consequential improvement in accuracy the HHI receives from the Gini index, also suggests the need to find it an accurate computational expression.

4.2 CALCULATION BY THE GINI INDEX

Gini (1936: 78), David (1968: 573-574; 1998: 373-374), as well as Piesch (2005: 285-289), provide several accounts which show that the identification of methods that reduce the computational complexity of the Gini index, in terms of the number of steps it takes to calculate it, is an ongoing area of research. Yitzhaki (1998: 24) gave the following summary of its developments:

> "While it is hard to make an accurate count of how many independent alternative definitions exist, there are clearly more than a dozen of them. This large number of alternative definitions explains why the Gini has been "reinvented" so often. It also explains why it is hard to work with the Gini."

In a nutshell, there is more than one way or technique by which to estimate, i.e. compute the Gini index, and by extension the same number of ways applies to the HHI in terms of its reformulation with the Gini index. Because of this connection each of these methods fulfils the Glasser inequality in the sense that as the number of observations, i.e. firms increases, they will produce the same results as the conventional HHI method, but by comparison, will not overstate business concentration when that number is small.

From among the available alternatives, a method, which has become exceptionally popular is that proposed by Lerman and Yitzhaki (1984: 365). It is simple, fast, and accurate. Its computational complexity has been further reduced by Ogwang (2000: 124). However both, Lerman and Yitzhaki, as well as Ogwang, give a much sanitised account of how the method is derived while promoting its use. The following provides a possible way by which the method is derived. This will show that it derives directly from the definition of the Gini index as twice the area of the Lorenz curve, as depicted in expression 2.1.

Lerman and Yitzhaki (1984: 365) begin by pointing out that because the determination of the Gini index is based on two variables – one comprising the values of observations, and the other their ranks – we can minimise the computation of its covariance component, by obtaining it as the slope of a regression (β_1). This is because the simplest way to compute the regression slope between two-variables is from the covariance between them and their standard deviations, or:

$$\beta_1 = \rho \bullet \frac{\sigma_x}{\sigma_i} = \frac{\mathrm{cov}\left(x_i, i\right)}{\sigma_x \sigma_i} \bullet \frac{\sigma_x}{\sigma_i} = \frac{\mathrm{cov}\left(x_i, i\right)}{\sigma_i^2} \qquad 3.5$$

Substituting expression 1.7 into 3.5 and thereafter re-arranging with respect to the covariance leads to:

$$\mathrm{cov}\left(x_i, i\right) = \beta_1 \sigma_i^2 = \frac{1}{12}\left(n^2 - 1\right)\beta_1 \qquad 3.6$$

Entering expressions 3.6 into that for the Gini index from expression 2.1 leads to:

$$G = \frac{2\,\mathrm{cov}(x_i, i)}{n\mu} = \frac{1}{6n}\left(n^2 - 1\right)\frac{\beta_1}{\mu} = \frac{1}{6n\mu}(n-1)(n+1)\beta_1 \qquad 3.7$$

Expression 3.7 is the method Lerman and Yitzhaki (1984: 365) propose for the computation of the Gini index. It is easy to see its appeal. All we have to do to calculate the Gini index is to read off the computer output for the mean of the values and the estimate of the regression slope. The regression is between the values of the observations (the dependant variable) and their ranks (the independent variable).

Ogwang (2000: 124) has proposed that the Lerman and Yitzhaki method can be simplified even further if the slope of the regression (β_1) is calculated directly from the observations:

$$\beta_1 = \frac{\sum_{i=1}^{n} i x_i - \frac{1}{n}\sum_{i=1}^{n} i \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} i\right)^2} \qquad 3.8$$

Expression 3.8 contains the following well-known natural number series:

$$\sum_{i=1}^{n} i = \frac{n}{2}(n+1) \qquad 3.9$$

$$\sum_{i=1}^{n} i^2 = \frac{n}{6}(n+1)(2n+1) \qquad 3.10$$

We also know that the mean can be re-expressed as follows:

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} \Rightarrow \sum_{i=1}^{n} x_i = n\mu \qquad 3.11$$

Then by substitution of expressions 3.9, 3.10, and 3.11 into expression 3.8, the regression slope is reduced to:

$$\beta_1 = \frac{\sum_{i=1}^{n} ix_i - \frac{1}{2}(n+1)\sum_{i=1}^{n} x_i}{\frac{n}{6}(n+1)(2n+1) - \frac{n}{4}(n+1)(n+1)}$$

$$= \frac{\sum_{i=1}^{n} ix_i - \frac{1}{2}(n+1)\sum_{i=1}^{n} x_i}{n(n+1)\left[\frac{1}{6}(2n+1) - \frac{1}{4}(n+1)\right]} = \frac{\sum_{i=1}^{n} ix_i - \frac{1}{2}(n+1)\sum_{i=1}^{n} x_i}{n(n+1)\left[\frac{2(2n+1) - 3(n+1)}{12}\right]}$$

$$= \frac{\sum_{i=1}^{n} ix_i - \frac{1}{2}(n+1)\sum_{i=1}^{n} x_i}{\frac{1}{12}n(n+1)(n-1)} \qquad 3.12$$

In turn substituting expression 3.12 into expression 3.7 leads to:

$$G = \frac{1}{6}\frac{1}{\sum_{i=1}^{n} x_i}(n-1)(n+1)\frac{\sum_{i=1}^{n} ix_i - \frac{1}{2}(n+1)\sum_{i=1}^{n} x_i}{\frac{1}{12}n(n+1)(n-1)}$$

$$= \frac{12}{6}\frac{1}{\sum_{i=1}^{n} x_i}\frac{2\sum_{i=1}^{n} ix_i - (n+1)\sum_{i=1}^{n} x_i}{2n} = \frac{2}{n}\frac{2\sum_{i=1}^{n} ix_i - (n+1)\sum_{i=1}^{n} x_i}{2\sum_{i=1}^{n} x_i}$$

$$= \frac{2}{n}\frac{2\sum_{i=1}^{n} ix_i - n\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i}{2\sum_{i=1}^{n} x_i} = \frac{2}{n}\frac{\sum_{i=1}^{n} ix_i}{\sum_{i=1}^{n} x_i} - \frac{2}{n}\frac{n}{2} - \frac{2}{n}\frac{1}{2} = \frac{2}{n}\frac{\sum_{i=1}^{n} ix_i}{\sum_{i=1}^{n} x_i} - 1 - \frac{1}{n} \qquad 3.13$$

As depicted in the last line of expression 3.13, with Ogwang's extension we do not even need a computer. We can just use a pocket calculator. In addition, running Ogwang's extension programmatically is no harder than the Lerman and Yitzhaki method. In fact we can see that there will be a point at which we have the option to omit from the calculations the reciprocal term for the number of observations. This is because that term $(1/n)$ approaches zero as that number grows.

If we apply Ogwang's extension to the data in proportional terms, then by substitution of expression 3.1 into expression 3.13, the latter reduces to:

$$G = \frac{2}{n}\sum_{i=1}^{n} ix_i - 1 - \frac{1}{n} \qquad 3.14$$

Taking the square of expression 3.14, we get:

$$\begin{aligned} G &= \left(\frac{2}{n}\sum_{i=1}^{n} ix_i - \left(\frac{n+1}{n}\right)\right)^2 \\ &= \frac{4}{n^2}\left[\sum_{i=1}^{n} ix_i\right]^2 - \frac{4}{n^2}(n+1)\sum_{i=1}^{n} ix_i + \frac{1}{n^2}(n+1)^2 \end{aligned} \qquad 3.15$$

Then by substitution of expression 3.15 into the reformulated expression for the HHI, we get:

$$\begin{aligned} HHI &= \frac{\frac{4}{n^2}\left[\sum_{i=1}^{n} ix_i\right]^2 - \frac{4}{n^2}(n+1)\sum_{i=1}^{n} ix_i + \frac{1}{n^2}(n+1)^2 + 1}{n} \\ &= \frac{4\left[\sum_{i=1}^{n} ix_i\right]^2 - 4(n+1)\sum_{i=1}^{n} ix_i + (n+1)^2 + n^2}{n^3} \end{aligned} \qquad 3.16$$

Expression 3.16 gives a new computational method for the HHI, which is compliant with the Glasser inequality. It clearly shows that the HHI can be converted into a robust estimator of business concentration – based on the fact that it includes the ranks of the observations – in the same way that robustness exists in the Gini index. As the Gini index is a robust estimator of relative variability, by analogy, it imparts the same on the HHI. We already know that robust estimators are accurate estimators, especially if the data conditions involve outliers, non-normal distributions, few observations, as well as considerable variation among observations. Likewise, we know that they are just as accurate, when these conditions do not exist. It follows then, in terms of the solution of the Glasser inequality, that expression 3.16 provides an accurate computational method for the HHI relative to the conventional method.

In practice however it is not unusual to be unsure about the market shares of firms, except to be confident only in the market shares of the largest and smallest firm. Because of their extremeness such firms attract more attention than their intermediary rivals. In turn we may know more about

them than their rivals. As we would not have all the market shares on all the firms we would be unable to compute the HHI from expression 3.16. In such a case, in order to fulfil the Glasser inequality, we need to compute the HHI from the range of the data while still keeping its robustness from the Gini index.

4.3 CALCULATION BY THE RANGE

The range of the data (r) is the difference between its maximum (x_i^{max}) and minimum (x_i^{min}) value:

$$r = x_i^{max} - x_i^{min} \qquad 3.17$$

But from the Chebyshev theorem, we also know very well, that the range is also the quadrupled standard deviation of the data's values corrected for over-measurement. For the data's *values* of any distribution, we know that this is given by:

$$r = 4\sigma_x \bullet \frac{n}{n+1} \qquad 3.18$$

By rearrangement of expression 3.18, the standard deviation is the quartered range corrected for under-measurement, or:

$$\sigma_x = \frac{r}{4} \bullet \frac{n+1}{n} \qquad 3.19$$

By substitution of expression 3.19 into the equality between the Gini index and the coefficient of variation, Glasser (1961a: 178, 180; 1961b: 399) shows that, the Gini index can be accurately computed from its range, by:

$$G = \frac{\sigma_x}{\mu} = \frac{r}{4\mu} \bullet \frac{n+1}{n} \qquad 3.20$$

Glasser (1961a: 178, 180) has also empirically verified this technique to be valid. It should be borne in mind that it operates on the premise that the range is the only reliable information we have

for the data. Assuming that we are working with the data in proportional terms, then by substitution of expression 3.1 into expression 3.20, the latter can be re-expressed as:

$$G = \frac{\frac{r}{4}}{\overline{n}} \bullet \frac{(n+1)}{n} = \frac{r}{4}(n+1) \qquad 3.21$$

By cancellation of terms, expression 3.21 becomes:

$$G = \frac{r}{4}(n+1) \qquad 3.22$$

Taking the square of expression 3.22, leads to:

$$G^2 = \frac{r^2}{16}(n+1)^2 \qquad 3.23$$

In turn by substitution of expression 3.23 into the reformulated expression for the Gini index, we get:

$$HHI = \frac{G^2+1}{n} = \frac{\frac{r^2}{16}(n+1)^2+1}{n} = \frac{r^2(n+1)^2+16}{16n} \qquad 3.24$$

To recall from our earlier discussion, robust estimators can be constructed from any number of quantities such as the ranks, range, median, median absolute deviation, inter-quartile range, and generally anything that can withstand outliers and skewed distributions. In the present situation, expression 3.24 provides a robust computational method for the HHI based on the range of the data and its number of observations. The method is derived directly from the Gini index ensuring that there is no migration from the Glasser inequality.

Thus far we have derived a number of estimation techniques for the HHI. The conventional method is revealed to have a questionable accuracy because it makes the HHI a potentially-biased measure of business concentration that is prone to overstate its actual level. The methods by the Gini index and the range are revealed to serve as robust estimation techniques for the HHI.

Provided that we have reliable information on the market shares of all firms we could estimate the HHI from the Gini index. If we are only sure of the reliability of the market share numbers for the best and worst performing firm we can estimate the HHI from the range. The latter two techniques are accurate in the theoretical sense of the word. As it happens, in practice, we want to know precisely how accurate their estimates are. For this, we turn to finding their confidence limits, by applying McKay's confidence intervals of the HHI to their values. Such demonstration will be when we deal with the practical uses of the HHI.

4.4 SUMMARY

In the present chapter a number of estimation techniques, i.e. computational methods have been considered for the HHI. The conventional method for the estimation of the HHI comes directly from the coefficient of variation, and suffers from the drawback that under imperfect data conditions it does not make the HHI an accurate measure of business concentration. If such conditions do not exist then the estimation of business concentration by the HHI is accurate in as much as it is if the Gini index is used for its formulation. Permanently retaining the Gini index in the HHI comes with the added advantage that it makes it a reliable measure of business concentration irrespective of the market share distribution encountered. If we have full confidence in the data before us, we can estimate the HHI from its ranks and values as per expression 3.16. If we only have confidence in the market share numbers of the largest and smallest firm, we can estimate the HHI from expression 3.24. What is important to realise is that these estimation techniques are of a different substance to the conventional estimation technique or its variations. By virtue of the existence of the Glasser inequality the latter leave the HHI as a potentially-biased measure of business concentration. Mandelbrot's criticism applies here too, in the sense that if the conventional computation of the HHI is done via the coefficient of variation, than its relevance to tell us what the actual level of business concentration is, is actually diminished. It is then certainly hard to believe why the HHI should be taken seriously. The necessary step to reverse this is to reformulate the HHI in terms of the Gini index as done already. From such reformulation, there is also more positive news concerning the estimation of the HHI. As highlighted by Yitzhaki, there are more than a dozen ways in circulation by which to compute the Gini index. By association the same number

of techniques extends to the HHI. It is safe to say that by showing that the HHI is a variant of the Gini index, we now have an explosion of prospective estimation techniques that can be equally applied to the HHI. This is a stark improvement to the single alternative offered by the conventional method. Consequently, while for simplicity only a handful of methods have been considered here, future research can just as well consider more.

4.5 APPENDIX: CHEBYSHEV'S THEOREM

This appendix is concerned with the main result of the Chebyshev theorem, widely-known, according to which the range is the quadrupled standard deviation of the data's values corrected for over-measurement. Most recently we saw the use of this result in deriving a computational expression for the HHI in terms of the data's range. Prior to this, the result was used in the derivation of the Stuart correlation, and so far has been an inseparable part of our discussion on the HHI. Its demonstration however is typically hindered by complicated derivations, on account of which the result is generally stated without its proof. Commenting on this state of affairs Touhey (1995: 139-140) remarked that:

> "Although Chebyshev's theorem is stated in almost all elementary statistics textbooks, few include a proof. The reason is that the usual algebraic proof is not very illuminating to students at this level."

On this basis, Touhey (1995: 140) has argued that there is a need for a simple proof of the main result of the Chebyshev theorem that gives:

> "…a better intuitive feel for the concepts of variance and standard deviation."

Building on the works of Thomson (1955: 268), Guterman (1962: 134-135), Book (1979: 332-333), and Hayes (2010: 54-55), such a proof is offered here. By providing a simplified derivation, it will become clear that the main result of the Chebyshev theorem can be seen as a restatement of the maximum value of the ratio between the data's standard deviation and its range when this ratio is expressed with respect to the range itself.

To begin with, recall that the *sample* standard deviation of any data (s_x) is asymptotically equal to its *population* standard deviation (σ_x). A proof of this is given by Book (1979: 332-333). As part of

our demonstration, it is reproduced here. Its starting point is the truism that the sum of the squared differences of sample values from their average is never more than its resultant value:

$$\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2 \le \sum_{i=1}^{n}\left(x_i-\mu+\mu-\bar{x}\right)^2 \qquad 3A.1$$

where x_i is the ith sample value, $\bar{x}$ the sample mean, and μ the population mean

We can re-express 3A.1 in ratio terms:

$$\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i-\mu+\mu-\bar{x}\right)^2} \le 1 \qquad 3A.2$$

Expression 3A.2 implies that in a sample, the sum of the squared differences of its values from their average is asymptotically equal to its resultant value:

$$\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2 = \sum_{i=1}^{n}\left(x_i-\mu+\mu-\bar{x}\right)^2 \qquad 3A.3$$

Rewriting the right-hand side of expression 3A.3 leads to:

$$\begin{aligned}\sum_{i=1}^{n}\left(x_i-\mu+\mu-\bar{x}\right)^2 &= \sum_{i=1}^{n}\left[\left(x_i-\mu\right)^2+2\left(x_i-\mu\right)\left(\mu-\bar{x}\right)+1\left(\mu-\bar{x}\right)^2\right]\\ &= \sum_{i=1}^{n}\left(x_i-\mu\right)^2+2\left(\mu-\bar{x}\right)\sum_{i=1}^{n}\left(x_i-\mu\right)+\left(\mu-\bar{x}\right)^2\sum_{i=1}^{n}1\\ &= \sum_{i=1}^{n}\left(x_i-\mu\right)^2+\left[2\left(\mu-\bar{x}\right)\left(\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}\mu\right)\right]+n\left(\mu-\bar{x}\right)^2\end{aligned} \qquad 3A.4$$

By rearrangement of the expression for the arithmetic mean, we know that the sum of the sample values is:

$$\sum_{i=1}^{n}x_i = n\bar{x} \qquad 3A.5$$

If the values in question themselves happen to be population means, then:

$$\sum_{i=1}^{n}\mu = n\mu \qquad \text{3A.6}$$

Substituting expressions 3A.5 and 3A.6 into expression 3A.4 leads to:

$$\begin{aligned}
&\sum_{i=1}^{n}\left(x_i-\mu+\mu-\bar{x}\right)^2 = \sum_{i=1}^{n}\left(x_i-\mu\right)^2 + \left[2\left(\mu-\bar{x}\right)\left(n\bar{x}-n\mu\right)\right] + n\left(\mu-\bar{x}\right)^2 \\
&= \sum_{i=1}^{n}\left(x_i-\mu\right)^2 + 2n\left(\mu-\bar{x}\right)\left(\bar{x}-\mu\right) + n\left(\mu-\bar{x}\right)^2 \\
&= \sum_{i=1}^{n}\left(x_i-\mu\right)^2 + n\left(\mu-\bar{x}\right)\left[2\left(\bar{x}-\mu\right)+\left(\mu-\bar{x}\right)\right] \\
&= \sum_{i=1}^{n}\left(x_i-\mu\right)^2 - n\left(\bar{x}-\mu\right)\left(\bar{x}-\mu\right) = \sum_{i=1}^{n}\left(x_i-\mu\right)^2 - n\left(\bar{x}-\mu\right)^2
\end{aligned} \qquad \text{3A.7}$$

By rearranging the expression for the population variance in terms of its product with the number of observations we get:

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i-\mu\right)^2 \Rightarrow n\sigma_x^2 = \sum_{i=1}^{n}\left(x_i-\mu\right)^2 \qquad \text{3A.8}$$

By rewriting the last term of expression 3A.7 we also get:

$$\begin{aligned}
&n\left(\bar{x}-\mu\right)^2 = n\left[\left(\frac{\sum_{i=1}^{n}x_i}{n}-\frac{\sum_{i=1}^{n}\mu}{n}\right)\right]^2 = n\left[\left(\frac{\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}\mu}{n}\right)\right]^2 \\
&= n\left[\left(\frac{\sum_{i=1}^{n}\left(x_i-\mu\right)}{n}\right)\right]^2 = \frac{n}{n^2}\sum_{i=1}^{n}\left(x_i-\mu\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i-\mu\right)^2 = \sigma_x^2
\end{aligned} \qquad \text{3A.9}$$

Re-entering expressions 3A.8 and 3A.9 into expression 3A.7 yields the following expression:

$$n\sigma_x^2 - \sigma_x^2 = \left(n-1\right)\sigma_x^2 \qquad \text{3A.10}$$

Expression 3A.10 is an asymptotic equality. Expressed as a one-sided non-strict inequality in accordance with expression 3A.1, we can rewrite expression 3A.10 as:

$$\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2 \leq (n-1)\sigma_x^2 \qquad 3A.11$$

Multiplying both sides of expression 3A.11 by the reciprocal of the number of observations leads to:

$$\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2 \leq \frac{1}{n}(n-1)\sigma_x^2 \qquad 3A.12$$

On balance, the above multiplication leaves expression 3A.12 unchanged. However rearranging 3A.12 leads to the following re-expression:

$$\frac{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}{\sigma_x^2} \leq \frac{n}{n} \qquad 3A.13$$

We would recognise that the numerator on the left-hand side of expression 3A.13 is that for the sample variance of the data's values (s_x^2). We can then rewrite expression 3A.13 as:

$$\frac{s_x^2}{\sigma_x^2} \leq 1 \Rightarrow \frac{s_x^2}{\sigma_x^2} = 1 \qquad 3A.14$$

It follows from expression 3A.14 that the ratio between the sample variance and the population variance approaches 1 as the number of observations increases. In the limit, with a large number of observations, the maximum value of this ratio is 1, implying that the sample variance is asymptotically equal to the population variance:

$$s_x^2 = \sigma_x^2 \qquad 3A.15$$

If we were to take the square root of both sides of expression 3A.15 this will also show that the sample standard deviation is asymptotically equal to the population standard deviation.

We will now see how expression 3A.15 enables us to derive the maximum value for the ratio of the data's standard deviation to its range. To this end use is made of the proof by Guterman (1962: 134-135), which has subsequently also been verified by Hayes (2010: 54-55). As part of our demonstration, it is reproduced here.

Suppose now that the data is *only* composed of its minimum and maximum value, or:

$$x_i = \left\{ x_i^{min}, x_i^{max} \right\} \qquad 3A.16$$

We know that the mean is always between these two values, or:

$$\mu = \frac{1}{2}\left(x_i^{min} + x_i^{max} \right) \qquad 3A.17$$

We also know that the difference between the minimum and maximum value is the range:

$$r = x_i^{max} - x_i^{min} \qquad 3A.18$$

It is easy to infer that the distance of either value from the data's mean is half the range. If the distance is measured from the maximum value this number is positive:

$$x_i^{max} - \mu = x_i^{max} - \frac{1}{2}\left(x_i^{min} + x_i^{max} \right) = \frac{1}{2}\left(x_i^{max} - x_i^{min} \right) = \frac{r}{2} \qquad 3A.19$$

If the distance is measured from the minimum value this number is negative:

$$x_i^{min} - \mu = x_i^{min} - \frac{1}{2}\left(x_i^{min} + x_i^{max} \right) = \frac{1}{2}\left(x_i^{min} - x_i^{max} \right) = -\frac{r}{2} \qquad 3A.20$$

If in turn the distance between each data point and the mean is considered by its absolute values, i.e. irrespective of its sign, then it follows that:

$$\left| x_i - \mu \right| \leq \frac{1}{2} r \Rightarrow \left| x_i - \mu \right| = \frac{1}{2} r \qquad 3A.21$$

Expression 3A.21 shows that the *maximum* or largest distance from the mean, whether taken from the minimum or maximum value of the data, is half its range. But because all other data points fall below these extremes, their distance away from the mean will be less than this. However if we work with the solution for the maximum distance and square this, we get:

$$\left|x_i - \mu\right|^2 = \left(x_i - \mu\right)^2 = \frac{r^2}{4} \qquad 3A.22$$

The modulus in expression 3A.22 is replaced with brackets because it concerns the same *squared* distance, which in either case is a positive number. Summing across *all* the distances from the mean yields:

$$\sum_{i=1}^{n}\left(x_i - \mu\right)^2 = \sum_{i=1}^{n}\left(1 \bullet r^2/4\right) = \frac{r^2}{4}\sum_{i=1}^{n}1 = \frac{nr^2}{4} \qquad 3A.23$$

Multiplying both sides of expression 3A.23 by the reciprocal of the number of observations subtracted from 1, results in:

$$\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \mu\right)^2 = \frac{1}{n-1}\frac{nr^2}{4} \qquad 3A.24$$

We would recognise that the left-hand side of expression 3A.24 is the sample variance. We already know that this variance is asymptotically equal to the population variance. In turn, substituting expression 3A.15 into expression 3A.24, followed by rearrangement of terms, leads to:

$$\sigma_x^2 = \frac{n}{n-1} \bullet \frac{r^2}{4} \Rightarrow \frac{\sigma_x^2}{r^2} = \frac{1}{4} \bullet \frac{n}{n-1} \qquad 3A.25$$

Taking the square root of both sides of expression 3A.25 gives the ratio between the data's standard deviation and its range:

$$\frac{\sigma_x}{r} = \frac{1}{2} \bullet \frac{n^{1/2}}{(n-1)^{1/2}} \qquad 3A.26$$

Expression 3A.26 is an asymptotic equality. If it is expressed in accordance with expression 3A.21 as a one-sided non-strict inequality it becomes:

$$\frac{\sigma_x}{r} \leq \frac{1}{2} \bullet \frac{n^{1/2}}{(n-1)^{1/2}} \qquad 3A.27$$

Concerning the result of expression 3A.27, Thomson (1955: 268) explains that we can think of the ratio of the data's standard deviation to its range in standalone terms, as something that has its own distribution. Whatever this distribution is, its maximum value is that identified in expression 3A.27. Were we to assume that this distribution is symmetric then the *minimum* value of this ratio will be exactly the same but with the opposite sign, resulting in the following re-expression for 3A.27:

$$\frac{\sigma_x}{r} \geq -\frac{1}{2} \bullet \frac{n^{1/2}}{(n-1)^{1/2}} \qquad 3A.28$$

There is another truism that we know quite well, namely that the maximum value of any data is always greater than its minimum value. This means that the outcome or solution of expression 3A.27 exceeds that of expression 3A.28:

$$\frac{1}{2} \bullet \frac{n^{1/2}}{(n-1)^{1/2}} > -\frac{1}{2} \bullet \frac{n^{1/2}}{(n-1)^{1/2}} \qquad 3A.29$$

Multiplying both sides of expression 3A.29 by a ratio involving the number of observations gives:

$$\left(\frac{1}{2}\frac{(n-1)^{3/2}}{n^{3/2}}\right) \bullet \frac{1}{2}\frac{n^{1/2}}{(n-1)^{1/2}} > -\frac{1}{2}\frac{n^{1/2}}{(n-1)^{1/2}} \bullet \left(\frac{1}{2}\frac{(n-1)^{3/2}}{n^{3/2}}\right) \qquad 3A.30$$

As both sides of expression 3A.30 are multiplied by the same term, on balance the expression is left unchanged. However a rearrangement results in:

$$\frac{1}{4}\frac{(n-1)^{\frac{3}{2}-\frac{1}{2}}}{n^{\frac{3}{2}-\frac{1}{2}}} > -\frac{1}{4}\frac{(n-1)^{\frac{3}{2}-\frac{1}{2}}}{n^{\frac{3}{2}-\frac{1}{2}}} \Rightarrow \frac{1}{4}\frac{(n-1)}{n} > -\frac{1}{4}\frac{(n-1)}{n} \qquad 3A.31$$

Between expression 3A.31 and 3A.27 we have two alternate depictions for the maximum value of the ratio of the standard deviation to the range. By extension we can rewrite expression 3A.27 by replacing the maximum value with that from expression 3A.31, which gives:

$$\frac{\sigma_x}{r} \leq \frac{1}{4}\frac{(n-1)}{n} \Rightarrow \frac{\sigma_x}{r} = \frac{1}{4}\frac{(n-1)}{n} \qquad 3A.32$$

Expression 3A.32 shows that the ratio of the data's standard deviation to its range is a one-sided non-strict inequality, with an exact solution given by its maximum value. Thus if we work with this solution, a re-arrangement of expression 3A.32 with respect to the range yields:

$$r = 4\sigma_x \bullet \frac{n}{n-1} \qquad 3A.33$$

Alternatively, if the inequality is kept, then expression 3A.33 can be re-written as:

$$r \geq 4\sigma_x \bullet \frac{n}{n-1} \qquad 3A.34$$

By now we should recognise expression 3A.34 as the main result of the Chebyshev theorem if the data is *even*. It shows that as a measure of dispersion the range will be more than four times greater than the standard deviation if it is not corrected, i.e. *reduced*, in order not to overstate the dispersion of the data. This reduction comes from the proportional term involving the number of observations, whose function is to lower the otherwise overstated value of the data's spread according to its range.

Let's return to expression 3A.29, and instead multiply both of its sides by another ratio involving the number of observations:

$$\frac{1}{2}\frac{(n-1)^{1/2}(n+1)}{n^{3/2}} \bullet \frac{1}{2}\frac{n^{1/2}}{(n-1)^{1/2}} > -\frac{1}{2}\frac{n^{1/2}}{(n-1)^{1/2}} \bullet \frac{1}{2}\frac{(n-1)^{1/2}(n+1)}{n^{3/2}} \qquad 3A.35$$

There is no change to expression 3A.35, as both sides are multiplied by the same term, which cancels out either way. However after an additional work-out, expression 3A.35 reduces to:

$$\frac{1}{4}\frac{(n+1)}{n^{\frac{3}{2}-\frac{1}{2}}} > -\frac{1}{4}\frac{(n+1)}{n^{\frac{3}{2}-\frac{1}{2}}} \Rightarrow \frac{1}{4}\frac{n+1}{n} > -\frac{1}{4}\frac{n+1}{n} \qquad 3A.36$$

Between expression 3A.36 and 3A.27 we have *another* two alternate depictions for the maximum value of the ratio of the standard deviation to the range. If we were now to replace the maximum value in expression 3A.27 with that from expression 3A.36, we get:

$$\frac{\sigma_x}{r} \leq \frac{1}{4}\frac{n+1}{n} \Rightarrow \frac{\sigma_x}{r} = \frac{1}{4}\frac{n+1}{n} \qquad 3A.37$$

Here too, expression 3A.37 shows that the ratio between the data's standard deviation and its range is a one-sided non-strict inequality, with an exact solution given by its maximum value. If we work with this solution, a re-arrangement of expression 3A.37 with respect to the range gives:

$$r = 4\sigma_x \bullet \frac{n}{n+1} \qquad 3A.38$$

If instead we keep up to the inequality, expression 3A.38 can be rewritten as:

$$r \geq 4\sigma_x \bullet \frac{n}{n+1} \qquad 3A.39$$

We would recognise that expression 3A.39 is the main result of the Chebyshev theorem if the data is *odd*. Here too we see that as a measure of dispersion the range is more than four times greater than the standard deviation if it is not corrected, i.e. *reduced*, when measuring the dispersion of the

data. Again, this reduction comes from a proportional term involving the number of observations, the function of which is to lower the otherwise overstated value of data spread according to the range.

After investigating the two correction factors in expressions 3A.34 and 3A.39, i.e. $n/(n-1)$ and $n/(n+1)$, Glasser (1961a: 179-180) found that, as the number of observations becomes large, they make no practical difference to either expression. In such cases both have a limiting value of 1, thereby neutralising their presence in the expressions. Thus, Glasser's finding indicates that either one of the expressions is usable without the need to be preoccupied as to the appropriateness of which expression is best for different data situations. This corroborates with the Chebyshev theorem, which generally requires no knowledge of the distribution of the data. For instance we can see this from expressions 3A.32 and 3A.37. In either of these, to find the ratio between the data's standard deviation and its range, we only require the number of observations without any regard for their distribution.

In the last round, we are left with demonstrating the main result of the Chebyshev theorem for the case when the data *comprises* the relative ranks of the observations as presented in the appendix to Chapter 3 dealing with the Stuart correlation.

Suppose now that the data is partitioned into quartiles. We know that the position of the first quartile (Q_1) is:

$$Q_1 = \frac{1}{4}(n+1) \qquad 3A.40$$

In addition, we know that the position of the fourth or last quartile (Q_4) is:

$$Q_4 = \frac{4}{4}(n+1) \qquad 3A.41$$

We also know that the position of the data's range according to the *positioning* of its quartiles is captured by the third quartile (Q_3):

$$Q_3 = Q_4 - Q_1 = \frac{4}{4}(n+1) - \frac{1}{4}(n+1) = \frac{3}{4}(n+1) \qquad 3A.42$$

If we multiply both sides of expression 3A.42 by the same ratio involving the number of observations, we get:

$$\frac{\frac{(n-1)}{2}}{\frac{(n+1)^2}{2}} \bullet Q_3 = \frac{3}{4}(n+1) \bullet \frac{\frac{(n-1)}{2}}{\frac{(n+1)^2}{2}} \qquad 3A.43$$

In effect this double-sided multiplication leaves expression 3A.43 unchaged. But by rearrangement and cancelattion of terms, we get:

$$Q_3\left(\frac{n-1}{(n+1)^2}\right) = \frac{3}{4}\left(\frac{\frac{n-1}{2}}{\frac{n+1}{2}}\right) = \frac{3}{4}\left(\frac{n-1}{n+1}\right) \qquad 3A.44$$

Expression 3A.44 still depicts the position of the third quartile, except that now this position is rebalanced according to how many observations are above the data's average position, $(n-1)/2$, and how many are below this position, $(n+1)/2$. For reminder the average position is that of the data's second quartile.

If we introduce expression 3A.44 into expression 3A.27 we find the spread of the relative ranks of the observations (i/n) around their average position *in terms* of the position of their quartile spread. To do this, multiply both sides of expression 3A.29 by yet another ratio involving the number of observations:

$$\frac{1}{2}\left[\frac{4}{3}\left(\frac{n+1}{n}\right)\right]^{1/2} \bullet \frac{1}{2}\frac{n^{1/2}}{(n-1)^{1/2}} > -\frac{1}{2}\frac{n^{1/2}}{(n-1)^{1/2}} \bullet \frac{1}{2}\left[\frac{4}{3}\left(\frac{n+1}{n}\right)\right]^{1/2} \qquad 3A.45$$

We see that on balance the multiplication has left expression 3A.45 unchanged. However, after regrouping, the following re-expression is reached:

$$\frac{1}{4}\left[\frac{4^{1/2}}{3^{1/2}}\frac{(n+1)^{1/2}}{n^{1/2}}\right]\bullet\frac{n^{1/2}}{(n-1)^{1/2}}>-\frac{1}{4}\left[\frac{4^{1/2}}{3^{1/2}}\frac{(n+1)^{1/2}}{n^{1/2}}\right]\bullet\frac{n^{1/2}}{(n-1)^{1/2}} \qquad 3A.46$$

An additional sorting of terms in expression 3A.46 gives:

$$\frac{1}{4}\frac{(4)^{1/2}}{(3)^{1/2}}\frac{(n+1)^{1/2}}{(n-1)^{1/2}}>-\frac{1}{4}\frac{(4)^{1/2}}{(3)^{1/2}}\frac{(n+1)^{1/2}}{(n-1)^{1/2}} \qquad 3A.47$$

Together with expression 3A.27, expression 3A.47 gives another alternate depiction for the maximum value of the ratio for the standard deviation to the range. By extension, substitution of the maximum value in the former expression with that in the latter expression, leads to:

$$\frac{\sigma_{i/n}}{r_{i/n}}\leq\frac{1}{4}\frac{(4)^{1/2}}{(3)^{1/2}}\frac{(n+1)^{1/2}}{(n-1)^{1/2}}\Rightarrow\frac{\sigma_{i/n}}{r_{i/n}}=\frac{1}{4}\frac{(4)^{1/2}}{(3)^{1/2}}\frac{(n+1)^{1/2}}{(n-1)^{1/2}} \qquad 3A.48$$

As the data consist of the relative ranks, expression 3A.48 reflects their standard deviation ($\sigma_{i/n}$) and range ($r_{i/n}$). After an additional rearrangement of expression 3A.48, the solution for the maximum value of the ratio with respect to the range becomes:

$$r_{i/n}=4\sigma_{i/n}\bullet\frac{(3)^{1/2}}{(4)^{1/2}}\frac{(n-1)^{1/2}}{(n+1)^{1/2}}=4\sigma_{i/n}\bullet\sqrt{\frac{3}{4}\left(\frac{n-1}{n+1}\right)} \qquad 3A.49$$

Expression 3A.49 is an asymptotic equality. In accordance with expression 3A.48, the counterpart one-sided non-strict inequality is:

$$r_{i/n}\geq 4\sigma_{i/n}\bullet\sqrt{\frac{3}{4}\left(\frac{n-1}{n+1}\right)} \qquad 3A.50$$

Again, as a measure of dispersion the range is more than four times greater than the standard deviation if it is not corrected, i.e. *reduced*, in order not to overstate the dispersion of the data. And again this reduction comes from a proportional term involving the number of observations. In the present case where the data is comprised of the relative ranks, its function is to bring down the otherwise overstated value in the spread of these ranks next to its value from the position implied by their number.

5. SECONDARY FINDINGS FOR THE HHI

5.1 EMPIRICAL APPROACH

In this chapter, secondary findings are provided in terms of giving an empirical confirmation of the implications of the Glasser inequality on the HHI, with the aid of simulation. Rather than showing the consequences of the asymptotic equality between the Gini index and the coefficient of variation, which has been demonstrated numerically already, the emphasis here is on showing the consequences of the Glasser inequality by its other outcome: this of the upward bias held by the coefficient of variation relative to the Gini index under imperfect data conditions, and how this feeds into the HHI.

As emphasised by Zickar (2006: 427-428), if there is nothing to suggest that simulation will be inferior for data analysis compared to doing the same analysis with real data from field collections or administrative records, there is no reason why it should not be used. There is nothing special about this emphasis, but it is important to keep it mind. It is well known, as pointed out by Van Belle (2008: 53-56), that simulation provides for controlled conditions of data analysis more easily than the undertaking of the same analysis via real data from field collections or administrative records. This is because it introduces fewer errors, if any, from measurement, data collection, and mathematical operations, all of which are endemic to real data. It accomplishes this by lessening the use of this data in the estimation process of a statistic. The ensuing discussion sheds light on this.

Synonymous with the Monte Carlo method, Sawilowsky (2006: 627) concisely defines simulation as a resampling approximation technique. While the brevity of this definition focuses attention on what simulation really is, it is somewhat too concise unless clarified. In a nutshell, simulation is a technique that generates artificial observations for a statistic from repetitive samples obtained by continually sampling a random seed of observations. This seed is either a random sample of real observations from field collections and administrative records, or a random sample from a pseudo-population. As Yung and Chan elaborate (1999: 87):

> "… a pseudo-population … is usually defined as the distribution of the sample data or of some appropriate transformation of the data."

In simplified terms, without the technical connotation, a pseudo-population is an outcome of random generation, which starts with any seed composed of randomly drawn real observations that are representative of reality. The requirement for representativeness is for the distribution of the randomly drawn real observations to resemble the shape of the distribution of the real population. This starting seed is then randomly shuffled many times just like a card deck. All the resultant reshufflings represent the pseudo-population.

To distinguish simulated data from real data, the simulated observations are sometimes called "fake" observations. This is to serve as a reminder that they are artificial. But it is important to emphasise that simulation is a *re-sampling* technique and therefore requires *real* data as a starting seed – albeit a limited amount of data. Simulation is conducted using various algorithms. By way of a reminder, an algorithm is simply a mechanical rule or procedure for the estimation of a statistic (Copeland, 1996: 335, 337). Diaconis and Efron (1983: 107-108) outline that bootstrapping, jack-knifing, and cross-validation, are the three main procedures or algorithms for simulation. Diaconis and Efron (1983: 107) also explain the similarities and differences between them in the following way:

> "Each of these procedures generates fake data sets from the original data and assesses the actual variability of a statistic from its variability over all the sets of fake data. The methods differ from … one another in the way the fake data sets are generated."

To clarify, the original data refers to the random seed or the basket of observations with which the algorithm is started. In the case of the bootstrap the same observations in the original data are repetitively drawn out at random to create cloned samples. The number of these repetitive samples, i.e. re-samples exceeds the number of observations of the original data. A statistic is calculated from each re-sample, and the procedure is terminated when additional re-sampling or

cloning no longer changes the variance of the statistic from all re-samples. In the case of the jackknife – following a random start, either at the top or the bottom of the original data, much like being at cross-roads before deciding which way to go – its observations are continually sampled until their total number is reached. This is done by removing the observations in the original data one at a time, beginning with the randomly identified starting point. A statistic is calculated from each resample and the procedure is terminated by the last calculation when the last observation is removed. Both the bootstrap and the jackknife, yield a repetitive selection of observations. The difference between the two is that the former generates more artificial data than the latter. As explained by Efron and Gong (1983: 40):

> "… the jackknife is, almost, a bootstrap itself. …The ordinary jackknife is the method of choice if one does not want to do the bootstrap computations."

In short the jackknife algorithm is the computationally light version of the bootstrap algorithm. In cross-validation the original data is randomly split in two halves, and each half is in turn also randomly split in half. This goes on until no further "half" splits are possible. A statistic is calculated from all halves.

Once the artificial estimates of a statistic are generated by anyone of the aforementioned algorithms, Efron (1988: 296) highlights they can be used to:

a) Conduct exploratory or confirmatory analysis of theory;
b) Confirm whether the sampling distribution of a statistic is subordinated to a determinate statistical distribution;
c) Find out what is the exact, if any, sampling distribution of a statistic; and
d) Empirically derive confidence intervals for the statistic if these cannot be obtained from a known statistical distribution.

We know that in the present case the last two uses of simulation are not applicable. We would recall that on account of the coefficient of variation or the Gini index, the HHI has no exact sampling distribution except that its shape is unimodal and positively skewed. As consequence its sampling distribution is approximated by the Chi-square distribution. Due to this, the HHI uses McKay's confidence intervals, which are based on this distribution. Resultantly the identification of empirical confidence intervals for the HHI is redundant because they are already obtainable from a known statistical distribution.

On the other hand, the first two uses of simulation are applicable to the present case. Up to now we have studied the Glasser inequality in terms of its limiting solution, showing that with 25-or more observations, the Gini index and the coefficient of variation are equal measures of relative variability. Incidentally such demonstration also reveals that the equality exists under the normal distribution, because as we know from the central limit theorem, sample sizes of approximately 30 observations are sufficient for the sampling distribution of the mean of a statistic to be approximately normal. In this sense, we have demonstrated the asymptotic equality between the coefficient of variation and the Gini index, provided that the data are normally distributed. However, in terms of the Glasser inequality, we know that when this no longer holds, the coefficient of variation is a biased measure of relative variability. It will be systematically higher than the Gini index. By extension, the bias will feed into the HHI, in the sense that the HHI calculated from the coefficient of variation will then exceed the HHI from the Gini index. Any simulation with less than 25 observations should confirm this. In addition, we already know, that either by association to the coefficient of variation or the Gini index, the sampling distribution of the HHI should by approximated by the Chi-square distribution. Any simulation irrespective of the number of observations it involves should confirm this.

To keep within the framework of the current enquiry, we can confirm the sampling distribution of the HHI from the Stuart correlation coefficient (ρ). *Firstly*, this will involve computing the Stuart correlation from the simulated data using expression 2A.4 as discussed in Chapter 3. *Secondly*, the Stuart correlation from the simulated estimates will be compared to its exact values for the Chi-

square distribution. O'Brien (1982: 150) has derived exact values of this coefficient for the Chi-square distribution when this distribution is extremely skewed, and when it is heavily skewed. The former case is typically identified as the Chi-square distribution with 1 degree of freedom. The latter case is typically identified as the Chi-square distribution with 4 degrees of freedom. Table 2 gives the exact Stuart correlations for the Chi-square distribution, as published in O'Brien (1982: 150).

Table 2: Exact Stuart correlations for the Chi-square distribution (%)

	Chi-square distribution	
Sample size (n)	**Extremely skewed ($df = 1$)**	**Heavily skewed ($df = 4$)**
2	100.00	100.00
3	93.81	95.06
4	91.73	94.23
5	89.97	93.79
6	88.70	93.47
9	86.68	93.27
18	83.73	93.18
27	82.59	92.53
36	81.40	92.46
72	79.74	92.36
$n \geq 73$	77.97	91.86

Because the Chi-square distribution is the approximating distribution, there is no expectation that the two correlations will match exactly. They are however expected to be in close proximity as a way of signalling that the fit by the Chi-square distribution closely agrees with the sampling distribution of the coefficient of variation, the Gini index, and by extension the HHI. Such confirmation will also validate the use of McKay's confidence intervals for the HHI, indicating that they should be based on a robust estimation of the HHI as discussed in Chapter 4.

5.2 SIMULATION RESULTS

To generate artificial data or to produce simulated estimates we have to run anyone of the aforementioned simulation algorithms. Efron and Tibshirani (1991: 390-391) indicate that there is no right or wrong decision to make regarding this. As they explain all of the algorithms (Efron and Tibshirani, 1991: 394):

"…differ in one important way from their classical predecessors: they substitute … algorithms for the traditional mathematical ways of getting a numerical answer."

This will now become apparent. To illustrate the point the computationally light version of the bootstrap algorithm will be run, i.e. the jackknife. It will be applied on the South African sugar market. Table 3 gives the markets shares by turnover of the firms making up this market, as published by the Competition Tribunal (2000: 13).

Table 3: Firms' market shares by turnover in the South African sugar industry (%)

Firms	***Market shares***
Tongaat-Hulett	36.0
Illovo	31.0
Transvaal Suiker Beperk	18.0
Swazi Sugar	15.0
TOTAL	**100.0**

Applying the jackknife algorithm to the data in table 3 led to the following calculation steps:

a) A random toss identified the last observation as the starting point of the algorithm;

b) Thereafter one observation at a time was removed from the data until its total number of observations was reached;

c) The outcome is four resamples or simulated markets, each comprising three observations made up of different combinations of firms and market shares. The simulated data is reported in Table 4;

Table 4: Simulated market shares from the South African sugar industry (%)

Firms	***1st resample***	***2nd resample***	***3rd resample***	***4th resample***
Tongaat-Hulett	42.0	44.0	52.0	
Illovo	36.0	38.0		48.0
Transvaal Suiker Beperk	22.0		26.0	28.0
Swazi Sugar		18.0	22.0	24.0
TOTAL	100.0	100.0	100.0	100.0

d) Each simulation gives different market shares, which vary from the original data. The new market shares are calculated by dividing their original values to the new totals for each simulated market;

e) The coefficient of variation, the Gini index, and the HHI are calculated from the values of each simulation, respectively from expressions 3.4 and 3.14 as discussed in Chapter 4. The estimates are reported in Table 5 together with their Stuart correlations, as well as their medians because they appear to be described by positively skewed distributions in respect of which the average is the inappropriate measure of central tendency.

Table 5: Simulated estimates of the HHI by different formulations

Simulation outcomes	*Coefficient of variation*	*Gini index*	*HHI from coefficient of variation*	*HHI from Gini index*
1st resample	30.79	13.33	36.49	33.93
2nd resample	40.84	17.33	38.89	34.33
3rd resample	48.87	20.00	41.29	34.67
4th resample	38.57	16.00	38.29	34.19
Median	**39.71**	**16.67**	**38.59**	**34.26**
Stuart correlations	**84.92**	**85.93**	**84.60**	**85.74**

f) With the calculation of the simulated or artificial estimates completed, the algorithm is terminated.

An examination of the results in Table 5 shows that the simulation confirms the existence of the Glasser inequality, in terms of which the coefficient of variation produces consistently higher estimates of relative variability than the Gini index under imperfect data conditions. As the medians show, on the average, the coefficient of variation overstates relative variability by 58% compared to the Gini index. We should note that this finding is specific to the current simulation because the Glasser inequality merely stipulates that there will be a systematic difference under imperfect data conditions. It does not tell what its actual magnitude will be, indicating that it will vary from one situation to the next. In the present situation it just happens to be high.

Just as expected, the simulation also shows that this difference sips-in into the HHI under such conditions. As a consequence the HHI obtained by the coefficient of variation overstates the

prevailing level of business concentration relative to that obtained by the HHI from the Gini index. The simulation confirms that the HHI obtained by the coefficient of variation is biased upwards when imperfect data conditions prevail. On the average, as captured by the medians, the HHI by the coefficient of variation systematically overstates the prevailing level of business concentration by 11% compared to the HHI from the Gini index. The presence of the number of observations or firms in the denominator of the HHI dampens the bias carried by the coefficient of variation, but it does not eliminate it. Furthermore as the Glasser inequality forewarns us, we should expect the systematic difference between the HHI from the coefficient of variation and the Gini index to vary from one situation to the next. In the present case this difference happens to be somewhat low, but it is not say that it cannot be higher, or lower. Either way what matters is that business concentration levels as measured by the HHI from the coefficient of variation will be overstated under imperfect data conditions when compared to the HHI from the Gini index. From the asymptotic equality between the coefficient of variation and the Gini index yielded by the Glasser inequality, we know that this would not be an issue under perfect data conditions. In such cases we can anyhow use the HHI from the Gini index as opposed to the HHI from the coefficient of variation because the two will then give the same estimates for the levels of business concentration. What is different here is that we have drawn attention to this much like Efron and Tibshirani advised we would, namely through the use of an algorithm in the place of mathematical proofs and their numerical evaluations.

Concerning the distribution of the HHI, we know that irrespective of the different abilities the coefficient of variation and the Gini index have in measuring relative variability, their approximating sampling distribution is the Chi-square distribution. By default the same extends to the HHI whether the coefficient of variation or the Gini index participates in its formulation. As expected the simulation confirms this. All estimates share a stable Stuart correlation, meaning that the HHI keeps the distribution it receives either from the coefficient of variation or the Gini index. Rounded off to the nearest common integer, all correlations are standing at 85%. They are strong. Comparing them to the exact Stuart correlations from Table 2 shows that the shape of the distributions of the estimates come closest to the expected fit by the extremely skewed Chi-square

distribution. The latter's Stuart correlation, when rounded off to the nearest integer, is 92%. The meaning of this is that while the Chi-square distribution does not give a perfect fit to the sampling distribution of the HHI, or for that matter the coefficient of variation or the Gini index, it still gives a strong approximate fit for each of these, just as expected. Another way of seeing this is to note that the difference of 8% between the correlations signals that the sampling distribution of the HHI mismatches the superimposed fit of the Chi-square distribution by 8%. Again this gives reinforcement that the latter distribution is a closely approximating sampling distribution for the HHI. By extension McKay's confidence intervals for the HHI are valid, in the sense that under imperfect data conditions, they must be derived by robust estimation of the HHI as discussed in Chapter 4. However this need not be done only for such conditions. We already know that the same estimation will also perform accurately in the case of perfect data conditions.

5.3 SUMMARY

The simulation results presented here have enabled us to confirm empirically what we have discussed mathematically in the previous chapters. They are really secondary findings because they do not give us anything new that we do not already expect from the mathematical illustrations. They do however highlight that it would be careless to work with the coefficient of variation under imperfect data conditions in the place of the Gini index. Then bias is inherently unavoidable in the measurement of relative variability. By extension the same carelessness will spread through to the HHI if it is calculated in terms of the coefficient of variation. The potential bias of the HHI with this formulation will differ from one market situation to the next. However, as seen here, the sheer fact that it exists, in turn implying that business concentration levels always run the risk of being overstated, should be sufficient grounds for wanting to find a way to eliminate it. To do so we need to work instead with the Gini-based formulation of the HHI. The latter has the virtue that it does just as well when the data conditions are perfect, making it unnecessary to change from it. Such a move also retains the sampling distribution of the HHI to the Chi-square distribution, implying that we retain the decision-making and hypothesis testing abilities of the index, while at the same time improving its measuring abilities as an accurate measure of business concentration. This is simply achieved by a change of its relative variability component. What this means for practical

applications of the HHI in terms of its uses as an indicator of business concentration and as a diagnostic tool for such concentration is considered next.

6. PRACTICAL USES OF THE HHI

In statistical contributions it should not necessarily be assumed that mechanical techniques of data analysis always suffice to tell the story, just because they are readily available or easy to implement. Tukey (1980: 24) for instance derided this practice, noting that:

> "Perhaps an equal effort, at least among statisticians, is needed to persuade us of the equally true statement, 'the usual bundle of techniques is not a field of intellectual activity'! …No catalogue of techniques can convey a willingness to look for what can be seen, whether or not anticipated."

It is from such considerations that Genichi Taguchi made use of transnumeration, which as Wu (1992: 140) has observed:

> "… is a very important practice he introduced to the inward-looking statistical community."

Wu's assessment, which was already referred to in the introductory chapter, suggests that Taguchi's transnumeration technique should not be ignored. Indeed, transnumeration has left a permanent footprint on the statistical analysis of data, showing that any mechanical analysis should be questioned when it is neither clear nor necessary why it should be performed. This is especially true if there is nothing to show that mechanical analysis by itself facilitates the understanding of data, or its real-world application.

A technical description of transnumeration is provided by Wild and Pfannkuch (1999: 227, 238; 2000: 137-139, 150). Put simply, it is a data analytic technique involving statistical story-telling that gives numerical literacy in the practical uses of a statistic. Blessing *et al*. (2005: 1) explain that:

> "A statistical story shows readers the significance, importance, and relevance of…information. In other words, it answers the question: Why should my audience want to read about this?
>
> …Statistical story-telling is about: catching the reader's attention with a headline or image; providing the story behind the numbers in an easily understood, interesting and entertaining fashion; and …how statistics might add impact to just about every story they have to tell."

In the current enquiry, flowing out of the technical findings discussed so far, there is a definite *illustrative* story to tell with the narrative of showing how the HHI can add impact to decision-making. To fulfil the characteristics of good statistical story-telling and make our story easily understood, interesting, and entertaining, we will tell an imaginary tale with the following fictitious characters:

John:	Government statistician at a statistical agency
Caroline:	John's director
Donald:	Business statistician at a major company in the chocolate industry
Amelia:	Chief executive officer of the chocolate company
Rebecca:	Chief economist of the chocolate company
Stan:	Senior civil servant working for the national competition regulator

6.1 CONCERNS

During a strategic planning meeting Caroline advises John that the statistical agency has decided to investigate the prospect of publishing the HHI as an official statistic because another branch of Government – the Competition Regulator – has expressed concerns that although the HHI is part and parcel of their decision-making the index is not officially available. As a consequence the Competition Regulator is feeling disadvantaged because it is forced to make decisions on the degree of competition in the economy in the dark. Caroline also advises that Stan at the Competition Regulator is working on an important case involving the chocolate industry, which is the country's biggest manufacturing industry. Due to the importance of this case, the Competition

Regulator has asked the statistical agency to pilot the publication of the HHI as an official statistic using this industry as an example.

During their discussion John reminds Caroline that just like any other official statistic the agency publishes, whether the HHI is produced and published, will depend entirely on whether the index can be accurately measured. This is because the public has to have confidence in the number released. John recommends to Caroline the work by Elvers and Rosn (1997: 622-626) to help substantiate his comments. John tells Caroline that the crux of the matter is that if the HHI is to be believable as an official statistic it must be reliable, and this can only be achieved if it is accurately measured.

The day after their conversation Caroline asks John to investigate further whether the statistical agency can compile the HHI. She also informs John that she will engage Stan only after the agency's investigation is concluded.

The following day, John extracts, from the business register, the latest available snapshots of the market shares by turnover, of the firms comprising the chocolate industry. The snapshots are recorded in Table 6.

Table 6: Market share by turnover in the chocolate industry (%)

Firms	*Year 1*	*Year 2*	*Rank (worst to best performing firm)*	*Annualised market share (obtained by geometric mean)*
Firm A	7.8	8.7	1	8.3
Firm B	15.0	15.7	2	15.4
Firm C	16.5	16.2	3	16.3
Firm D	18.0	17.3	4	17.7
Firm E	18.6	18.5	5	18.5
Firm F	24.1	23.6	6	23.8
TOTAL	100.0	100.0		100.0

With this information John sets about calculating the HHI. After much research John finds Herfindahl's (1955: 96, 98) and Hirschman's (1964: 761) articles on the HHI. The first one tells him that the HHI is expressed in terms of the coefficient of variation by:

$$HHI = \frac{c^2 + 1}{n} \qquad 4.1$$

The second one tells him that the computation for expression 4.1 is the sum of squared market shares:

$$\sum_{i=1}^{n} x_i^2 = \frac{\frac{\sigma_x^2}{\mu^2} + 1}{n} = \frac{c^2 + 1}{n} \qquad 4.2$$

After referencing the 2002 edition of the *New Palgrave Dictionary of Economics and The Law*, John finds that these technical formulations of the HHI have not changed. However he becomes very troubled. The measures appear biased because they overlook two famous results in mathematical statistics known as the De Vergottini and Glasser inequalities. John is reminded of them after coming across a paper by Piesch published in a 2005 issue of the prestigious *Metron* journal. According to these inequalities the coefficient of variation will always give elevated measurements of relative variability compared to the Gini index, except in the limit when with many observations, both measures of relative variability will be equal. John also comes across an article by Taguchi and Clausing (1990b: 66, 67-68, 69-70), and now fully remembers that he was taught about the Glasser inequality in applied statistics in the topics on measurement. Here the coefficient of variation and the Gini index are referred to as the noise-to-signal ratios because they describe the relative precision of estimates. They show how much of their typical value, i.e. the average, is corrupted by its dissimilarity from individual estimates, whether captured by the standard deviation in the case of the coefficient of variation, or the covariance in the case of the Gini index. Thus the typical value is regarded as the "signal", and the dissimilarities from it are the "noise". In terms of Glasser's inequality John is able to recall that the coefficient of variation produces more noisy estimates of relative variability compared to the Gini index. In response John begins to

contemplate how the HHI can be an accurate measure of business concentration if it is inherently susceptible to inflating its level.

John decides to investigate further. He comes across a review by Mandelbrot (1997: 215-218) that indeed confirms his fears. There Mandelbrot advises that the HHI is inconsiderately injected with the wrong second moment. John of course knows that Mandelbrot is right because he must be referring to the Glasser inequality for the Gini index in terms of which the right second moment for measuring relative variability is the Gini index, as well as because the Gini index is known to be a robust estimator of relative variability, which unlike the variance or the coefficient of variation, will accurately estimate this variability *even* when the data conditions are not perfect. In the present case John is concerned about this too where the chocolate industry has only a small number of firms. The number of observations is too small to assume any normality in the distribution of the shares, and besides, it has been known at least since the 1940s that, in practice, with small and with large data sets this assumption will seldom, if ever, be met. In this case John, is sure of this because during his research he comes across well substantiated studies from Gibrat (1931 [1957]: 57-58), Lawrence (1988: 231-233, 241-242, 251), Axtel (2001: 1819-1820), as well as Gaffeo *et al*. (2003: 119-121), that find that the typical distribution of firms' market shares in industries is positively skewed. John sees the same thing in his snapshots. Besides, even if hypothetically it could be assumed that the industry had more firms, it does not follow from this that their distribution would be normal. This is because the central limit theorem applies to the averages of observations obtained from repeated sampling. John knows there is no repeated sampling here because the market shares for each period are from single – not multiple – snapshots. And so reality shutters even the ideal hypothetical case.

After weighing the practical shortcomings John confronts, he tentatively concludes that it will be erroneous to calculate the HHI either by expression 4.1 or 4.2 because the resultant estimate of business concentration will be inaccurate. It does not help matters much when John is also unable to find any research that could show that the HHI is estimated with corresponding confidence limits. In paper after paper, John cannot find any confidence interval being reported for the HHI. He

comes across one lonely study by Adelman (1969: 101) that makes a plea for someone to find such intervals because then it would be possible to hypothesise and test the significance of the HHI. But beyond this John finds no confidence intervals for the HHI. To be sure that he is not missing something, John goes back to Mandelbrot's review, which informs him that the existence of an expectation for the HHI is controversial. John clicks. He immediately realises that Mandelbrot is aware that the distribution of market shares is already positively skewed just as confirmed by the studies he came across. For such distributions the coefficient of variation is not an accurate measure of relative variability while the Gini index is. Then, knowing that we can use the familiar McKay approximation to derive confidence intervals for the coefficient of variation from the Chi-square distribution, does not help in any way. The intervals become controversial because the expected range and expected value they will yield for the coefficient of variation are biased in the sense that the resultant values are higher than they should be. If the expected range or expected value for the coefficient of variation is unreliable, it is not at all surprising that the HHI will be defunct in the same way, and that Mandelbrot cautions against its use. For John, this only serves to highlight why, when formulated by the coefficient of variation, the HHI should be suspected to be an inaccurate measure of business concentration. In spite of this sceptical outlook for the HHI, John at least is able to learn that the sampling distribution of the HHI is approximated by the Chi-square distribution for which no prior knowledge is to be found anywhere.

John's views have just about cemented and he writes Caroline a memo sharing his findings, as well as recommending that the statistical agency should not calculate the HHI, because the index is disconnected from the relevant statistical concept of relative variability. John also writes that by extension the HHI cannot be measured accurately because the resultant confidence limits for it will also be inaccurate. John warns that this is outright problematic, because whatever the HHI's number, the agency cannot inform the public about the level of confidence with which it can be regarded.

On reading the memo, Caroline contacts John to thank him for his work, as well as to let him know that now more so than ever, if the HHI remains unpublished, it would not be anything the public

would miss, except for braches of Government with vested interest in its production. Indeed, Caroline shares in John's sentiment, that it will be damaging for a statistical agency to produce a statistic with unverifiable accuracy. This will reflect poorly on the credibility image of the agency. For this reason, as well as those technical reasons John found, Caroline thinks that the Competition Regulator will have to look elsewhere for obtaining their HHI numbers. Caroline decides to communicate this to Stan after she returns from an upcoming workshop on business concentration.

6.2 REMEDIES

Caroline attends a workshop on business concentration. She meets Donald, who has been asked by his employer to learn as much as he can on how to measure business concentration using the HHI, because the firm he works for, is about to merge with another firm from the same industry. In the course of the conversation, it becomes known that Donald works for one of the firms in the chocolate industry, on which Caroline had just completed her investigation. Donald's firm has familiarised with the HHI regulatory thresholds, and wants to make sure that the merger will not create contentious HHI levels that can lead to the Competition Regulator prohibiting the merger, or to the firm being compelled to divest of some of its operations as a condition to merging. This is why Donald's boss – Amelia – has made it clear to him, that because a lot rides on the merger to make their firm more profitable, she wants to know beyond any doubt what the actual HHI numbers for her industry are, and she wants Donald to be pretty sure what these numbers are.

Donald informs Caroline that he still needs to learn what the HHI is. He hopes to get all the answers at the workshop's Question and Answer session. The purpose of the session is to discuss some statistical aspects about the HHI. Caroline agrees to join David in the session, but also shares her knowledge about the deficiencies of the HHI. David becomes discouraged. Much to his surprise, and that of Caroline's, during the session, both David and Caroline are astounded to learn that:

a) Due to the asymptotic equality between the Gini index and the coefficient of variation, which is the limiting solution of the Glasser inequality, the HHI can be simply reformulated in terms of the Gini index by substituting it for the coefficient of variation in the initial HHI formulation:

$$HHI = \frac{G^2 + 1}{n} \qquad 4.3$$

b) The HHI exists in terms of the Gini index because it replicates the reflection of the Gini index about the Lorenz curve, as well as thereby also replicating the symmetry of the Lorenz curve. This is captured in its re-expression in terms of the Gini index:

$$G^2 = nHHI - 1 \Rightarrow G = \sqrt{nHHI - 1} \qquad 4.4$$

c) In practical terms the HHI can then be calculated from the Gini index. This will then make it a robust measure of business concentration that can accurately estimate such concentration under perfect as well as imperfect data conditions. One accurate, simple, and fast estimation technique for the HHI, which uses the ranks and values of the data in the same way as the Gini index, is:

$$HHI = \frac{4\left[\sum_{i=1}^{n} ix_i\right]^2 - 4(n+1)\sum_{i=1}^{n} ix_i + (n+1)^2 + n^2}{n^3} \qquad 4.5$$

d) If for some or another reason, we have no reliable information on all market shares, we can keep the robust status of the HHI, by estimating it from the range, using the market shares of the largest and smallest firms, based on the following expression:

$$HHI = \frac{r^2(n+1)^2 + 16}{16n} \qquad 4.6$$

e) Because in the limit, as the number of observations increases, the Gini index and the coefficient of variation are equal, McKay's approximation in terms of its confidence intervals also extends to the Gini index. Firstly this means that the approximate sampling distribution of the HHI is the Chi-square distribution. Secondly it means that we can actually derive reliable confidence limits for the HHI, from McKay's confidence intervals. McKay's original confidence interval for the HHI is given by:

$$\Lambda_1^{HHI} = \left(\frac{\sqrt{nHHI-1}}{\sqrt{(nHHI-1)\bullet\left|\left(\frac{X_u^2}{n}-1\right)\right|+\frac{X_u^2}{n-1}}}, \frac{\sqrt{nHHI-1}}{\sqrt{(nHHI-1)\bullet\left|\left(\frac{X_l^2}{n}-1\right)\right|+\frac{X_l^2}{n-1}}} \right) \quad 4.7$$

McKay's modified confidence interval for the HHI is given by:

$$\Lambda_2^{HHI} = \left(\frac{\sqrt{nHHI-1}}{\sqrt{(nHHI-1)\bullet\left|\left(\frac{2+X_u^2}{n}-1\right)\right|+\frac{X_u^2}{n-1}}}, \frac{\sqrt{nHHI-1}}{\sqrt{(nHHI-1)\bullet\left|\left(\frac{2+X_l^2}{n}-1\right)\right|+\frac{X_l^2}{n-1}}} \right) \quad 4.8$$

For both intervals the degrees of freedom are the number of firms subtracted from one ($df = n-1$). Critical Chi-square values for the lower and upper confidence limits can be obtained from any Chi-square table such as that produced in the Appendix to Chapter 2.

f) To find out if the observed HHI estimate is accurate, take the width (W) of the confidence intervals, as the difference between their lower (L) and upper (U) limits:

$$W = U - L \quad 4.9$$

We know that the interval with the smallest width is the accurate one. To find out the bias (B) of an estimate, take half the width of anyone or both of the intervals:

$$B = \frac{1}{2}W \quad 4.10$$

To obtain the expected value of the HHI (E) from the intervals, take the average of their limits:

$$E = \frac{L+U}{2} \quad 4.11$$

g) If either of the confidence intervals yield bias in excess of 25%, then it is necessary to correct their limits, by adding to the lower limit and subtracting from the upper limit a corrective amount (K), given by:

$$K = \frac{|W - 50\%|}{2} \quad 4.12$$

This adjustment to the limits decreases the bias in the estimate of the expected value in terms of pooling it down to the tolerable limit of 25%. At this level, we know that the bias is still negligible not to have any real influence on the estimate.

h) The HHI is a statistical decision-making tool for business concentration, because its sampling distribution is approximated by the Chi-square distribution. Because the HHI subscribes to this distribution, either under the Gini index or the coefficient of variation, there is no reason or excuse not to work with the HHI in terms of the Gini index.

i) With McKay's confidence intervals for the HHI, we can statistically determine its accuracy, or test hypotheses on business concentration in terms of: the degree of monopolisation in markets; the nature of competition; and the likely forms such competition assumes. Such evaluations of accuracy can be especially relevant to a statistical agency wanting to disclose the reliability of HHI estimates as an official statistic. In terms of hypotheses tests, the intervals can be particularly valuable to economists, business executives, and regulators, all of which use the index interpretively.

Donald is happy with the findings that emerge out of the Question and Answer session. He can now give Amelia the information she has asked him for. As for Caroline – she is relieved. She feels that her decision to withhold the HHI discussion with Stan until after the workshop has now given her an opportunity to have this discussion on positive rather than negative terms.

The day after the workshop, Caroline calls a meeting with John, to discuss with him the discoveries from the Question and Answer session, and share with him all the new findings on the HHI that his investigation did not reveal. As John becomes exposed to this knowledge he becomes excited. He learns that there are things about the HHI he knew nothing about, and at the same time he also learns that he has to change his perceptions about the index. Together with Caroline he is eager to apply the new findings on the chocolate industry snapshots before they meet Stan. They want to advise Stan that as an official statistic the HHI can be determined precisely with minimal error.

Caroline and John want to give Stan the confidence that the official HHI number is reliable or trustworthy. As a result they decide to work with a margin of error (or significance level) of 1%, giving them a 99% surety (or confidence level) in their range of estimates for the HHI. Because they have the market shares of all firms in the industry, they decide to estimate the HHI from the ranks and values of the data. Using a handheld calculator, after 30 minutes, Caroline and John produce their calculations in Table 7. They are satisfied with the results. They agree that the use of both McKay intervals is necessary for the accurate estimation of the HHI since it is by their joint use that it becomes possible to calculate an accurate estimate of the true HHI value. On the basis of the width and bias of the intervals they find that the original McKay confidence interval is more accurate than the modified McKay interval. In turn they decide to focus their data analysis on the former interval. From this interval, they find that the pilot estimates the agency has produced are below the bias criterion of 25%. This is the practical cut-off limit in terms of which the bias of an expected estimate should not exceed 25% from its true value. Their bias numbers are 19.9% for year 1, 18.3% for year 2, and 19.0% for the data of the annualised market shares. All these numbers are within the 25% cut-off limit making it safe to conclude that the agency's pilot estimates are not biased.

Table 7: HHI for the chocolate industry (%)

Calculations in respect of:	***Year 1***	***Year 2***	***Annualised market shares***
HHI (from expression 4.5)	17.1	16.9	17.0
Number of firms	6	6	6
Degrees of freedom	5	5	5
X_l^2 at 1% significance level	0.41	0.41	0.41
X_u^2 at 1% significance level	16.75	16.75	16.75
99% confidence interval for HHI (original McKay interval from expression 4.7)	8.5 , 48.3	7.6 , 44.2	7.9 , 45.9
Width (from expression 4.9)	39.8	36.6	38.0
Bias (from expression 4.10)	19.9	18.3	19.0
99% confidence interval for HHI (modified McKay interval from expression 4.8)	8.5 , 50.3	7.6 , 45.7	7.9 , 47.7
Width (from expression 4.9)	41.8	38.1	39.8
Bias (from expression 4.10)	20.9	19.0	19.9
Expected value (of original McKay interval, from expression 4.11)	28.4	25.9	26.9

The results in Table 7 bode very well for the agency. They show that it can produce reliable HHI estimates because it can do so without introducing systematic error. Caroline and John are particularly happy about this outcome. In the present case it means that they can be 99% confident that the range of values they have calculated for the HHI represent its true value. By extension this makes them equally confident that the expected value of the HHI, as calculated from these intervals, has been accurately determined. This is important because the expected value is the true value, on the average, in the absence of bias. It is closely approximated whenever the influence of bias is trivial. Accordingly there is a 99% chance or confidence that this expectation, or average, is a precise estimate of the HHI's true value.

Emboldened, John and Caroline meet with Stan to share their results. Stan is impressed. What he finds most interesting is that the HHI has confidence intervals. This is an important development because the Competition Regulator will now be able to monitor its HHI thresholds with the help of statistical decision-making reliant on the HHI itself. Such a development is scientific, bringing with it objectivity to decision-making, by introducing considerations that have to be made, analysed, checked and rechecked with recourse to the balance of probabilities, as opposed to feelings, intuition, preconceptions, or prejudices. Stan is concerned about all of these. He wants to ensure that the monitoring and enforcement of competition regulations by the HHI is known to be impartial, and the decisions or judgments by the Competition Regulator based on it are known to be fair, and as much as possible, beyond reproach.

He wants to find out how this could work in practice with the statistical decision-making abilities of the HHI. With the knowledge he has received from Caroline and John he decides to test this out on his current investigation of the chocolate industry. Stan is not alone in his quest. Amelia – the chief executive officer to one of the companies in that industry – has equally become interested in the HHI. The business daily she reads every morning had been running a number of stories, advising business executives not to be complacent about their companies conduct, and especially to monitor how such conduct promotes or stifles rivalry in the markets they work in. The business daily had warned that if companies ignore this, they risk being the subject of enquiries or

investigations by the Competition Regulator. From the stories Amelia also learns that business concentration is monitored by the HHI. Amelia becomes particularly interested in the HHI because her firm – the largest by market share in the chocolate industry – is in the middle of merger talks with the second largest firm in the same industry. She does not know how that prospective merger affects the HHI. In addition she also does not know what the HHI is, except that it is some kind of an index for business concentration. This is why Amelia asked Donald to prepare a brief with the HHI's industry numbers that can help her answer these questions. Donald too did not know anything about the HHI. But after attending the workshop things became clearer, and he was able to produce estimates for the HHI at a 99% confidence level, using the latest information on firm market shares, from the last two years of operation in the chocolate industry. The estimates are identical to those by Caroline and John in Table 7. Donald is intent on giving Amelia numbers that are as much as possible free of error. He too knows that Amelia is in merger negotiations and does want to take the risk of providing numbers that somehow could mislead Amelia into taking the wrong decision.

On the way to handing over his brief to Amelia, Donald invites Rebecca – the firm's chief economist – to join her. Donald feels unequipped about handling possible questions concerning the use of the HHI into economic and regulatory analysis, while Rebecca's expertise is in these areas. He also knows that economists refer to the HHI quite often whenever they discuss competition issues or anything that has to do with monopolisation and the practices of the Competition Regulator. Rebecca is surprised why would there be a need for special brief and a special meeting with Amelia, on what is after all, an ordinary index of business concentration that is really quite simple to understand, and very easy to calculate. She even tells Donald that if he would give her the information on market shares he has, she would quite gladly calculate the index herself by summing their squares, and then just send everyone a quick email with the results, saving everyone the time and trouble for a meeting. But as they walk to Amelia's office, and Donald tells her about the new findings on the HHI from a workshop he had attended, Rebecca's attitude begins to change. She comes to understand that there are things about the index that her student textbooks or economic articles did not tell her anything about, as well as that her

profession was unaware of them. She is particularly excited that the HHI can be a hypothesis testing instrument with which to probe the type of competition markets operate in, as well as to probe what form of rivalry exists between firms. She feels that if she is in a position to give an objective quantitative view on these matters to Amelia, then her firm will be in a position to make smart strategic decisions about whether to enter new markets, or whether to keep its presence in existing ones. Donald and Rebecca have now reached Amelia's office and are getting ready to go in.

6.3 DECISION-MAKING

Amelia, Donald, and Rebecca are locked in their meeting. At the start, Donald tells Amelia that the yearly numbers are not nearly as important as those for the annualised market shares, because in any given year market shares fluctuate, so that in order to obtain a stable picture of their behaviour, they should be looked at when they are annualised. This has the effect of ironing out the fluctuations, and allows for the study of numbers that are not time-dependent. Rebecca agrees. Based on Donald having used a 1% significance level, she tells Amelia that she can be 99% sure that his *figures* are accurate, with only a 1% chance that they are not. But she sees that Donald's technical-looking table contains too much information, which Amelia is finding hard to read. She recreates Table 7 so that it includes only the applicable information on annualised market shares (Table 8).

Table 8: Summarised account of HHI for the chocolate industry (%)

Calculations in respect of:	***Annualised market shares***
HHI	17.0
99% confidence interval for HHI	7.9 , 45.9
Width	38.0
Bias	19.0
Expected value (average of the interval's limits)	26.9

She also simplifies Donald's technical presentation on how the HHI is calculated, explaining to Amelia that the HHI – which aims to show how alike or how different firms' market shares are – is just a slightly different version of the Gini index. Rebecca tells Amelia that concentration measures

like the HHI are widely used in economic analysis and for the purpose of applying Competition Policy, as they indicate the degree of competition or monopolisation in markets. She recalls coming across the following abbreviated explanation on market concentration in the 2006 edition of the *Collins Dictionary of Economics*:

> "The significance of market concentration for market analysis lies in its effect on the nature and intensity of competition. Structurally, as the level of seller concentration in a market progressively increases, "competition between the many" becomes "competition between the few" until, at the extreme, the market is totally monopolised by a single supplier. In terms of market conduct, as supply becomes concentrated in fewer and fewer hands (oligopoly), suppliers may seek to avoid mutually ruinous price competition and channel their main marketing efforts into sales promotion and product innovation, activities that are more profitable and effective way of establishing competitive advantage over rivals."

Using the above economic terminology, Rebecca explains to Amelia that economists *expect* that when the level of business concentration in a market increases – that is, when the HHI goes up – competition between many firms turns into competition between few firms, and at the extreme, may result in one firm monopolising the industry. She also points out to Amelia that in terms of market conduct, as supply becomes concentrated in fewer and fewer hands – what economists call oligopoly – suppliers are more *likely* to seek to avoid price wars as well as more *likely* to channel their main marketing efforts into sales promotions, various branding tactics, and product innovations. Amelia knows that too. In the business world these are activities proven to be more profitable. They are known to create *possibilities* that may give more effective ways to establish a competitive advantage over rivals. Donald interrupts Amelia and Rebecca to advise that with the 99% confidence interval they could test their theories on the chocolate industry.

After a short course in econometrics, Rebecca immediately recognises how to do this. She sets about to test her hypothesis that, as there are only six firms operating in the chocolate industry,

she is convinced that there is only competition among the few – and that therefore the market structure of the industry is one of oligopoly. She proposes to test the following hypothesis:

Ho: My firm's industry is characterised by "competition between the many", vs.

Ha: My firm's industry is characterised by "competition between the few"

From Table 8 she takes the observed HHI value for this test and for the industry as a whole, which is 17.0%. Rebecca recognises that *if* the firms' market shares are all equal, implying that they have a uniform distribution, then the minimum value for the HHI would be the reciprocal of the number of firms, or 16.7%. Rebecca rightly treats this value as invalid for the present case since the market share distribution here is positively skewed. In turn she proceeds to take the range of HHI values from the 99% confidence interval, which is between 7.9% and 45.9%. As the observed HHI value falls within the confidence interval limits the null hypothesis cannot be rejected. Rebecca is surprised to have her preconceptions so convincingly challenged, but has no choice but to accept a result that, with reference to probabilities, demonstrates objectively what is scientifically most *probable.*

For Amelia, the finding is not as surprising because she knows that even among a small number of firms rivalry can be intense. However, Amelia has also believed for some time that because of this intensity, firms in the industry have engaged in too many ruinous price wars that have affected the profitability of all the firms and therefore the industry itself. She has been asking her board of directors to abandon crippling price competition and adopt a new, more forward-looking strategy. However the board is reluctant to act on her proposal, arguing that they need concrete, objective, scientific evidence on which to base their decision to make what seems like an unprecedented and risky move. Amelia now starts to think that that she can convince them if she uses a statistical test based on the HHI to provide the required evidence. She decides to test whether her firm indeed operates in an industry environment of crippling price competition, or whether firms are actually competing on the basis of other non-price related tactics. Depending on the outcome of

this test she will know what strategic direction to recommend to the board. She tests the following hypothesis:

Ho: Our firm operates in an industry characterised by price rivalry, vs.

Ha: Our firm operates in an industry characterised by non-price rivalry

Like Rebecca, Amelia refers to Table 8 and takes the observed HHI value for this test, which is the HHI for the industry as a whole. The HHI is 17.0%. She also takes the range of HHI values from the 99% confidence interval, which is between 7.9% and 45.9%. As the observed HHI value falls within the confidence interval limits, the null hypothesis cannot be rejected. There is a 99% chance that Amelia is right about this conclusion. She considers this convincing enough to argue at the next boardroom meeting that the firm must change its strategic sales direction to non-price rivalry if it is to become more profitable and still be in business in the next decade. In spite of this she has reservations that the board of directors may regard the HHI test as an unorthodox tool by which to argue for a change in strategic sales direction. To counteract this, Amelia decides that she will advise the board that this is an unusual application of the test, in spite of which it seemingly suggests that there is a need for a change in strategic sales direction. Ultimately whether the direction suggested by the test is practically significant is up to the professional judgment of the board.

Amelia is also reassured by Rebecca's test that she is unlikely to have to battle with the Competition Regulator over the merger between her firm and the second biggest firm in the industry. This is because Rebecca's hypothesis test found that there is competition between many firms in the industry. However, she knows that the small number of operators in the industry y, which is about to get even smaller if the merger takes place, could create perceptions to the contrary. In response, she asks Rebecca to test the hypothesis whether the degree of monopolisation in the industry is harmful to competition. Rebecca responds by testing at the 1% significance level in order to assure the Competition Regulator that utmost care has been taken to examine this question. She tests the following hypothesis:

Ho: The degree of monopolisation in the industry is not harmful to competition in the industry, vs.

Ha: The degree of monopolisation in the industry is harmful to competition in the industry

Using the results in Table 8, Rebecca compares the HHI value of 17.0% to the range of the confidence interval and finds that it falls between its limits of 7.9% and 45.9%. She concludes that the null hypothesis cannot be rejected and advises Amelia that the HHI statistical test for monopolisation indicates that the extent of monopolisation does not stifle competition. *On the basis of the confidence interval* in Table 8, she also advises Amelia that the expected HHI value of the test is 26.9% and that in effect the statistical test *does not* find this value to be significantly different from the 25% threshold set by Regulators. This is because the threshold and the expected value happen to lie inside the confidence interval within a distance of 2 percentage points from each other. Rebecca knows that for Regulators, highly-concentrated markets in which competition has the potential to be stifled outright, have HHI levels above 25%. Being conservative she has made up her mind to test for the degree of monopolisation under the most stringent threshold level, and in response she tests at that level only. She is aware that testing at less stringent levels can also be done, although she is concerned that this may lead her to give an over-promising prognosis to Amelia in terms of the Regulator's strictest position on the degree of monopolisation. She is unprepared to commit to this.

Returning to the 25% threshold, *which is substitutable for any other threshold level*, Rebecca notes that if the degree of monopolisation is to be regarded as not harmful to the rest of the industry players, there must be no evidence that this threshold has been breached. The statistical test does not show such evidence.

Amelia is now confident that the results of the HHI statistical tests provide solid scientific support with 99% confidence that the merger between her firm and the second biggest firm, will not breach the HHI threshold of the Competition Regulator, and that the resulting degree of monopolisation will not debilitate other industry players. She submits these arguments to the Competition

Regulator, together with the accompanying results from the tests, asking that the merger be permitted to take place.

6.4 CONSEQUENCES

Stan has received Amelia's submission. His meeting with Caroline and John has also taught him how to calculate and test by the Gini-based method of HHI. He has concerns about Amelia's submission, because in his view she has conducted the statistical tests on the assumption that the merger in the industry is not taking place since all firms are still separately accounted for. He thinks that the relevant analysis is the one that applies to the industry under the scenario that the merger has occurred. In this case the composition of the industry changes. He observes that in that case the number of firms will decline from 6 to 5, and the firms' shares will change. From Table 6 he constructs Table 9.

Table 9: HHI of the chocolate industry after the merger of its biggest firms (%)

Firms	*Annualised market shares*	*Calculations in respect of:*	*Annualised market shares*
Firm A	8.3	HHI (from expression 4.5)	21.6
Firm B	15.4	Number of firms	5
Firm C	16.3	Degrees of freedom	4
Firm D	17.7	X_l^2 at 1% significance level	0.21
Firms E+F	42.3	X_u^2 at 1% significance level	14.86
TOTAL	100.0	99% confidence interval (original McKay)	14.3 , 78.5
		Width	64.2
		Bias	32.1
		Corrected 99% confidence interval - original McKay interval (by expression 4.7)	21.4 , 71.4
		Width	50.0
		Bias	25.0
		Expected value (original McKay interval)	46.4
		99% confidence interval (modified McKay)	14.2 , 90.5
		Width	76.3
		Bias	38.2
		Corrected 99% confidence interval - modified McKay interval (by expression 4.8)	27.4 , 77.3
		Width	49.9
		Bias	24.9
		Expected value (modified McKay interval)	52.4

In Table 9, Stan includes the calculations of the HHI and its confidence limits in response to the changed market shares. Stan takes 40 minutes to do the calculations with a handheld calculator. He has not done them before and has had to rely on the notes by Caroline and John showing him step-by-step how to do them. He thinks that they could be done much faster if they were programmed.

During the calculations Stan realises that his initial estimates for the range of the HHI are biased. He notes that the original McKay and modified McKay confidence intervals show larger than acceptable bias – in excess of 25% – in respect of the expected value of the HHI. He responds properly by correcting the limits of the confidence intervals, increasing the lower limit and decreasing the upper limit by equal amounts. He does this until the bias arising from either interval is decreased within the limit of 25% where it is small enough not to matter. He finds that this adjustment makes the modified McKay interval slightly more accurate than the original McKay interval, with its width and bias being just a notch lower. Like Caroline and John this leads Stan to conclude that both intervals need to be calculated in practice in order to determine which will give the most accurate estimates. Because the corrected version of the modified McKay interval produces the least amount of bias and closest range in the expected value for the HHI, Stan decides to do his analysis and hypotheses on this interval. First, Stan decides to formulate his hypothesis and to do so he does some research. He finds that according to the 2006 edition of the *Collins Dictionary of Economics*:

> "Market theory predicts that market performance will differ according to whether there are many suppliers in the market, each accounting for only a small fraction of total supply (Perfect Competition), or only a few suppliers, each accounting for a substantial portion of total supply (Oligopoly) or a single supplier (Monopoly)."

Stan understands the meaning of what he has just read because the Competition Regulator considers the degree of monopolisation in a market to stifle or eliminate its degree of competition whenever the HHI exceeds 25%. Stan also knows that the Competition Regulator is of the view

that lower HHI values are more consistent with perfect competition, and are thereby deemed to indicate that the degree of monopolisation in a market is either less harmful or not harmful to rivalry among suppliers (i.e. firms). Stan therefore sets about formulating and testing two hypotheses. The first hypothesis is about market structure:

> Ho: The market structure of the chocolate industry is one of perfect competition, vs.
>
> Ha: The market structure of the chocolate industry is different from perfect competition

Stan refers to Table 9 to read off the values for this test at the 1% significance level. He too wants to be accurate in his decision-making, in order to be on par with Amelia. He takes the estimated HHI value for the chocolate industry on the assumption that the merger has occurred. This value is 21.6%. He then compares this value to the 99% confidence limits of the bias-corrected confidence interval for the HHI, which are between 27.4% and 77.3%. As the estimated HHI value falls outside the range of the confidence interval he concludes that the null hypothesis of perfect competition must be rejected. This is a straightforward application of statistical testing with a confidence interval. As Smithson (2000: 177) explains:

> "In general, any value contained inside a confidence interval around a sample statistic is …a plausible value of the … population statistic, and any value outside the interval is … an implausible value."

Under the present *illustrative* statistical test, there is a 99% chance that the industry structure of perfect competition would not exist if the merger takes place. Apart from this, Stan also concludes that the test indicates that the merger will make the industry an oligopoly rather than monopoly because there is still more than one firm operating in it. In any event Stan is concerned that there is now a degree of monopolisation that has set-in because the market structure has moved away from perfect competition. He wants to test whether this degree of monopolisation will harm competition, i.e. rivalry, in the industry. He has two equivalent hypotheses to consider in this regard. The one hypothesis he could test is this:

Ho: The degree of monopolisation in the chocolate industry is not harmful to competition, vs.

Ha: The degree of monopolisation in the chocolate industry is harmful to competition

The other – equivalent – hypothesis is based on whether the borderline 25% regulatory threshold for the HHI is met. Here the hypothesis will be:

Ho: The degree of monopolisation in a market is borderline harmful to its degree of competition, vs.

Ha: The degree of monopolisation in a market may or may not exceed its degree of competition

Stan decides to test the latter hypothesis because it is based on the regulatory thresholds for the HHI. Keeping up with his 1% margin of error, Stan refers to Table 9 and compares the observed HHI estimate of 21.6% to the confidence limits for the HHI, which are between 27.4% and 77.3%. He concludes that the null hypothesis can be rejected. This tells him that there is a 99% chance that the degree of monopolisation in the industry is different from the 25% benchmark required by regulation. He again refers to Table 9 and finds that the expected HHI value for the test is 52.4%. He correctly concludes this to mean that the HHI test tells him with 99% certainty, that this expected HHI value does not equal the regulatory threshold of 25%. This threshold lies outside the confidence interval relative to the expected value, being twice as far away from it than it should be if they were the same. This tells Stan that the expected degree of monopolisation the merger will produce is significantly higher than the regulatory threshold. It exceeds the intensity or degree of competition in the industry, and in effect it will retard competition. The statistical test indicates that there is a 99% probability that this will happen under the assumption of the merger having occurred.

On the basis of the statistical evidence, which on the balance of probabilities Stan considers is convincing, he decides to call a meeting with Amelia to present his findings and outline the

conclusions they lead to. He wants to explain why the Competition Regulator will not allow the merger to go through. He is relieved that the statistical analysis produces findings that are impersonal, and allow him to objectively outline with reference to factual probabilities, why the Competition Regulator will not permit the merger to proceed. He feels that in this way, while Amelia may dislike what she hears, she will at least appreciate that by paying attention only to what is probable and likely, the Competition Regulator is being objective and impartial as to how it has reached its decision.

Amelia too has some decisions to make. At the next boardroom meeting she could advise her board that the road to profitability by merging her firm with the second largest firm is now closed, and it is imperative that the board approve the change in the strategic sales direction she has been asking them to approve for some time. Or she could approach the Competition Regulator to find out whether the merger would be allowed if the joint firm dilutes its market share, by divesting from some of its operations. On the sidelines she has already asked Donald and Rebecca to conduct the necessary HHI hypothesis tests for these scenarios.

This concludes the statistical story aimed at illuminating the usefulness and practical applications of a Gini-based method of calculating HHI. As Blessing *et al.* (2005: 2) observed:

> "Statistics can tell people something about the world they live in. But not everyone is adept at understanding statistics by themselves. Consequently, statistical stories can, and must, provide a helping hand."

This is what our statistical story here does. It provides a helping hand to show without complexity that the HHI is indeed a statistical decision-making tool. It is an index with many possibilities of relevance to decision-makers in different contexts whenever they have to deal with the issue of business concentration or its estimation. We find that far from having the virtue of simplicity it is assumed to have, the HHI is actually an index intimately connected to the Gini index, and therefore one for which expectations (expected values) do exist. These values can be obtained by reference

to the Chi-square distribution, which approximates the sampling distribution of the HHI. The index is subordinate to the balance of probabilities to the extent that its statistical significance can be verified in any practical situation it is applied to, without the need to doubt whether the numbers are believable. And, perhaps most importantly, it is an index that can test the statistical significance of the context it is subjected to.

6.5 SUMMARY

By engaging in transnumeration, i.e. statistical story-telling, in this chapter, a number of novel contributions have been made and shown:

a) The HHI is a statistical decision-making tool. It is an index with many possibilities of relevance to decision-makers in different contexts whenever they have to deal with the issue of business concentration or its estimation.
b) The HHI is intimately connected to the Gini index. It is an index for which expectations exist, and their associated values can be obtained by referral to the Chi-square distribution, which approximates the sampling distribution of the HHI.
c) The index is subordinate to the balance of probabilities to the extent that its statistical significance can be verified in any practical situation it is applied to without the need for doubt about the credibility or plausibility of the numbers.
d) In terms of measurement, the HHI will be mis-reported if it is only reported by itself without its confidence limits. There is now a way to determine the accuracy of the estimates.
e) Estimation with confidence intervals supports and influences decision-making through hypothesis testing. As a decision-making tool, the HHI makes it possible to study business concentration directly in terms of tests and hypotheses related to the Chi-square distribution.
f) Ultimately, because the HHI follows the Chi-square distribution, it must be regarded and treated as a statistic and a test procedure all in one. For this alone it more than adequately qualifies as an official statistic of business concentration.

g) Most important of all: it is an index that can test the statistical significance of the context it is subjected to. In terms of this, the HHI can be accurately estimated together with confidence limits, and this is quickly and effectively achieved when the HHI is treated as a robust measure of business concentration. This in effect creates a situation in terms of which only the Gini representation of the HHI should be used for the measurement of business concentration. This should not be hard to accept, because as seen, the HHI is just a variant of the Gini index, or to put it differently, the Gini index is just another version of the HHI.

The crux of the matter is that we now know that the HHI can be accurately estimated with confidence limits. It should by now be clear that this is preferable if done using the Gini index. In any other depiction the HHI has neither reliable statistical representation nor reliable confidence intervals by which to determine the accuracy of the estimates for its expected values. The accuracy improves if the Gini representation is used because the Gini index is a known robust estimator of relative variability. This in effect creates a situation in terms of which only the Gini representation of the HHI should be used for the measurement of business concentration. Considering that after all the HHI is just a variant of the Gini index, the only alternative would be to face up to Mandelbrot's critique and accept that HHI estimates derived from the coefficient of variation contain an inherent upward bias.

We have also seen that the HHI will be misreported if it is only reported by itself without its confidence limits. There is nothing much that can be done about past practices of this sort, and they are not to be faulted for having done so, because up to now they have operated on the premise that no method or methods for such reporting exist. Future practices however should be held accountable to such reporting requirement in light of the demonstration in this story that there is a way to ascertain the accuracy of the estimates. We also saw that these estimates support and influence decision-making through hypothesis testing. Their use in this way will also serve to answer Mandelbrot's concern as to why the HHI should be taken seriously and its numbers believed. As a decision-making tool related to the Chi-square distribution, the HHI makes it possible to study business concentration directly in terms of what is statistically probable rather

than presumed. As such, it has the potential to be an invaluable tool for decision-makers in terms of allowing them to examine their preconceptions and test their assumptions about business concentration.

Thus concludes the transnumerative illustration of the statistical relevance and practical applications of the HHI index of business concentration. As we saw from this illustration, Taguchi's transnumeration epitomises Bellman's description referred to earlier, that "mathematics is not a science, it is an art". Chapter 7 follows in summarising the findings and conclusions of this enquiry.

7. CONCLUSION

This enquiry began by noting that this will be a contribution in the field of business statistics, which is a branch of applied statistics. In that vein, we employed statistical theory to formulate and solve the problem of the enquiry, which was to examine or determine whether the Herfindahl-Hirschman Index (HHI) of business concentration, is a statistically-relevant index. It was recognised from the start that the nature of the problem and the enquiry itself will have applications in areas such as economics, business administration, and public administration as concerns the publication of the HHI as an official statistic by a statistical agency. The enquiry, and specifically the problem it deals with, found an early companion in the critique by Benoit Mandelbrot. In terms of this critique, Mandelbrot demonstrates the inherent weakness of the HHI and dismisses it as a statistically-flawed measure in that it overstates business concentration levels – raising the question whether it should be used at all.

At the heart of the criticism is the legitimate concern that by its original or conventional formulation the HHI is questionably formulated in using the coefficient of variation, which overlooks two renowned results in mathematical statistics known as the De Vergottini and Glasser inequalities. According to these inequalities the coefficient of variation will always give elevated measurements of relative variability compared to the Gini index, except in the limit when with many observations, both measures of relative variability will be equal. The coefficient of variation and the Gini index are aptly also known as the noise-to-signal ratios because they describe the relative precision of estimates. They show how much of their typical value, i.e. the average, is corrupted by its dissimilarity from individual estimates, whether captured by the standard deviation in the case of the coefficient of variation, or the covariance in the case of the Gini index. Thus the typical value is regarded as the “signal”, and the dissimilarities from it are the “noise”.

In terms of the Glasser inequality the coefficient of variation is known to produce estimates with more “noise” compared to the Gini index. It then certainly becomes nigh impossible to accept the HHI is an accurate measure of business concentration because it is potentially susceptible to

inflated levels, even if it could be argued that this shortcoming is not an issue with a growing number of observations. But the point simply is that it if we formulate the HHI in terms of the Gini index it will simply not overstate the measurement of business concentration irrespective of the data conditions encountered.

Fortunately, it does not take much to resolve this and reformulate the index to eliminate the problem. The solution of the Glasser inequality that the Gini index and the coefficient of variation are equal as the number of observations increases leads us straight to the finding – heretofore unknown – that the HHI is a variant of the Gini index. In effect this finding is an extension of the Glasser inequality. It is thus another important and special case of that inequality. In demonstrating this finding, a number of other new things were also demonstrated:

a) Milanovic's incomplete proof for the asymptotic equality between the Gini index and the coefficient of variation was completed. In this way a new simplified proof for the Glasser inequality was reached. For example, Piesch's thorough survey of the Glasser inequality shows that no other equally simple proof has so far been derived.
b) McKay's original and modified confidence intervals based on the Chi-square distribution, which are so far only applied to the coefficient of variation, are immediately extendable to the Gini index, and by analogy to the HHI. As concerns the Gini index this helps resolve a puzzle about its sampling distribution. Kamat and Ramasubban long ago found this to be the Chi-square distribution. But they judged this to be an empirical regularity. It is in fact inherited from the coefficient of variation. Because the two are asymptotically equal, in turn they share the same sampling distribution. Furthermore neither Kamat nor Ramasubban were ever able to provide confidence intervals for the Gini index based on this distribution. The extension of McKay's confidence intervals to the Gini index closes this gap. As concerns the HHI, prior to this enquiry neither its sampling distribution nor the fact that it has confidence limits was known.
c) From McKay's confidence intervals for the HHI we can infer that the approximate sampling distribution of the HHI is the Chi-square distribution with n-1 degrees of freedom. The

subordination of the HHI to the Chi-square distribution diminishes the need for simulation studies that seek to find what this distribution is. As we saw in the present enquiry, they do not yield anything that we do not already know from the asymptotic equality between the Gini index and the coefficient of variation.

d) The accuracy of any HHI estimate can now be studied in terms of its precision and the extent of its bias. In short the HHI is not just a descriptive index. It can be fully used as a statistical decision-making tool to test the statistical significance of the context it is subjected to. Through this statistical ability it is enabled to serve as another tool by which to conduct confirmatory and exploratory studies of economic and business theories as well as regulatory enforcement in the area of business concentration. In this respect the current work responds to Adelman's neglected call for a statistical framework that can do this. The HHI confidence intervals and their relation to the Chi-square distribution provide such a framework.

e) While the HHI confidence intervals do not prohibit its calculation by the conventional formula as the sum of squared market shares, it is now apparent that every time we opt for such a calculation, we also diminish its accuracy, as well as that of its confidence levels, because the index then has no robustness. This only leaves calculation of the HHI by its reformulation in terms of the Gini index. In essence we have to accept that if we want to make use of the statistical abilities of the index we also have to accept that our calculation of it has to change accordingly. This is neither impossible nor difficult to do.

In addressing the question of the statistical relevance of the HHI it was found that because the HHI is a variant of the Gini index, by mere rearrangement the latter can be depicted as an amended version of the HHI. Not only was a new formula for the HHI obtained, but at the same time a new formula for the Gini index. This alternative depiction enables us to recognise that the HHI replicates the symmetry of the Gini index on both sides of the Lorenz curve, which reaffirms the connection between the two. This intertwined relationship between the two indices also made it easy to understand that the supposed distinction between absolute and relative measures of business concentration does not apply in the case of the HHI simply because the HHI is connected to the

Lorenz curve in the same way as the Gini index is. The implication of this is that as an example of an absolute measure of business concentration, the HHI can be converted into a relative measure of business concentration, and vice versa. There are no boundaries between the two. By extension, because by formulation the Gini index is a robust measure of relative variability, the HHI can be made equally robust. In this way we confirmed that the HHI will not be affected by imperfect data conditions, and will perform equally accurately under perfect data conditions, when reformulated by the Gini index as opposed to the coefficient of variation. In this regard a number of estimation techniques for the HHI were outlined, which preserve its robustness. According to one of them, the HHI can be estimated from the ranks and values of the data in the same way as the Gini index is. In the process a new, simple proof for the very popular Lerman and Yitzhaki method was reached, as well as that of its extension by Ogwang – proofs or derivations that were absent from these original contributions.

According to the other technique, which combines the equality between the Gini index and the coefficient of variation with the relationship between the range and the standard deviation of the data, the HHI can be estimated from the range whenever we have no confidence in the market shares of all firms, except for their numbers from the largest and smallest firm. Principally it was highlighted that there are more than a dozen alternative techniques by which the Gini index can be estimated, and this implies that there are that many for the HHI. Compared to the single conventional method for its estimation, there is now an explosion of estimation techniques by which the HHI can be computed. This was not the subject of investigation here, but is not to say that it may not be of interest for a future investigation, as a way of giving a catalogue of the available alternatives.

What matters is that by finding out that a robust representation of the HHI exists in terms of the Gini index, we also have demonstrated that the HHI is an accurate measure of business concentration. As it happens we come to learn of this from the introspective examination that Mandelbrot's critique forces on the HHI. Without it we would not know that the HHI and the Gini index can be interconnected. The practical implications of this are twofold. Firstly, whether on

account of Mandelbrot's critique or otherwise, we have now learned that the HHI can be accurately estimated and used to test hypotheses. Because it can be accurately measured in terms of computable and reportable confidence limits it qualifies as a publishable official statistic. Secondly, while the HHI is frequently used by competition regulators, economists, and business executives, its use no longer has to be predicated on treating the index as a descriptive instrument. It is a formal statistical decision-making tool connected to the Chi-square distribution. It needs to be treated as such. With its confidence intervals it provides a probability mechanism for testing any confirmatory and exploratory theories dealing with business concentration. In the same way it gives competition regulators an objective standard for monitoring the likelihood of their HHI thresholds being met. In short by knowing that the index is a statistical decision-making tool, the approach to using and calculating the index must evolve. It is no longer necessary to be assumed that the index has a single value, since the range of all its possible values can be determined – which has useful applications for scenario analysis under different business concentration conditions.

Because the HHI follows the Chi-square distribution, the HHI should be acknowledged as a statistic and a test procedure all in one. For this alone it more than adequately qualifies as an official statistic of business concentration. Consequentially the storyline of the present enquiry can be summarised in the following points:

a) The coefficient of variation is asymptotically equal to the Gini index;
b) As a result of this equality a Gini version of the HHI is constructed by replacing the coefficient of variation with the Gini index in the original formulation of the HHI;
c) The reformulation of the HHI in terms of the Gini index is a significant improvement on the regular HHI because it contains an unbiased measure of relative variability in terms of the Gini index;
d) Thus under skewed market share distributions prospective overstatement of the levels of business concentration is eliminated, while anyhow these levels remain accurately detected when, in the rare practical cases, the distributions are normal or symmetrical;

e) The reformulation of the HHI in terms of the Gini index preserves the Chi-square distribution as its sampling distribution, as well as its abilities to serve as a reliable statistic and a test procedure for business concentration;

f) In light of Mandelbrot's critique, revising the HHI in terms of the Gini index is not optional. It is necessary if the HHI is to have a reputation as a reliable index of business concentration.; and lastly

g) Re-expressed in terms of the Gini index, the HHI warrants a more elevated status as a frequently publishable official statistic by statistical agencies.

REFERENCES

Academic Press Dictionary of Science and Technology 1992, *Asymptotic formula*, Elsevier Science and Technology, Oxford, UK.

Acar, W. & Sankaran, K. 1999, "The myth of the unique decomposability: specializing the Herfindahl and entropy measures?", *Strategic Management Journal,* vol. 20, no. 10, pp. 969-975.

Adelman, M.A. 1969, "Comment on the “H” concentration measure as a numbers-equivalent", *Review of Economics and Statistics,* vol. 51, no. 1, pp. 99-101.

Andreosso, B. & Jacobson, D. 2005, *Industrial economics and organization: a European perspective,* 2nd edn, McGraw-Hill, Berkshire, UK.

Axtell, R.L. 2001, "Zipf distribution of U.S. firm sizes", *Science,* vol. 293, no. 5536, pp. 1818-1820.

Bellman, R. 1954, "Inequalities", *Mathematics Magazine,* vol. 28, no. 1, pp. 21-26.

Bendel, R.B. *et.al.* 1989, "Comparison of skewness coefficient, coefficient of variation, and Gini coefficient as inequality measures within populations", *Oecologia,* vol. 78, no. 3, pp. 394-400.

Bender, R.*et.al*. 2005, "Using confidence intervals in medical research", *Biometrical Journal,* vol. 47, no. 2, pp. 237-247.

Blessing, C. *et.al.* 2005, *Making data meaningful: a guide to writing stories about numbers*, United Nations Economic Commission for Europe, Geneva, Switzerland.

Boddy, R. & Smith, G. 2009, *Statistical methods in practice: for scientists and technologists,* John Wiley and Sons, Chichester, UK.

Book, S.A. 1979, "Why n-1 in the formula for the sample standard deviation? ", *Two-Year College Mathematics Journal,* vol. 10, no. 5, pp. 330-333.

Box, G. 1990, "A question of quality", *Harvard Business Review,* vol. 68, no. 2, pp. 225-228.

Box, G. 1988, "Signal-to-noise ratios, performance criteria, and transformations", *Technometrics,* vol. 30, no. 1, pp. 1-17.

Bronk, B.V. 1979, "Some inequalities for moments and coefficients of variation for a large class of probability functions", *Journal of Applied Probability,* vol. 16, no. 3, pp. 665-670.

Business Week 1998, *What do the trustbusters want?*.

Cabral, L. 2000, *Introduction to industrial organization,* 1st edn, MIT Press, Cambridge, USA.

Cai, T. 2005, "One-sided confidence intervals in discrete distributions", *Journal of Statistical Planning and Inference,* vol. 131, no. 1, pp. 63-88.

Carlton, D. & Israel, M. 2010, "Will the new guidelines clarify or obscure antitrust policy?", *Antitrust Source,* vol. 10, no. 1, pp. 1-4.

Carlton, D.W. & Perloff, J.M. 2000, *Modern industrial organization,* 3rd edn, Addison-Wesley, New York, USA.

Ceriani, L. & Verme, P. 2011, "The origins of the Gini index: extracts from Variabilita e Mutabilita (1912) by Corrado Gini", *Journal of Economic Inequality,* Special Issue, pp. 1-23.

Church, J. & Ware, R. 2000, *Industrial organisation: a strategic approach,* 1st edn, McGraw-Hill, Singapore.

Collins Dictionary of Economics 2006, *Concentration measures*, Collins, London, UK.

Collins Dictionary of Economics 2006, *Market concentration*, Collins, London, UK.

Collins Dictionary of Economics 2006, *Seller concentration*, Collins, London, UK.

Competition Commission. 2009, *Unleashing rivalry: ten years of enforcement by the South African competition authorities*, Competition Commission, Pretoria, SA.

Competition Commission. 2000, *Competition Commission report to the South African Reserve Bank on the proposed merger between NEDCOR and STANBIC*, Competition Commission, Pretoria, SA.

Competition Tribunal 2000, *Large merger between the Tongaat-Hulett Group and Transvaal Suiker Beperk*, Case No: 83/LM/Jul00, South Africa, Pretoria.

Conover, W.J. & Iman, R.L. 1981, "Rank transformations as a bridge between parametric and nonparametric statistics", *American Statistician,* vol. 35, no. 3, pp. 124-129.

Copeland, B.J. 1996, "What is computation?", *Synthese,* vol. 108, no. 3, pp. 335-359.

Croux, C. & Filzmoser, P. 2007, "Discussion of Morgenthaler's "A survey of robust statistics", *Statistical Methods and Applications,* vol. 15, no. 3, pp. 280-282.

Cumming, G. & Finch, S. 2005, "Confidence intervals and how to read pictures of data", *American Psychologist,* vol. 60, no. 2, pp. 170-180.

Cumming, G. & Finch, S. 2001, "A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions", *Educational and Psychological Measurement,* vol. 61, no. 4, pp. 532-574.

David, F.N. 1949, "Note on the application of Fisher's k-statistics", *Biometrika,* vol. 36, no. 3/4, pp. 383-393.

David, H.A. 1998, "Early sample measures of variability", *Statistical Science,* vol. 13, no. 4, pp. 368-377.

David, H.A. 1968, "Gini's mean difference rediscovered", *Biometrika,* vol. 55, no. 3, pp. 573-575.

De Vergottini, M. 1950, "Sugli indici di concentrazione", *Statistica,* vol. 10, no. 4, pp. 445-454.

Diaconis, P. & Efron, B. 1983, "Computer-intensive methods in statistics", *Scientific American,* vol. 248, no. 5, pp. 96-108.

Doyle, E. 2005, *The economic system,* John Wiley and Sons, Chichester, UK.

Du Plessis, P.G. 1979, "An international comparison of economic concentration: a note", *South African Journal of Economics,* vol. 47, no. 3, pp. 304-310.

Efron, B. 1988, "Bootstrap confidence intervals: good or bad?", *Psychological Bulletin,* vol. 104, no. 2, pp. 293-296.

Efron, B. & Gong, G. 1983, "A leisurely look at the bootstrap, the jackknife, and cross-validation", *American Statistician,* vol. 37, no. 1, pp. 36-48.

Efron, B. & Tibshirani, R. 1991, "Statistical data analysis in the computer age", *Science,* vol. 253, no. 5018, pp. 390-395.

Ehrenberg, A.S.C. 1982, "Writing technical papers and reports", *American Statistician,* vol. 36, no. 4, pp. 326-329.

Ehrenberg, A.S.C. 1981, "The problem of numeracy", *American Statistician,* vol. 35, no. 2, pp. 67-71.

Ehrenberg, A.S.C. 1977, "Rudiments of numeracy", *Journal of Royal Statistical Society,* vol. 140, no. 3, pp. 277-297.

Elvers, E. & Rosn, B. 1997, "Quality concept for official statistics " in *Encyclopaedia of statistical sciences*, eds. S. Kotz, C. Read, N. Balakrishnan & B. Vidakovic, John Wiley and Sons, New York, USA.

European Union. 2004, *Guidelines on the assessment of horizontal mergers under the Council Regulation on the control of concentrations between undertakings*, European Union, Luxembourg, EU.

Everitt, B.S. & Dunn, G. 1982, *An introduction to mathematical taxonomy,* Cambridge University Press, Cambridge, UK.

Fedderke, J. & Naumann, D. 2011, "An analysis of industry concentration in South African manufacturing, 1972-2001", *Applied Economics,* vol. 43, no. 22, pp. 2919-2939.

Fedderke, J. & Szalontai, G. 2009, "Industry concentration in South African manufacturing industry: trends and consequences, 1972-96", *Economic Modelling,* vol. 26, no. 1, pp. 241-250.

Fieller, E.C. 1932, "A numerical test of the adequacy of A.T. McKay's approximation", *Journal of Royal Statistical Society,* vol. 95, no. 4, pp. 699-702.

Forkman, J. 2009, "Estimator and tests for common coefficients of variation in normal distributions", *Communications in Statistics - Theory and Methods,* vol. 38, no. 2, pp. 233-251.

Forkman, J. & Verrill, S. 2008, "The distribution of McKay's approximation for the coefficient of variation", *Statistics and Probability Letters,* vol. 78, no. 1, pp. 10-14.

Fourie, F. & Smith, A. 2001, "Revisiting the concentration-profits relationship in South Africa: moving the debate beyond deadlock", *Journal of Studies in Economics and Econometrics,* vol. 25, no. 2, pp. 25-60.

Gaffeo, E. *et.al.* 2003, "On the size distribution of firms: additional evidence from the G7 countries", *Physica A,* vol. 324, no. 1-2, pp. 117-123.

Geary, R. 1947, "Testing for normality", *Biometrika,* vol. 34, no. 3/4, pp. 209-242.

George, F. & Kibria, B.G. 2012, "Confidence intervals for estimating the population signal-to-noise ratio: a simulation study", *Journal of Applied Statistics,* vol. 39, no. 6, pp. 1225-1240.

Gerber, L. 2007, "A quintile rule for the Gini coefficient", *Mathematics Magazine,* vol. 80, no. 2, pp. 133-135.

Gerstenkorn, T. & Gerstenkorn, J. 2003, "Gini's mean difference in the theory and application to inflated distributions", *Statistica,* vol. 63, no. 3, pp. 469-488.

Gibrat, R. 1957 [1931], "On economic inequalities" in *International Economic Papers*, eds. A.T. Peacock, W.F. *et.al.*, 7th edn, Macmillan and Company, London, UK, pp. 53-70.

Gini, C. 2005 [1914], "On the measurement of concentration and variability of characters", *Metron,* vol. 63, no. 1, pp. 3-38.

Gini, C. 1965, "On the characteristics of Italian statistics", *Journal of the Royal Statistical Society,* vol. 128, no. 1, pp. 89-109.

Gini, C. 1947, "Statistical relations and their inversions", *Review of International Statistical Institute,* vol. 15, no. 1/4, pp. 24-42.

Gini, C. 1936, *On the measure of concentration with special reference to income and wealth*, Cowles Commission for Research in Economics, Colorado, USA.

Gini, C. 1921, "Measurement of inequality of incomes", *Economic Journal,* vol. 31, no. 121, pp. 124-126.

Glasser, G. 1962a, "A statistical game and sample size", *Mathematics Teacher,* vol. 55, no. 8, pp. 626-629.

Glasser, G. 1962b, "Variance formulas for the mean difference and coefficient of concentration", *Journal of American Statistical Association,* vol. 57, no. 299, pp. 648-654.

Glasser, G. 1961a, "Relationships between the mean difference and other measures of variation", *Metron,* vol. 21, no. 1-4, pp. 176-180.

Glasser, G. 1961b, "Tchebycheff-type inequalities in terms of the mean deviation", *Sankhya,* vol. 23, no. 4, pp. 397-400.

Goldreich, O. & Wigderson, A. 2008, "Computational complexity" in *The Princeton companion to mathematics*, ed. T. Gowers, Princeton University Press, Princeton, USA.

Grubbs, F. 1973, "Errors of measurement, precision, accuracy and the statistical comparison of measuring instruments", *Technometrics,* vol. 15, no. 1, pp. 53-66.

Gulhar, M., *et al*. 2012, "A comparison of some confidence intervals for estimating the population coefficient of variation: a simulation study", *SORT,* vol. 36, no. 1, pp. 45-68.

Guterman, H.E. 1962, "An upper bound for the sample standard deviation", *Technometrics,* vol. 4, no. 1, pp. 134-135.

Hart, P.E. 1975, "Moment distributions in economics: an exposition", *Journal of Royal Statistical Society,* vol. 138, no. 3, pp. 423-434.

Hart, P.E. 1971, "Entropy and other measures of concentration", *Journal of Royal Statistical Society,* vol. 134, no. 1, pp. 73-85.

Hay, D.A. & Morris, D.J. 1991, *Industrial economics and organisation: theory and evidence,* 2nd edn, Oxford University Press, New York, USA.

Hayes, K. 2010, "On the attainability of bounds on the standard deviation", *Teaching Statistics,* vol. 32, no. 2, pp. 54-56.

Herfindahl, O.C. 1955, "Comment on Rosenbluth's measures of concentration" in *Business concentration and price policy*, ed. G. Stigler, Princeton University Press, Princeton, USA.

Hirschman, A.O. 1964, "The paternity of an index", *American Economic Review,* vol. 54, no. 5, pp. 761-762.

Hudson, R. 2011, "Obituary: Benoit Mandelbrot", *IMS Bulletin,* vol. 40, no. 4, pp. 8-9.

Hürlimann, W. 1998, "Coefficient of variation" in *Encyclopedia of statistical sciences*, eds. S. Kotz, C.B. Read & D.L. Banks, Volume 2, John Wiley & Sons, New York, USA, pp. 127-130.

Hürlimann, W. 1995, "A uniform approximation to the sampling distribution of the coefficient of variation", *Statistics and Probability Letters,* vol. 24, no. 3, pp. 263-268.

Hwang, T.Y. & Lin, Y.K. 2000, "On the distribution of the sample heterogeneity of molecular polymer", *Tamsui Oxford Journal of Mathematical Sciences,* vol. 16, no. 2, pp. 133-149.

Iglewicz, B. 1983, "Robust scale estimators and confidence intervals for location" in *Understanding robust and exploratory data analysis*, eds. D.C. Hoaglin *et.al.*, John Wiley and Sons, New York, USA.

Iglewicz, B., Myers, R. & Howe, R. 1968, "On the percentage points of the sample coefficient of variation", *Biometrika,* vol. 55, no. 3, pp. 580-581.

Iglewicz, B. & Myers, R.H. 1970, "Comparisons of approximations to the percentage points of the sample coefficient of variation", *Technometrics,* vol. 12, no. 1, pp. 166-169.

Jacobson, P.E. 1970, "Some comments to console Edgar F. Borgatta", *Sociological Quarterly,* vol. 11, no. 2, pp. 265-269.

Jansen, H. 1992, "Gini's coefficient of mean difference as a measure of adoption speed: theoretical issues and empirical evidence from India", *Agricultural Economics,* vol. 7, no. 3/4, pp. 351-369.

Kamat, A.R. 1961, "A note on Gini's mean difference", *Metron,* vol. 21, no. 1-4, pp. 170-175.

Kamat, A.R. 1953, "The third moment of Gini's mean difference", *Biometrika,* vol. 40, no. 3/4, pp. 451-452.

Kelly, W.A. 1981, "A generalized interpretation of the Herfindahl index", *Southern Economic Journal,* vol. 48, no. 1, pp. 50-57.

Kendall, M. & Stuart, A. 1977, *The advanced theory of statistics: distribution theory,* 4th edn, Charles Griffin, London, UK.

Krishnamoorthy, K. 2006, *Handbook of statistical distributions with applications,* Chapman and Hall, Boka Raton, USA.

Lange, L.H. 1963, "Some inequality problems", *Mathematics Teacher,* vol. 56, no. 7, pp. 490-494.

Lange, L.H. 1959, "On two famous inequalities", *Mathematics Magazine,* vol. 32, no. 3, pp. 157-160.

Lawrence, R.J. 1988, "Applications in economics and business" in *Lognormal distributions: theory and applications*, eds. E.L. Crow & K. Shimizu, Marcel Dekker, New York, USA, pp. 229-266.

Leach, D.F. 1997, "The concentration-profit, monopoly vs. efficiency debate: some new South African evidence", *Contemporary Economic Policy,* vol. 15, no. 2, pp. 12-23.

Lerman, R.I. & Yitzhaki, S. 1984, "A note on the calculation and interpretation of the Gini index", *Economics Letters,* vol. 15, no. 3/4, pp. 363-368.

Lorenz, M. 1905, "Methods of measuring the concentration of wealth", *Journal of American Statistical Association,* vol. 9, no. 70, pp. 209-219.

Mandelbrot, B.B. 1997, *Fractals and scaling in finance: discontinuity, concentration, risk,* Springer Publishing, New York, USA.

Marfels, C. 1972, "The Gini ratio of concentration reconsidered", *Statistical Papers,* vol. 13, no. 2, pp. 160-179.

Marfels, C. 1971, "Absolute and relative measures of concentration reconsidered", *Kyklos,* vol. 24, no. 4, pp. 753-766.

Marron, J.S. 1999, "Effective writing in mathematical statistics", *Statistica Neerlandica,* vol. 53, no. 1, pp. 68-75.

Maturi, T.A. & Elsayigh, A. 2009, "The correlation between variate values and ranks in samples from complete fourth power exponential distribution", *Journal of Mathematics Research,* vol. 1, no. 1, pp. 14-18.

McDonald, J. 1981, "Some issues associated with the measurement of income inequality" in *Statistical distributions in scientific work*, eds. C. Taillie, G. Patil & B. Baldessari, D. Reidel Publishing, Dordrecht, The Netherlands.

McKay, A.T. 1932, "Distribution of the coefficient of variation and the extended "t" distribution", *Journal of the Royal Statistical Society,* vol. 95, no. 4, pp. 695-698.

Milanovic, B. 1997, "A simple way to calculate the Gini coefficient, and some implications", *Economics Letters,* vol. 56, no. 1, pp. 45-49.

Mooney, C.Z. & Duval, R.D. 1993, *Bootstrapping: a nonparametric approach to statistical inference,* Sage Publications, Inc., Newbury Park, USA.

Morgenthaler, S. 2007, "A survey of robust statistics", *Statistical Methods and Applications,* vol. 15, no. 3, pp. 271-293.

O.E.C.D. 2008, *Glossary of statistical terms,* Organization for Economic Co-operation and Development, Paris, France.

O.E.C.D. 2006, *Structural and demographic business statistics, 2006*, Organization for Economic Co-operation and Development, Paris, France.

O'Brien, R.M. 1982, "Using rank-order measures to represent continuous variables", *Social Forces,* vol. 61, no. 1, pp. 144-155.

Ogwang, T. 2000, "A convenient method of computing the Gini index and its standard error", *Oxford Bulletin of Economics and Statistics,* vol. 62, no. 1, pp. 123-129.

Olkin, I. & Yitzhaki, S. 1992, "Gini regression analysis", *International Statistical Review,* vol. 60, no. 2, pp. 185-196.

Olkin, I. & Yitzhaki, S. 1991, *Concentration indices and concentration curves*, Institute of Mathetmatical Statistics, Hayward, USA.

Ott, R.L. 1993, *An introduction to statistical methods and data analysis,* 4th edn, Duxbury Press, Belmont, USA.

Oxford Concise Dictionary of Mathematics 2009, *Asymptotic functions*, Oxford University Press, New York, USA.

Oxford Dictionary of Statistical Terms 2003, *Concentration coefficient*, Oxford University Press, New York, USA.

Pearson, E.S. 1932, "Comparison of A.T. McKay's approximation with experimental sampling results", *Journal of Royal Statistical Society,* vol. 95, no. 4, pp. 703-704.

Phadke, M. 1992, "The role of interactions, SN ratios, and selection of quality characteristics", *Taguchi's parameter design: a panel discussion*, ed. V. Nair, Technometrics, pp. 137.

Piesch, W. 2005, "A look at the structure of some extended Ginis", *Metron,* vol. 63, no. 2, pp. 263-296.

Ramasubban, T.A. 1959, "The generalised mean differences of the binomial and Poisson distributions", *Biometrika,* vol. 46, no. 1/2, pp. 223-229.

Ramasubban, T.A. 1956, "A Chi-square approximation to Gini's mean difference", *Journal of the Indian Society of Agricultural Statistics,* vol. 8, no. 1/2, pp. 116-122.

Reekie, W.D. 1989, *Industrial economics: a critical introduction to corporate enterprise in Europe and America,* Edward Elgar Publishing, Aldershot, UK.

Reh, W. & Scheffler, B. 1996, "Significance tests and confidence intervals for coefficients of variation", *Computational Statistics and Data Analysis,* vol. 22, no. 4, pp. 449-452.

Rhoades, S.A. 1993, "The Herfindahl-Hirschman index", *Federal Reserve Bulletin,* vol. 79, no. 3, pp. 188-189.

Rosenbluth, G. 1955, "Measures of concentration" in *Business concentration and price policy*, ed. G. Stigler, Princeton University Press, Princeton, USA.

Salop, S.C. & O'Brien, D.P. 2000, "Competitive effects of partial ownership: financial interest and corporate control", *Antitrust Law Journal,* vol. 67, no. 3, pp. 559-614.

Sawilowsky, S. 2006, "Monte Carlo methods" in *Encyclopaedia of measurement and statistics*, ed. N.J. Salkind, Sage Publishing, Thousand Oaks, USA.

Schaaf, W. 1951, "The use of simple algebra in business arithmetic", *Mathematics Teacher,* vol. 44, no. 1, pp. 22-25.

Shalit, H. 1985, "Calculating the Gini index of inequality for individual data", *Oxford Bulletin of Economics and Statistics,* vol. 47, no. 2, pp. 185-189.

Sharma, J.K. 2010, *Fundamentals of business statistics,* Pearson Education, New Delhi, India.

Smith, A. & Du Plessis, S. 1996, "Concentration in South African manufacturing industry: measuring both blades of the Marshallian Scissors", *Journal of Studies in Economics and Econometrics,* vol. 20, no. 2, pp. 1-24.

Smithson, M. 2000, *Statistics with confidence,* 1st edn, Sage Publishing, London, UK.

Stats SA. 2008, *Compendium of industrial statistics, 2008*, Statistics South Africa, Pretoria, SA.

Stats SA. 1999, *Census of manufacturing, 1996*, Statistics South Africa, Pretoria, SA.

Stigler, S.M. 2010, "The changing history of robustness", *American Statistician,* vol. 64, no. 4, pp. 277-281.

Stuart, A. 1954, "The correlation between variate-values and ranks in samples from a continuous distribution", *British Journal of Statistical Psychology,* vol. 7, no. 1, pp. 37-44.

Taguchi, G. & Clausing, D. 1990a, "Reply to George Box", *Harvard Business Review,* vol. 68, no. 2, pp. 228-229.

Taguchi, G. & Clausing, D. 1990b, "Robust quality", *Harvard Business Review,* vol. 68, no. 1, pp. 65-75.

The Economist. 1998, *The trustbusters' new tools*.

The New Palgrave Dictionary of Economics and The Law 2002, *Horizontal mergers*, Palgrave Macmillan, New York, USA.

Theron, N. 2001, "The economics of competition policy: merger analysis in South Africa", *South African Journal of Economics,* vol. 69, no. 4, pp. 614-658.

Thomson, G.W. 1955, "Bounds for the ratio of range to standard deviation", *Biometrika,* vol. 42, no. 1/2, pp. 268-269.

Tomkins, C. 2006, "An introduction to non-parametric statistics for health scientists", *University of Alberta Health Sciences Journal,* vol. 3, no. 1, pp. 20-26.

Touhey, P. 1995, "Chebyshev's theorem: a geometric approach", *College Mathematics Journal,* vol. 26, no. 2, pp. 139-141.

Tukey, J.W. 1986, "Sunset salvo", *American Statistician,* vol. 40, no. 1, pp. 72-76.

Tukey, J.W. 1980, "We need both exploratory and confirmatory", *American Statistician,* vol. 34, no. 1, pp. 23-25.

U.S. Census Bureau. 2006, *Concentration ratios: 2002*, United States Census Bureau, Washington DC, USA.

U.S. Department of Justice & Federal Trade Commission. 2010, *Horizontal merger guidelines*, U.S. Department of Justice and the Federal Trade Commission, Washington DC, USA.

U.S. Department of Justice & Federal Trade Commission. 1997 [1992], *Horizontal merger guidelines*, U.S. Department of Justice and the Federal Trade Commission, Washington DC, USA.

Umphrey, G.J. 1983, "A comment on McKay's approximation for the coefficient of variation", *Communications in Statistics: Simulation and Computation,* vol. 12, no. 5, pp. 629-635.

Van Belle, G. 2008, *Statistical rules of thumb,* 2nd edn, John Wiley and Sons, Hoboken, USA.

Vangel, M.G. 1996, "Confidence intervals for a normal coefficient of variation", *American Statistician,* vol. 50, no. 1, pp. 21-26.

Walther, B.A. & Moore, J.L. 2005, "The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance", *Ecography,* vol. 28, no. 6, pp. 815-829.

Warren, W.G. 1982, "On the adequacy of the Chi-squared approximation for the coefficient of variation", *Communications in Statistics: Simulation and Computation,* vol. 11, no. 6, pp. 659-666.

Wild, C.J. & Pfannkuch, M. 2000, "Statistical thinking and statistical practice: themes gleaned from professional statisticians", *Statistical Science,* vol. 15, no. 2, pp. 132-152.

Wild, C.J. & Pfannkuch, M. 1999, "Statistical thinking in empirical enquiry", *International Statistical Review,* vol. 67, no. 3, pp. 223-248.

Wong, A. & Wu, J. 2002, "Small sample asymptotic inference for the coefficient of variation: normal and non-normal models", *Journal of Statistical Planning and Inference,* vol. 104, no. 1, pp. 73-82.

Wu, J. 1992, "The role of interactions, SN ratios, and selection of quality characteristics", *Taguchi's parameter design: a panel discussion*, ed. V. Nair, Technometrics, pp. 139.

Yitzhaki, S. 1998, "More than a dozen alternative ways of spelling Gini" in *Research on economic inequality*, ed. D.J. Slottje, JAI Press, Stamford, USA.

Yung, Y.F. & Chan, W. 1999, "Statistical analyses using bootstrapping: concepts and implementation" in *Statistical strategies for small sample research*, ed. R.H. Hoyle, Sage Publications, Thousand Oaks, USA.

Zickar, M. 2006, "Computational modelling" in *Psychology research handbook*, eds. F.T. Leong & J.T. Austin, 2nd edn, Sage Publishing, Thousand Oaks, USA.