Spring 2006

Measuring inequality: Statistical inference theory with applications

Mihaela Paun
Louisiana Tech University

MEASURING INEQUALITY: STATISTICAL INFERENCE THEORY WITH APPLICATIONS

by

Mihaela M. Paun, B.S., M.S.

A Dissertation Presented in Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

COLLEGE OF ENGINEERING AND SCIENCE

LOUISIANA TECH UNIVERSITY

May 2006

LOUISIANA TECH UNIVERSITY

THE GRADUATE SCHOOL

March 23, 2006
Date

We hereby recommend that the dissertation prepared under our supervision

by Mihaela M. Paun

entitled MEASURING INEQUALITY: STATISTICAL INFERENCE THEORY WITH APPLICATIONS

be accepted in partial fulfillment of the requirements for the Degree of

Doctor of Philosophy

Raja Nassar
Supervisor of Dissertation Research

R. Zitikis
Co-supervisor of Dissertation Research

Richard J. Greechie
Head of Department

CAM
Department

Recommendation concurred in:

Weizhong Dai

Richard J. Greechie

Advisory Committee

Approved:

Bala Ramachandran
Director of Graduate Studies

Stan Napper
Dean of the College

Approved:

Terry M. McConathy
Dean of the Graduate School

GS Form 13
(5/03)

ABSTRACT

In this dissertation we develop statistical inference for the Atkinson index, one of the measures of inequality used in studying economic inequality.

Specifically, we construct empirical estimators for the Atkinson index, both in the parametric and nonparametric case, and derive formulas for the asymptotic variances for the estimators. These statistics are used for testing hypothesis and constructing confidence intervals for the Atkinson index. We test the validity and the robustness of the asymptotic theory, by simulations (using R, a language and environment for statistical computing and graphics), in the case of one and two populations. In addition to proving asymptotic normality for the theory, we develop a nonparametric bootstrap theory, as an alternative to the asymptotic theory, and present some of the advantages for this method.

It is natural, when studying income inequality, to analyze the distributions of the data sets and make statistical inference about various parameters of interest, such as means, medians, variances, etc. In trying to condense the information into a few parameters, one certainly faces a problem of constructing measures or indices that would give a proper idea about what happens in the society under consideration. The mean, as a statistical measures of distribution is useful in some instances, but not particularly relevant when, for example, we have outliers and/or skewed distributions. In addition, the mean does not tell us if inequality changes by transfers of wealth from the rich to the poor or from the poor to the rich. Hence, the need for constructing measures with various properties like transfer sensitivity, scale and/or location invariance.

Indeed, some measures, like the Gini index, are not sensitive to the transfers at the lower and upper ends of the distribution, whereas other measures like the Atkinson index are more sensitive to such transfers. In this dissertation we have chosen to work with the Atkinson index because of its econometric properties described in Chapter 1 and the lack of inferential statistics results in the literature pertaining to this index.

Keywords: Economic inequality, Gini index, Atkinson index, consistency, asymptotic normality, bootstrap, confidence intervals.

APPROVAL FOR SCHOLARLY DISSEMINATION

The author grants to the Prescott Memorial Library of Louisiana Tech University the right to reproduce, by appropriate methods, upon request, any or all portions of this Dissertation. It is understood that "proper request" consists of the agreement, on the part of the requesting party, that said reproduction is for his personal use and that subsequent reproduction will not occur without written approval of the author of this Dissertation. Further, any portions of the Dissertation used in books, papers, and other works must be appropriately referenced to this Dissertation.

Finally, the author of this Dissertation reserves the right to publish freely, in the literature, at any time, any or all portions of this Dissertation.

Author ______________________

Date April 20, 2006

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

The successful completion of this dissertation is due in great part to the support, encouragement and understanding of my two supervisors Dr. Raja Nassar and Dr. Ricardas Zitikis. They have been a continuous inspiration for me over the years.

My special thanks are also due to the my advisory committee: Dr. Richard Greechie, Dr. Weizhong Dai and Dr. Chokchai Leangsuksun.

Finally, I would like to acknowledge with great appreciation the support I have received from the College of Engineering and Science at the Louisiana Tech University.

DEDICATION

To my husband for his support and understanding in accomplishing my goals and to my parents for always believing in me.

CHAPTER 1

OVERVIEW OF THE LITERATURE

1.1 Introduction

The Atkinson index, was introduced by Sir Anthony Atkinson in 1970 to study income distribution, see [Atkinson, 1970]. However, its use was hampered by the lack of any methodology for estimation and hypothesis testing. In the present dissertation we fill this gap by developing statistical methodology for making inference concerning this index.

One of the most attractive feature the Atkinson index has is that it allows one to specify the social welfare function which characterizes the income distribution. By specifying the welfare function for the Atkinson index, the researcher may choose to emphasize the lower, middle or upper end of the income distribution. The Atkinson index parameter a, with $(0 < a < 1)$, determines the level of income inequality in the population. The lower the value of a the more society is concerned about inequality.

In this dissertation, inferential statistics (namely, test statistics and confidence intervals) were developed for the Atkinson index in the case of large samples (asymptotic results). This was done in the parametric case where the income distribution was specified. Also, the same inferential statistics were developed in the nonparametric or distribution free case using the bootstrap method.

In the parametric case, three distributions have been considered: Pareto, Exponential and Lognormal distribution and for each parametric family we have derived the estimators for the Atkinson index, as well as for the variances. These expressions are needed when testing hypothesis and setting up confidence limits.

The bootstrap approach is a resampling procedure that employs simulation methods to evaluate a parameter estimate. This approach requires no theoretical calculations, applies the same to any inequality measure that we use, and it is available no matter how mathematically complicated the parameter estimate or its asymptotic standard error may be. Monte Carlo simulation was performed to check on the adequacy and robustness of the asymptotic results for relatively small sample size. The simulation was done using R, a language and environment for statistical computing and graphics, and the code for the programs is provided in Appendix B. Results from simulation (parametric and nonparametric) showed that the asymptotic theory was adequate for sample size as small as 100. The parametric results were more efficient than the nonparametric results.

The dissertation is structured as follows:

In Chapter 1 we give the main definitions and notations used in subsequent chapters and we define a number of known measures (indices) of income inequality. We discuss the properties of these indices from the economics point of view, as well as the statistics point of view. We introduce the Atkinson index and discuss special properties that this index possesses which make it a better index to use than the currently used Gini index.

In Chapter 2 we investigate the theoretical Atkinson index, introduce its empirical estimator, and prove its consistency.

In Chapter 3 we present a large sample statistical inference theory for the Atkinson index in the case of one population (parametric and nonparametric). Bootstrap approximations are also presented. The bootstrap approximations are of particular interest when it is not easy to derive and/or estimate the true asymptotic variance of the index estimate.

Chapter 4 presents a simulation study for the theory presented in Chapter 3. The chosen distributions for modelling income are the Pareto, Exponential, and Lognormal distribution.

In Chapter 5 we extend the theory presented in Chapter 3 to two populations. We consider the cases of independent and paired samples. The theory for independent samples is applicable when one compares Atkinson indices for the incomes of two populations A and B. The case of paired samples is applicable when comparing Atkinson indices for the same population but over two time periods. For example, one might be interested in comparing the Atkinson indices for university A salaries in 1990 and 2005.

Chapter 6 presents a simulation study for the theory presented in Chapter 5.

Chapter 7 concludes the dissertation and presents future problems of interest, such as comparing Atkinson indices for more than two populations.

1.2 Income Inequality Indices - Economics Point of View

When one considers income inequality and indices of income inequality, several questions come to mind: “What is inequality?” “How can one measure inequality?”

"Is there one measure of inequality that is better than other measures?" These questions will be answered in this chapter.

When considering two societies, one with perfect equality and the other with an unequal distribution of incomes, it is easy to distinguish which one has the bigger inequality. However, when both societies are unequal, it is hard to decide which one has a more unequal distribution.

There are several known measures of inequality and researchers have used them based on convenience and/or familiarity.

For the rest of this section we suppose that we have a population of n individuals where each individual receives an annual income x_i, with $i = 1, 2, \dots n$. We assume that the incomes are arranged in ascending order. Therefore, we have $x_1 \leq x_2 \leq \cdots \leq x_n$. We need one measure of inequality that can characterize every possible set x_i with regard to inequality.

The measures that we are looking for should have some properties. For instance, they should be zero when all individuals have the same income and positive when there are at least two different incomes.

At this point, all common measures of inequality, including the dispersion measures (variance, standard deviation) satisfy these conditions. The measures that are considered for measuring income inequality are the following:

- Variance

- Standard Deviation

- Coefficient of Variation

- The Gini Index

- The Theil Index

- The Atkinson Index

Measures that are scale invariant but not location invariant are desirable. As will be seen, not all these measures satisfy these properties. We remind the reader, that the income inequality measures are techniques used by economists to measure the distribution of income among members of a society. In particular these techniques are used to measure the inequality, or equality of income within an economy.

Definition 1.2.1 *An inequality measure* I *is* ***location invariant*** *if for any scalar* $a > 0$ *we have* $I(x) = I(x+a)$.

The measures of inequality discussed above are not location invariant.

To explain this property, we consider four individuals with incomes of $5,000, $20,000, $50,000 and $80,000. We can see a big difference in these incomes. However, if we add $500,000 to each one of the four incomes, the individuals will have the following incomes: $505,000, $520,000, $550,000 and $580,000. It is easy to see that the difference in incomes is now almost negligible and we can conclude that the inequality declined. After this argument, one can eliminate the standard deviation because it will stay exactly the same, and we have made the argument that it should decrease.

One quantity, known as the **Mean Gini difference**, is a location invariant statistic. This quantity is given as

$$GMD = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|.$$

The second property that the measures need to satisfy is the scale invariance principle. The *Principle of scale invariance* requires the inequality measure to be invariant to proportional changes. If each individual's income changes by the same proportion (as it happens, for example, when changing the currency unit), then the inequality should not change.

Formally, this principle is stated as follows:

Definition 1.2.2 *An inequality metric I is* ***scale invariant*** *if for any scalar $\lambda > 0$ we have $I(x) = I(\lambda x)$.*

We verify this principle for the following measures:

- Variance

- Coefficient of Variation

- The Gini Index

- The Theil Index

- The Atkinson Index

This rule eliminates the variance, because if a person's income is doubled, then the variance becomes now four times greater than before. Formally, $Var(\lambda x) = \lambda^2 Var(x)$, for any scalar λ. Therefore the variance does not satisfy the scale invariance principle.

The measures that satisfy this criterion also satisfy the rule that if we change the unit in which the incomes are measured (for example we change U.S. dollars to Euros), then the distribution of the incomes does not change. We say that the measure is invariant to such changes.

To convert a measure of dispersion into a scale invariant measure of inequality, we have to divide it by the mean or a function of the mean. Therefore, we will consider the measures defined below.

The coefficient of variation V is defined as the standard deviation divided by the mean:

$$V_n = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}}{\overline{X}}.$$

One can verify that this measure is scale invariant using the rule in Def. (1.2.2) for the coefficient of variation

$$V_n(\lambda x) = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\lambda x_i - \lambda\overline{x})^2}}{\lambda\overline{X}} = V_n(x).$$

The most common used measure of inequality, **the Gini index** G, is defined in terms of the Lorenz curve. This is a measure of dispersion divided by twice the mean. The empirical Gini index is defined below,

$$G_n = \frac{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2\overline{X}}.$$

To calculate the Gini index, we take the average of all absolute differences between all pairs of incomes and divide it by twice the mean of the incomes.

We can verify using Def. (1.2.2) that the Gini index is satisfying the scale invariance principle, because $G_n(\lambda x) = G_n(x)$.

$$\begin{aligned} G_n(\lambda x) &= \frac{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2\overline{X}} = \frac{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|\lambda x_i - \lambda x_j|}{2\lambda\overline{X}} \\ &= \frac{\frac{\lambda}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2\lambda\overline{X}} = G_n(x) \end{aligned}$$

One other measure used in the literature is the **Theil index,** T, defined as follows:

$$T_n = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i log x_i - \overline{X} log \overline{X}}{\overline{X}}.$$

We can verify using Def. (1.2.2) that the Theil index is satisfying the scale invariance principle, because $T_n(\lambda x) = T_n(x)$.

$$\begin{aligned} T_n(\lambda x) &= \frac{\frac{1}{n}\sum_{i=1}^{n} \lambda x_i log \lambda x_i - \lambda\overline{X} log \lambda\overline{X}}{\lambda\overline{X}} \\ &= \frac{\frac{\lambda}{n}\sum_{i=1}^{n} x_i log x_i + \frac{\lambda}{n}\sum_{i=1}^{n} x_i log\lambda - \lambda\overline{X} log\overline{X} - \lambda\overline{X} log\lambda}{\lambda\overline{X}} \\ &= \frac{\frac{1}{n}\sum_{i=1}^{n} x_i log x_i + \frac{1}{n}\sum_{i=1}^{n} x_i log\lambda - \overline{X} log\overline{X} - \overline{X} log\lambda}{\mu} \\ &= \frac{\frac{1}{n}\sum_{i=1}^{n} x_i log x_i + \overline{X} log\lambda - \overline{X} log\overline{X} - \overline{X} log\lambda}{\overline{X}} = T_n(x). \end{aligned}$$

The **Atkinson index,** A is defined as follows:

$$A_n = 1 - \frac{1}{\overline{x}}\left(\frac{1}{n}\sum_{i=1}^{n} x_i^{1-a}\right)^{\frac{1}{1-a}}.$$

One can verify using Def. (1.2.2) that the Atkinson index is satisfying the scale invariance principle, because $A_n(\lambda x) = A_n(x)$.

$$
\begin{aligned}
A_n(\lambda x) &= 1-\frac{1}{\lambda\overline{x}}\left(\frac{1}{n}\sum_{i=1}^{n}(\lambda x_i)^{1-a}\right)^{\frac{1}{1-a}}\\
&= 1-\frac{1}{\lambda\overline{x}}\left(\frac{\lambda^{1-a}}{n}\sum_{i=1}^{n}x_i^{1-a}\right)^{\frac{1}{1-a}}\\
&= 1-\frac{\lambda}{\lambda\overline{x}}\left(\frac{1}{n}\sum_{i=1}^{n}x_i^{1-a}\right)^{\frac{1}{1-a}}\\
&= 1-\frac{1}{\overline{x}}\left(\frac{1}{n}\sum_{i=1}^{n}x_i^{1-a}\right)^{\frac{1}{1-a}} = A_n(x).
\end{aligned}
$$

So which one of these four measures would one choose when studying the income inequality? To answer this question we will apply what the literature calls **the transfer principle**. [Dalton, 1920] made the argument that if we transfer income from a poorer individual to a richer individual (regardless of how poor or how rich the individuals are, or of the amount that we are transferring), then the measures of inequality should increase in this case. This argument is intuitively correct, since one individual becomes poorer and the other richer.

Definition 1.2.3 *We consider the vector x' which is the transformation of the vector x obtained by a transfer of a positive amount 'a' from x_i to x_j, where $x_j > x_i$ and $x_j + a > x_i - a$. Then **the transfer principle** is satisfied if and only if $I(x') \geq I(x)$. This means that because we transfer from a poor individual to a richer individual, the income inequality measure I rises. Alternatively, if we transfer an amount 'a' from a richer individual to a pooper individual, the inequality measure falls.*

We verify this principle for the following measures:

- Coefficient of Variation

- The Gini Index

- The Theil Index

- The Atkinson Index

Subsequent studies indicate that this principle is used when one discusses the Lorenz curve or the inequality measures that use social welfare functions (like the Atkinson index).

All the measures defined above satisfy the principle of transfer. However, some differences exists in the sensitivity of transfers at different points on the scale, according to [Atkinson, 1970]. Consider two individuals from a given society with n individuals, with incomes x_i and x_j, with $x_i \leq x_j$. We wish to transfer an amount $'a'$ only from the i^{th} individual to the j^{th} individual in the population.

The coefficients of variations are V_1 before the transfer and V_2 after the transfer. From [Dalton, 1920] one can see that

$$V_2^2 - V_1^2 = ca(x_j - x_i) + ca^2,$$

where c is positive and depends only on the mean and number of observations.

We can interpret this result as follows: The coefficient of variation V is equally sensitive to transfers at all income levels. Therefore, if one transfers \$100 from a person who has an income of \$10,000 to a person who has an income of \$11,000 one will have the same impact as if one transfers the \$100 from a person who has an income of \$60,000 to a person who has an income of \$61,000.

Note 1.2.4 *The* ***coefficient of variation*** *attaches equal weight to transfers anywhere in the distribution.*

It is of interest to determine how the Gini index behaves when one transfers income.

The Gini coefficients are G_1 before the transfer and G_2 after the transfer. To check the transfer sensitivity, we use an equivalent formula of the Gini index, given by [Dasgupta et al., 1973]

$$G = \frac{2}{\overline{X}n^2}\sum_{i=1}^{n} ix_i - \frac{n+1}{n}.$$

Therefore, one can show that

$$G_2 - G_1 = c'a(j-i),$$

where $c' = \frac{2}{\mu n^2}$ is positive and depends only on the mean and number of observations.

One immediate observation is that the sensitivity of transfer depends not on the value of the incomes, but on the ranks of the individuals (i is the rank of individual with income x_i, and j is the rank of individual with income x_j).

In other words, the change in the Gini index depends on the number of individuals with incomes lower than x_j and higher than x_i. Today, the U.S. has more individuals with incomes in the \$30,000 - \$35,000 interval than in the \$90,000 - \$95,000 interval. Therefore, a transfer from an individual earning \$30,000 to another earning \$35,000 will affect the Gini coefficient more than if we transfer an equal amount from an individual earning \$90,000 to another earning \$95,000. We also have less individuals with incomes in the interval \$5,000 - \$10,000.

Thus we can conclude that the Gini index is more sensitive to transfers around the middle of the distribution (middle class) and less sensitive to transfers among very poor or very rich individuals.

Note 1.2.5 *The* ***Gini index*** *attaches more weight to transfers at the middle of the distribution than in the tails.*

For the Theil index we have, using the above definition, that for T_1 before the transfer and T_2 after the transfer:

$$T_2 - T_1 = \frac{1}{n\overline{X}} \left[x_j log \frac{x_j + a}{x_j} + a \cdot log \frac{x_j + a}{x_i - a} + x_i log \frac{x_i - a}{x_i} \right].$$

From the above expression, the effect of a transfer on the Theil's index is not very clear. Since a can be any amount, for any value of a the effect should be the same. We consider a to be very small and we use a limiting argument. We now have that the change in the Theil's index resulting from a transfer from x_i to x_j is of the following form:

$$T_2 - T_1 = \frac{a}{n\overline{X}} log \frac{x_j}{x_i}.$$

In other words for the Theil index the change in incomes depends on the ratio of the incomes. Therefore, one can conclude that transferring \$500 from an individual that earns \$20,000 to an individual that earns \$30,000 has almost the same effect on the Theil index as a transfer of still \$500 from an individual that earns \$60,000 to an individual that earns \$90,000. However, this change in the Theil index is approximatively 24 times larger than the change in the Theil index that happens when we transfer \$500 from an individual earning \$60,000 to an individual earning \$61,000. We can conclude that the lower the level of income, the more sensitive the Theil index is to transfers.

The sensitivity of Atkinson index depends on how we choose the welfare function. We may choose to emphasize the lower, middle or upper end of the distribution. In the next section we will discuss at large the properties that the Atkinson index has.

After discussing each index and highlighting their behavior, we were left with the of whether we now had a criterion for choosing among the indices.

The answer is that it depends on the assumption that we make when we start analyzing our data. Let us also remember that the Gini index depends on the shape of the frequency distribution. Since most distributions are bell-shaped, the Gini index will be most sensitive in the middle of the distribution. The index of choice is Gini if changes among middle class individuals are of interest to the researcher.

In his well-known paper, [Atkinson, 1970] compared income inequality among 12 different countries. He concluded that measures which were more sensitive in the lower range of the incomes show relatively less inequality in developing countries and more inequality in developed countries. Atkinson's explanation was due to the fact that developing nations have a large number of poor individuals and also a great inequality among the rich individuals.

For an hypothetical case when one individual has everything and the rest of the population has nothing, the coefficient of variation goes to ∞, the Theil index goes to ∞ and the Gini index is 1.

One can use these indices for a number of situations such as measuring the inequality of incomes, the difference in age for a given population, the inequality in students grades of a school. One can also measure the inequality for two distributions.

As we have mentioned above, the Gini index is usually defined in terms of the Lorenz curve. Actually, all invariant-scale measures of inequality can relate to the Lorenz curve, which allows us to formulate a rule for greater or lesser inequality.

In what follows we try to explain how to obtain the Lorenz curve. We have, as before, a population of n individuals and we rank their incomes from lowest to highest. For each rank we calculate the proportion of the population at that rank or below the rank, and we construct a vector $'a'$ with these values. We also calculate the proportion of incomes at that rank or below the rank and we construct a vector $'b'$ with these values. If one plots all the pairs (a, b) for every rank, one obtains the Lorenz curve.

The Gini index is equal to twice the area between the Lorenz curve and the egalitarian line (the line of perfect equality, plotted as the diagonal). Consider the following problem: Given a fixed total income that one wishes to distribute among n individuals, we assume that W is a number for every possible distribution of incomes that will indicate the preference for that distribution. This W is called the social welfare function.

In [Atkinson, 1970] this approach was adopted and it was proved that if one places some constraints on W one will obtain an important relationship between W and the Lorenz curve and as well as between W and the principle of transfers. If we define the utility function of an income x_i to be $U(x_i)$ then we can say that the welfare function W is

$$W = \sum_{i=1}^{n} U(x_i).$$

Atkinson also assumed that the utility function is the same for all individuals and it is a concave and increasing function. The important idea Atkinson presented in his paper was that a ranking of the Lorenz curves implies a ranking of social welfare.

For example, if we have two distributions X and Y and the Lorenz curve for X is somewhere above the Lorenz curve for Y but never below, then the welfare function $W(X)$ is greater than the welfare function $W(Y)$. We can say that the distribution X is preferable to the distribution Y.

Atkinson has chosen an additive separable, symmetric and concave welfare function and defined the following measure of inequality:

$$A = 1 - \frac{1}{\overline{x}} \left(\frac{1}{n} \sum_{i=1}^{n} x_i^{1-a} \right)^{\frac{1}{1-a}},$$

where $a > 0$. The parameter a reflects the strength of society's preference for equality and can take values ranging from zero to infinity. As $'a'$ increases, $'A'$ becomes more sensitive to transfers at the lower end of the distribution (lower incomes) and less sensitive to transfers among larger incomes. What is important about the Atkinson index is that we can choose a to follow one's decision about what portions of the distribution are more relevant to the analysis.

In conclusion, when measuring inequality one can choose one of the three measures of inequality discussed: Gini, Theil or the coefficient of variation. However, if one wants to incorporate judgements about social welfare, then one should choose the Atkinson index.

In [Atkinson, 1970] it was argued that the Atkinson index should be used in place of the conventional indices (Gini, Theil or the coefficient of variation). He presented some of the strengths of the Atkinson index. One cannot reach a complete ranking of distributions without specifying the form of the social welfare function. If we examine the implicit welfare functions of these measures, we see that in a number of cases they have properties that sometimes are not in accord with the social values.

1.3 Income Inequality Indices - Statistics Point of View

Inequality indices embody explicitly or implicitly social values on income distributions. So it is common for those who are concerned with comparing distributional changes over time to draw conclusions from comparisons of estimates of mean income and inequality.

One of the problems is to compare two frequency distributions F_1 and F_2. The conventional approach in nearly all empirical work is to adopt some summary statistic of inequality such as the variance, the coefficient of variation or the Gini coefficient.

We define below some of the inequality measures used by the economists, starting with the conventional statistic formulas.

Definition 1.3.1 *The* **variance** *of a random variable X is denoted by $Var(X)$ and is defined by*

$$Var(X) = \mathbf{E}[(X - \mathbf{E}[X])^2]. \tag{1.1}$$

Definition 1.3.2 *A measure of the relative variability, that is, variability relative to the magnitude of data, is the* **coefficient of variation, COV** *and it is the ratio of the sample standard deviation to the mean.*

$$COV = \frac{\sigma}{\mu}. \tag{1.2}$$

In 1905, Max Otto Lorenz suggested a method based on a convex function that has been widely used in comparing distributions of incomes. The Lorenz curve was developed as a graphical representation of an income distribution. It portrays observed income distributions and compares the results to a state of perfect income equality.

Suppose one is interested in the distribution of income in a society. One randomly selects n individuals from the society and records their incomes: $X_1, X_2, ...X_n$. The incomes can be negative (if the individuals selected are in debt) or they can be non-negative incomes.

One wishes to determine the inequality among the n incomes. We assume without the loss of generality that all incomes are positive. For our discussion, we would like these incomes to be ordered in ascending order: $X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$. If all incomes are equal, there is no inequality among the incomes. If one assumes all incomes are equal with a constant c, then $X_{1:n} = X_{2:n} \cdots = X_{n:n} = c$. Using the Lorenz curve, one can demonstrate graphically the distribution of income in a population. If each person in a country earns c per year, this means that the income is evenly distributed.

As shown in Fig. (1.1), the plot is a 45 degree line. On the $y - axis$ one has the percent income. On the $x - axis$ we have the population also expressed as a percentage. Obviously, the upper limits of each scale must equal 100. One hundred percent of the population earns 100 percent of the income. Similarly, zero percent of the population earn zero percent of the income. If income is distributed so that everyone earns the same, then 20% of the population earns 20% of the income, 40% of the households earn 40% of the income, and so on. This is called the **line of absolute equality** (egalitarian line).

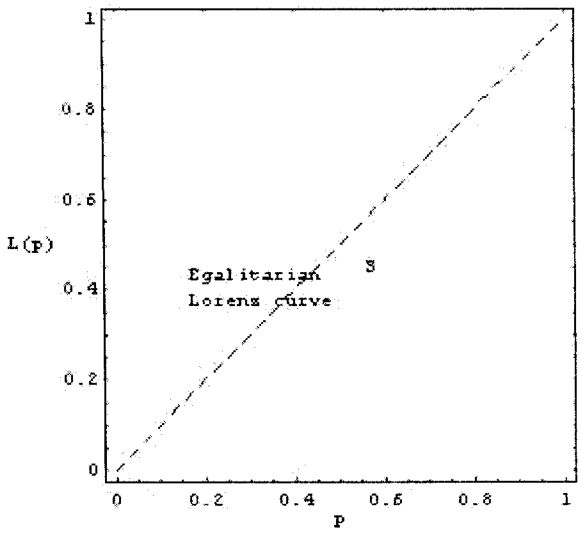

Figure 1.1 Lorenz curve - equal incomes

If the incomes are not equal, given the ordered values previously obtained, we calculate, for any $k = 0, 1, \ldots, n$, the proportion of income that the least fortunate $(k/n) \times 100\%$ individuals possess.

Mathematically, we can represent these proportions as follows:

$$l_{k,n} := \left(\sum_{i=1}^{k} X_{i:n}\right) / \left(\sum_{i=1}^{n} X_{i:n}\right), k = 0, 1, \ldots, n.$$

We assume that the denominator in the above expression is different from zero. We note that when $k = 0$ the sum $\sum\limits_{i=1}^{k} X_{i:n}$ is empty; therefore, $l_{0,n} = 0$. Also, when $k = n$, we have $l_{n,n} = 1$.

For a better understanding, we plot the points $(k/n, l_{k,n}), k = 0, 1, \ldots n$ on a real plane and connect them. What one obtains is the empirical Lorenz curve, L_n. This curve, as one can see in Fig. (1.2), is well-defined on the entire interval $[0, 1]$ and $L_n(0) = 0, L_n(1) = 1$.

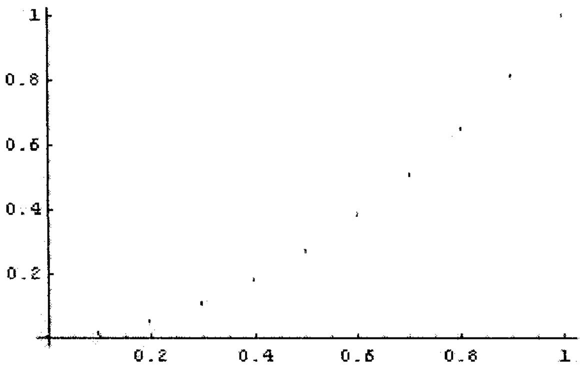

Figure 1.2 Empirical Lorenz curve

As can be seen, when all the incomes are equal, they are plotted on the diagonal I. The diagonal is also an empirical Lorenz curve known as the "egalitarian" Lorenz curve, as one can see in Fig. (1.3). The empirical Lorenz curve, L_n is either below or above the diagonal.

When incomes are not equal, the case of perfect inequality would be the case when all $n-1$ individuals have 0 income and the n^{th} individual has all the wealth.

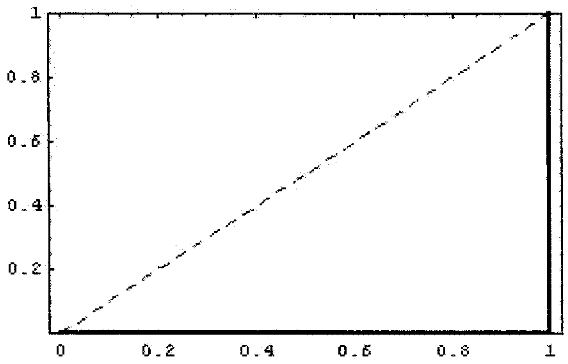

Figure 1.3 Egalitarian Lorenz curve

Thus we define the empirical Lorenz curve as follows:

$$L_n(t) = \frac{1}{\mu} \int_0^t F_n^{-1}(s) ds. \tag{1.3}$$

We can write the Lorenz curve as a sum of n integrals, defined on smaller intervals as follows:

$$L_n(t) = \frac{1}{\overline{X}}\left(\int_0^{1/n} F_n^{-1}(s)ds + \int_{1/n}^{2/n} F_n^{-1}(s)ds + \cdots + \int_{[nt]/n}^{t} F_n^{-1}(s)ds\right)$$
$$= \frac{1}{\overline{X}}\left(\frac{1}{n}X_{1:n} + \frac{1}{n}X_{2:n} + \cdots + \frac{1}{n}X_{[nt]:n} + \left(t - \frac{[nt]}{n}\right)X_{[nt]+1:n}\right),$$

where $\overline{X} = \frac{X_1+X_2+\ldots X_n}{n}$.

Therefore, L_n can be rewritten as follows:

$$L_n(t) = \begin{cases} M_n(t)/(n\overline{x}) & 0 \le t < 1, \\ 1, & t = 1 \end{cases}$$

where

$$M_n(t) := \sum_{i=1}^{[kn]} X_{i:n} + (tn - [tn])x_{[tn]+1:n},$$

with $[tn]$ denoting the largest integer less than or equal to tn.

After comparing this formula with the proportions defined previously, we see that $L_n(k/n) = l_{k,n}$, for any $k = 0, 1, \ldots, n$.

Note that we have

$$0 \le \int_0^1 (t - L_n(t))dt \le \frac{1}{2}. \tag{1.4}$$

In [Gastwirth, 2002] we find the most general definition of the Lorenz curve, which we will present below.

Definition 1.3.3 *Let X be a random variable and $F(x)$ its cumulative distribution function. Let $p = F(x)$. The* **Lorenz curve** *L(p) is the graph of $L(p)$ against p, where*

$$L(p) = \frac{1}{E[X]}\int_0^p F^{-1}(u)du, \tag{1.5}$$

$F^{-1}(s) = inf\{u : F(u) \le s\}$, *and* $0 < s < 1$ *is the quantile function of* F.

Remark. In discussions of personal income, we frequently make statements such as, "the bottom twenty five percent of all households have ten percent of the total income." The Lorenz curve is based on such statements; every point on the curve represents one such statement. A perfectly equal income distribution in a society would be one in which every person has the same income.

The next plot in Fig. (1.4) shows the Lorenz curve, for a Pareto distribution.

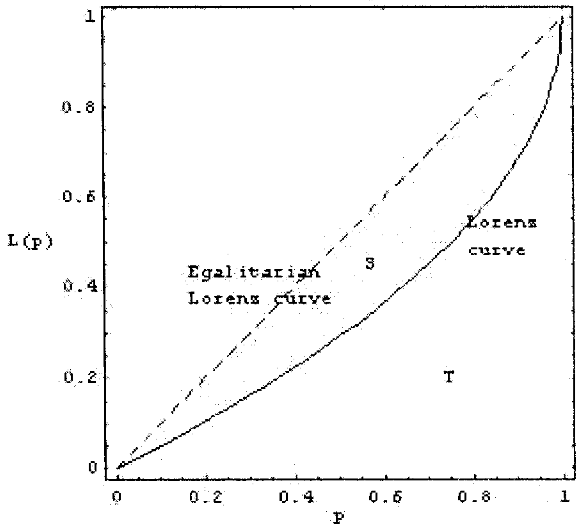

Figure 1.4 Lorenz curve - Pareto distribution

The horizontal axis plots the cumulative percentage of the population whose inequality is under consideration, starting from the poorest and ending with the richest. The vertical axis plots the cumulative percentage of income (or expenditure) associated with the units on the horizontal axis. If the inequality between incomes is higher, then the Lorenz curve will look similar to the plot in Fig. (1.5).

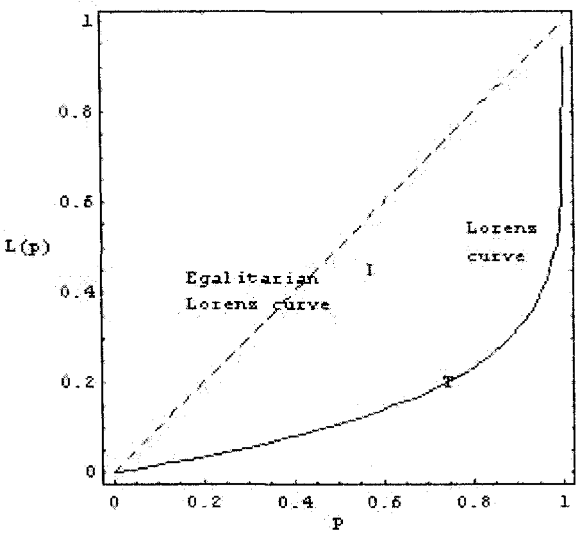

Figure 1.5 Lorenz curve - Pareto distribution

Note that the more evenly spread incomes are, the closer the Lorenz curve will look like that in Fig. (1.1) (the line of equal distribution). The more uneven the distribution of incomes are, the more the curve will look like the one in Fig. (1.5).

We will present below an example obtained from a study conducted by the World Bank. "In Brazil and Hungary, for example, Gross National Product per capita levels are quite comparable, but the incidence of poverty in Brazil is much higher. In Hungary the richest 20 percent (quantile) of the population receives about 4 times more than the poorest quantile, while in Brazil the richest quantile receives at least 30 times more than the poorest quantile" (World Development Index 2002, The World Bank). One may conclude that the Gini index for Hungary is smaller than the Gini index in Brazil.

Figure (1.6) shows that one Lorenz curve deviates from the hypothetical line of absolute equality much more than that of the other Lorenz curve. This means that the first has the highest income inequality.

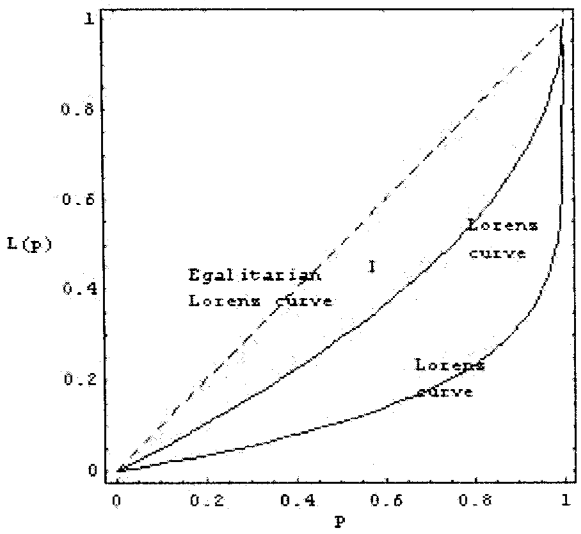

Figure 1.6 Lorenz curves - two countries A and B

In 1912, Corrado Gini made far-reaching contributions to the area by suggesting an index that – just like the Lorenz curve – has been widely used in comparing economic inequality. Intuitively, the Gini index is the ratio of the area between the Lorenz curve $L(p)$ and the diagonal line I , to the area under the diagonal I (which is 1/2). Thus the classical Gini index is twice the area between I and L(p)

$$G_F := 2\int_0^1 (t - L(t))dt. \tag{1.6}$$

If one wishes to obtain a measure of the amount of inequality in the income distribution, one may use the Gini index.

To compute the Gini coefficient, we first measure the area between the Lorenz curve and the 45 degree equality line. For a perfectly equal distribution, there would be no area between the 45 degree line and the Lorenz curve – this would have a Gini coefficient of zero. For complete inequality, in which only one person has any income (if that were possible), the Lorenz curve would coincide with the straight lines at the lower and right boundaries of the curve; thus, the Gini coefficient would be one.

Definition 1.3.4 *The* **Gini index** *G is a a measure of inequality in a population, and is computed by the following formula:*

$$G_F = \frac{1}{2\mu}\mathbf{E}(|X_1 - X_2|) = \frac{1}{2\mu}\int_R\int_R |x_1 - x_2| dF(X_1)dF(X_2), \tag{1.7}$$

where X_1 and X_2 are independent random variables.

We can derive from the formula above, the empirical Gini coefficient:

$$G_n = \frac{1}{2\overline{X}}\int_R\int_R |x-y|dF_n(x)dF_n(y)$$

$$= \frac{1}{2n^2\overline{X}}\sum_{i=1}^{n}\sum_{j=1}^{n}|X_i - X_j|,$$

where

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x).$$

Several other variants of the Gini index were introduced in the literature. In [Donaldson, 1980] the Extended Gini index was introduced, **E-Gini index**, defined as follows:

$$G_{F,\alpha} = 2\left(\int_0^1 (t - L_F(t))^\alpha dt\right)^{1/\alpha}, \tag{1.8}$$

where $\alpha > 1$ and L_F is the Lorenz curve.

The empirical E-Gini estimator formulated [Chakravarty, 1988] is of the form:

$$G_{n,\alpha} := 2\left(\int_0^1 (t - L_n(t))^\alpha dt\right)^{1/\alpha}. \tag{1.9}$$

Note that when $\alpha = 1$, the empirical $G_{n,\alpha}$ is the classical Gini coefficient, G_n.

A year later the Generalized Gini Index, **S-Gini**, was introduced in [Weymark, 1981] and is defined as follows:

$$I_{F,\nu} := 1 - \frac{\nu}{\mu} \int_0^1 F^{-1}(t)(1-t)^{\nu-1} dt, \tag{1.10}$$

where $\nu > 0$ is a parameter.

When $0 < \nu < 1$ (equality aversion), the S-Gini index $I_{F,\nu}$ is well-defined given that $E(|X|^r)$ is finite for some $r > 1/\nu$. When $\nu = 1$ (equality neutrality), then $I_{F,\nu}$ is identically 0. When $\nu > 1$ (inequality aversion) the S-Gini index $I_{F,\nu}$ is also well-defined.

The empirical S-Gini estimator is defined as follows:

$$I_{n,\nu} := 1 - \frac{\nu}{\mu_n} \int_0^1 F_n^{-1}(t)(1-t)^{\nu-1} dt. \tag{1.11}$$

In the above formula, F_n^{-1} denotes the empirical quantile function based on independent random variables $X_1, X_2, \ldots, X_n$, each having the same cumulative distribution function as X, and μ_n is the sample mean of these n random variables. Note that $I_{n,\nu}$ can be expressed by the following formula:

$$I_{n,\nu} = 1 - \frac{1}{\mu_n} \sum_{i=1}^{n} \left((1 - \frac{i-1}{n})^\nu - (1 - \frac{i}{n})^\nu \right) X_{i:n}, \tag{1.12}$$

which is the formula of choice in the econometric literature.

As mentioned above, the Gini index can be used in different areas. We would like to mention below some of the advantages and disadvantages the Gini index presents, from the economist's point of view, that the statisticians keep in mind when developing the theoretical framework:

1. The main advantage of the Gini coefficient is that it is a measure of inequality, not a measure of average income or some other variable which is unrepresentative of most of the population, such as gross domestic product.

2. The Gini index can be used to compare income distributions across different population sectors as well as countries. For example, the Gini index value for urban areas differs from that of rural areas in many countries.

3. The Gini coefficient is sufficiently simple that it can be compared across countries and be easily interpreted.

4. We can use the Gini coefficient to indicate how the distribution of income has changed within a country over a period of time.

5. The Gini coefficient satisfies four important principles:

- it does not matter who the high and low earners are.

- the Gini coefficient does not consider the size of the economy, the way it is measured, or whether it is a rich or poor country on average.

- it does not matter how large the population of the country is.

- if we transfer income from a rich person to a poor person, the resulting distribution is more equal.

The disadvantages of the Gini coefficient as a measure of inequality are listed below:

1. If the Gini coefficient is measured for a large geographically diverse region, then the result is a much higher coefficient than that of each of its composing regions has. For this reason, the scores calculated for individual countries within the E.U. are difficult to compare with the score of the entire U.S.

2. It may be difficult to compare income distributions among countries because the benefit systems may be different in different countries. For example, some countries give benefits in the form of money, others use food stamps, which may or may not be counted as income in the Lorenz curve and therefore are not taken into account in the Gini coefficient.

3. If we apply the Gini index to individuals instead of households, then we get different results. When different populations are not measured with consistent definitions, comparison is not meaningful.

4. It is claimed that the Gini coefficient is more sensitive to the income of the middle classes than to that of the extremes.

A large number of papers analyze the indices of economic inequality, especially the Gini indices, from a mathematical point of view. Asymptotic consistency and normality of the classical Gini coefficient can be found in [Hoeffding, 1948], whereas in [Barrett and Donald, 2001] the empirical and quantile processes point of view was employed and obtained desired asymptotic results for a large class of indices, including the S-and E-Gini indices. Worth mentioning is the work of [Gastwirth, 2002], [Zitikis, 2002] and [Zitikis, 2003] concerning asymptotic results on indices under minimal assumptions on the cumulative distribution function F. [Gastwirth, 2002] and [Zitikis, 2002] proved, for example, that the theory of L-statistics is a most natural tool for investigating the S-Gini index.

Studies dealing with the Gini index are extensive and the index is one of the principal inequality measure used in economics. However, in reality no explicit reason is given for preferring one measure of inequality over another. As such, we focus our attention next on the Atkinson index.

1.4 Atkinson Index Among Other Indices

[Dalton, 1920] suggested that when approaching the question of comparing the distributions (which an economist refers to as income), one should consider directly the form of the welfare function $W(U)$ to be employed.

Note 1.4.1 *The welfare function is used to measure poverty/wealth using inequality measures. It is an absolute equality measure. The inequality is shown as a number, not a percentage. Inequality indices are relative inequality measures. They show inequality in percentages.*

To obtain a specific inequality measure one needs to impose more structure on the welfare function. In [Atkinson, 1970], some assumptions about this welfare function are made before ranking the distributions. It was assumed that this function will be an additively separable and symmetric function (see definitions below) of individual incomes.

Definition 1.4.2 *A function $U : X \to \Re$ is a* **utility function** *representing preference relation $\succeq$ if, for all $x, y \in X$,*

$$x \succeq y \Leftrightarrow U(x) \geq U(y).$$

We note that $\overline{y}$ is the highest observed income and $F(y)$ is the distribution function.

The welfare function relates to the utility derived by an individual or a group to the goods and services that it consumes.

Note 1.4.3 *The extent to which a dollar is deemed to be worth more to a poorer individual that a richer individual depends on the utility and the welfare function W(U).*

Economists employ different utility functions. Some examples are as follows. In portfolio problems the utility functions used are

Example 1.4.4 $U_1(x) = 1 - e^{-kx}$ *and* $U_2(x) = logx$.

In the first example it is very common to consider $k = 1$. We will see later in this section that the welfare function that Atkinson chose, when discussing the Atkinson index, is of the form $U(x) = x^a$, with $0 < a < 1$. When we plot the three functions mentioned above, we see that the Atkinson's utility function lies in between the previous two.

Definition 1.4.5 *An utility function U(x) is* **additively separable** *if it has the form* $U(x) = \Sigma_l U_l(x_l)$, *where* $U_l = U_l(X_l)$ *is the utility of the* l^{th} *person.*

Definition 1.4.6 *A symmetric function on n variables is a function that is unchanged by any permutation of its variables.*

Choosing the welfare function, the ranking of the distributions is done according to

$$W = \int_0^{\overline{y}} U(y)dF(y) \tag{1.13}$$

$$= \int_0^1 U(F^{-1}(s))ds. \tag{1.14}$$

Definition 1.4.7 *A decision maker is risk averse if for any cumulative distribution function* $F(\cdot)$, *the degenerate distribution function* $F^*(\cdot)$ *that yields the mean* $\mu = \int xdF(x)$ *with certainty is at least as good as the distribution function* $F(\cdot)$ *itself.*

This means that a person with concave utility function U is risk averse.

It follows from the definition of risk aversion that the decision maker is risk averse if and only if the inequality

$$\int_{-\infty}^{\infty} U(x)dF(x) \leq U\left(\int_{-\infty}^{\infty} xdF(x)\right),$$

holds for all distribution functions F, provided that the utility function U is a concave function.

In [Dalton, 1920], the author suggested using as a measure of inequality the ratio of the actual level of social welfare to that which would be achieved if income were equally distributed:

$$\frac{1}{U(\mu)} \int_0^\infty U(y) dF(y) = \frac{1}{U(\mu)} \int_0^1 U(F^{-1}(y)) dy.$$

It is seen that this formula is not invariant with respect to linear transformations of the function $U(y)$. For example, in the case of the logarithmic utility function, $U(y) = log(\mu) + c$, Dalton's measure is

$$\frac{1}{\log(\mu) + c} \int_0^{\overline{y}} \log(y) f(y) dy + c,$$

the value of which depends on c.

Using the equally distributed equivalent level of income, in [Atkinson, 1970] a new welfare index is defined. In the case of discrete distributions, this new measure of inequality becomes

$$\begin{aligned} I &= 1 - \left[\sum_{i=1}^{n} \left(\frac{y_i}{\mu} \right)^a f(y_i) \right]^{\frac{1}{a}} \\ &= 1 - \frac{1}{\mu} \left(\sum_{i=1}^{n} y_i^a f(y_i) \right)^{\frac{1}{a}} \\ &= 1 - \frac{1}{\mu} \left(\int_0^\infty y^a dF(y) \right)^{\frac{1}{a}}. \end{aligned}$$

Without using discrete random variables, this could be generally rewritten as

$$I = 1 - \frac{1}{\mu} \left(\mathbf{E}(Y^a) \right)^{\frac{1}{a}}.$$

The Atkinson index is especially useful for confronting the distributional impact of inflation. By setting the social welfare function for the Atkinson index, the researcher may choose to emphasize the lower, middle, or upper end of the income distribution. The Atkinson index's social welfare function, which may be interpreted as the level of inequality aversion, depends on a parameter between 0 and 1. As the parameter approaches its lower limit (i.e., as aversion declines), the index gives more weight to the upper end of the income distribution. However, as the parameter approaches its upper limit, the index measure gives more weight to the lower end of the income distribution.

From [Atkinson, 1970] we obtain the following definition.

Definition 1.4.8 *The* **Atkinson index** *is of the form* $A_{F,a} = 1 - \frac{1}{\mu}\left(\mathbf{E}(Y^a)\right)^{\frac{1}{a}}$.

The utility function used by Atkinson in defining his measure is of the form $U(x) = x^a$. Note that the Atkinson's index is 0 when incomes are equally distributed and converges to 1 as inequality increases. The index increases in a.

The distinguishing feature of the Atkinson index is its ability to measure movements in different segments of the income distribution. Researchers can place greater weight on changes in a given portion of the income distribution by adjusting the a parameter.

All the measures mentioned in this chapter are sensitive to transfers at all income levels. In what follows, we would like to briefly summarize the most important characteristics of each measure, from the economic point of view, in what follows.

Note 1.4.9 *For the* ***Atkinson index*** *the larger a is, the more weight the index attaches to transfers at the low end of the distribution. As a result, less weight is* *attached to transfers at the high end of the distribution.*

In the extreme case where $a \to \infty$, transfers at the lowest end dominate. When $a = 1$, the utility function is linear in income and the distribution of income does not affect the welfare index ($I = 0$ for any income vector).

We note that the sampling distributions of some inequality indices, such as the Gini coefficient are known. However, some researchers may prefer the Atkinson index, because of its properties. The Atkinson index has seen increased applications in empirical analysis of income distributions.

[Ravallion, 1997] modelled the impact of growth and inequality on carbon emissions, using the Atkinson index.

In [Gusenleitner et al., 1998] the author analyzed the distribution of earnings in Austria, over almost three decades. They compared various inequality measures and looked at the trends they uncovered. Authors analyzed trends among various groups of workers and changes over time in the distribution, using both Gini and Atkinson indices. Similar analysis was conducted in [Atkinson, 1970] concerning the income inequality in the U.K.

In [Mayer, 2000] the Gini index was used to estimate the effects income inequality has on mean educational attainment and on the difference in educational attainment between rich and poor children. As an alternative inequality measure the author suggested the Atkinson index.

In [Golan et al., 2001] the minimum wage discussed (which unlike most government transfer programs, lowered welfare in the 1980s and 1990s) is measured by commonly used welfare or inequality measures, including various indices, the Atkinson index, the Gini index, the standard deviation of logarithms, and others. The authors were interested in demonstrating that the minimum wage – in contrast to most government transfer programs – lowers welfare. The Atkinson welfare index has some desirable properties. It is derived as a weighted combination of average income changes and distributional changes. First, the Atkinson welfare index has a dollar-denominated interpretation. Second, the measure for the entire population can be decomposed into within-group and between-group welfare measures for subgroups of the population. Third, changing the single parameter of the Atkinson index, one changes the weight the welfare index places on relative increases of wealth at the lower end of the income distribution. Thus, by varying this parameter, one can examine the effects of government policies over a range of social welfare functions.

In [Lovell, 1998] the author looked at both within country and among country inequalities. In the spirit of [Dalton, 1920] and [Atkinson, 1970] this paper reports estimates of the welfare loss arising from inequality.

In [Londoo, 2000] the changes in aggregate poverty was assessed and inequality that have taken place in Latin America during the past 26 years. With this objective, they examined the income distribution for the region for the period from 1970 - 1995. In their analysis they suggest, following [Atkinson, 1970], that different inequality measures give different weight to different sections of the distribution. As such, it is possible to check on the validity of the results and on the choice of the index.

A large body of literature is devoted to the measurement of income inequality, yet little attention is given to the question "Why measure inequality?" In [Kaplow, 2005] the author considered measures of poverty and emphasized the importance of inequality indices, especially the Atkinson index.

The aim of this study is to develop estimators of the Atkinson index and investigate their asymptotic distributions for hypotheses testing and for building confidence intervals. Also, for application purposes, a check of robustness of the asymptotic theory for finite samples will be investigated by simulation.

CHAPTER 2

EMPIRICAL ATKINSON ESTIMATOR

2.1 Theoretical Atkinson Index

Let X be a non-negative random variable with distribution function F, and let F^{-1} quantile function be defined as

$$F^{-1}(t) = \inf\{x : F(x) \geq t\},\ 0 < t < 1.$$

In [Atkinson, 1970] the following measure of economic inequality was proposed:

$$A_F = 1 - \frac{1}{\mu}\left(\int_0^1 (F^{-1}(t))^a dt\right)^{\frac{1}{a}}, \tag{2.1}$$

where $a > 0$ is a parameter, and μ is the mean of X. Here,

$$0 < \mu < \infty. \tag{2.2}$$

Recall that the Atkinson index A_F can be rewritten in the following form:

$$A_F = 1 - \frac{1}{\mu}(\mathbf{E}(X^a))^{\frac{1}{a}}. \tag{2.3}$$

From Eq. (2.3) it is obvious that the Atkinson index $A_{F,a}$ is well-defined if, in addition to Eq. (2.2), we also have that

$$\mathbf{E}(X^a) < \infty. \tag{2.4}$$

The Atkinson index A_F can be negative, zero, or positive depending on the parameter a.

$$A_F = \begin{cases} \leq 0 & \text{when} \quad a > 1, \\ = 0 & \text{when} \quad a = 1, \\ \geq 0 & \text{when} \quad 0 < a < 1. \end{cases}$$

Therefore, we can conclude the following:

(1) When $0 < a < 1$, then $E(X^a) \leq \mu^a$ which, in turn, implies that $A_F \geq 0$.

(2) Obviously, when $a = 1$, then $A_F = 0$.

(3) When $a > 1$, then $\mu \leq (E(X^a))^{\frac{1}{a}}$, and we thus have that $A_F \leq 0$.

Since in the majority of individuals in a population are risk averse, we will consider only the last case in our simulation, that is $0 \leq a \leq 1$.

2.2 Empirical Atkinson Index

In this section, we discuss the empirical estimator for the Atkinson index A_F, where F refers to a nonparametric distribution or a parametric one. In the case of a parametric distribution, we consider the Pareto, the Exponential and the Log-normal distributions.

One of the most natural ways to obtain an empirical estimator for the Atkinson's index A_F is to replace μ and F^{-1} in Eq. (2.1) by their empirical estimators. Let $X_1, X_2, \ldots X_n$ be independent and identically distributed random variables. Let $\bar{X}$ be the sample mean of $X_1, X_2, \ldots X_n$, that is,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

and let F_n be the empirical distribution function. Then, it is seen that

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x),$$

where $-\infty < x < \infty$.

The corresponding empirical quantile function is defined as

$$F_n^{-1}(t) = \inf\{x : F_n(x) \geq t\}.$$

Hence, the empirical Atkinson's index is defined as follows:

$$A_n = 1 - \frac{1}{\bar{X}} \left(\int_0^1 (F_n^{-1}(t))^a dt \right)^{\frac{1}{a}}. \tag{2.5}$$

We shall now rewrite A_n in a slightly different way, which will be more computationally convenient. For this reason, we first note that F_n^{-1} allows the following explicit representation:

$$F_n^{-1}(t) = X_{i:n}, \ \frac{i-1}{n} < t \leq \frac{i}{n} \tag{2.6}$$

for any $i = 1, \ldots n$, where $X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$ are the order statistics of $X_1, X_2, \ldots X_n$. Introducing Eq. (2.6) into Eq. (2.5) one obtains

$$A_n = 1 - \frac{1}{\bar{X}} \left(\frac{1}{n} \sum_{i=1}^{n} X_i^a \right)^{\frac{1}{a}}. \tag{2.7}$$

2.2.1 Consistency of the Estimator

Definition 2.2.1 *Let $X_1, X_2 \ldots$ and X_n be random variables on a probability space $(\Omega, \mathcal{A}, P)$. We say that X_n* **converges in probability** *to X if for any $\epsilon > 0$*

$$\lim_{n \to \infty} \mathbf{P}(|X_n - X| < \epsilon) = 1.$$

This is written as $X_n \to^p X$.

Definition 2.2.2 *Let $\xi_1, \xi_2 \ldots$ and ξ_n be random variables on a probability space $(\Omega, \mathcal{A}, P)$. We say that ξ_n* **converges with probability 1** *to ξ if*

$$\mathbf{P}(\lim_{n \to \infty} \xi_n = \xi) = 1. \tag{2.8}$$

This is written as

$$\xi_n \rightarrow^{wp1} \xi.$$

Note that Eq. (2.8) means that there exists a set $\Omega_0 \subseteq \Omega$ such that $\mathbf{P}(\Omega_0) = 1$, $\lim_{n\to\infty} \xi_n(\omega) = \xi(\omega)$ for every $\omega \in \Omega_0$.

Definition 2.2.3 *An estimator θ of Y is a* **strong consistent estimator** *if it becomes almost certain that the value of θ gets closer to the value of Y as the sample size increases. We can formalize this as follows: For almost every elementary event $\omega \in \Omega$ we have*

$$|Y(\omega) - \theta| \to 0 \textit{ as } n \to \infty,$$

where n is the sample size.

Definition 2.2.4 *An estimator θ of Y is a* **weakly consistent estimator** *if for any $\epsilon > 0$, $\mathbf{P}(|Y - \theta| > \epsilon) \to 0$, when $n \to \infty$.*

Theorem 2.2.5 *$\bar{X}$ and $\frac{1}{n}\sum X_i^a$ converge (strongly and weakly) to μ and $\mathbf{E}(X^a)$, respectively. Under the assumptions in Eqs. (2.2) and (2.4), A_n is a (strongly and weakly) consistent estimator of A_F.*

Proof:

It is enough to prove only that A_n is a strong consistent estimator of A_F, because weak consistency follows from the strong one. That is, we want to prove that there exists a set $\Omega_0 \subseteq \Omega$ such that $\mathbf{P}(\Omega_0) = 1$ and

$$A_n(w) - A_F \to 0,$$

for every $\omega \in \Omega$. We proceed with the representation:

$$A_n(w) - A_F = -\left(\frac{\bar{X}_a^{\frac{1}{a}}}{\bar{X}} - \frac{\mu_a{}^{\frac{1}{a}}}{\mu} \right)$$

$$= -\left(\frac{\bar{X}_a^{\frac{1}{a}} - \mu_a^{\frac{1}{a}}}{\bar{X}} + \mu_a{}^{\frac{1}{a}}\left(\frac{1}{\bar{X}} - \frac{1}{\mu}\right)\right). \tag{2.9}$$

Note that by The Strong Law of Large Numbers, we have the following two statements:

(1) There exists a set $\Omega_1 \subseteq \Omega$ such that $\mathbf{P}(\Omega_1) = 1$ and

$$\bar{X}(w) \to \mu, \tag{2.10}$$

for every $\omega \in \Omega_1$.

(2) There exists a set $\Omega_2 \subseteq \Omega$ such that $\mathbf{P}(\Omega_2) = 1$ and

$$\frac{1}{n}\sum_{i=1}^{n} X_i^a(w) \to \mathbf{E}(X^a) = \mu_a, \tag{2.11}$$

for every $\omega \in \Omega_2$.

Let us define

$$\Omega_0 = \Omega_1 \cap \Omega_2.$$

For every $\omega \in \Omega_0$, we have both Eqs. (2.10) and (2.11). By Eq. (2.11),

$$\bar{X}(\omega)_a{}^{\frac{1}{a}} = \left(\frac{1}{n}\sum_{i=1}^{n} X(\omega)_i^a\right)^{\frac{1}{a}} \to \mathbf{E}(X^a)^{\frac{1}{a}}$$

for every $\omega \in \Omega_0$, which gives

$$\frac{\bar{X}(\omega)_a{}^{\frac{1}{a}} - \mu_a^{\frac{1}{a}}}{\bar{X}(\omega)} \to 0. \tag{2.12}$$

Furthermore, by Eq. (2.10),

$$\bar{X}(\omega) \to \mu,$$

for every $\omega \in \Omega_0$ which gives

$$\mu_a{}^{\frac{1}{a}}\left(\frac{1}{\bar{X}(\omega)} - \frac{1}{\mu}\right) \to 0. \tag{2.13}$$

By Eqs. (2.9), (2.12), (2.13) it is seen that

$$A_n(w) - A_F \to 0,$$

for every $\omega \in \Omega_0$. Consequently, we have shown that $A_{n,a}$ is a strongly consistent estimator of $A_{F,a}$, provided that $\mathbf{P}(\Omega_0) = 1$. To prove the latter, one may show that $\mathbf{P}(\bar{\Omega}_0) = 0$. It can be seen that

$$\bar{\Omega}_0 = \overline{\Omega_1 \cap \Omega_2} = \bar{\Omega}_1 \cup \bar{\Omega}_2.$$

Note that $\mathbf{P}(\bar{\Omega}_1) = 0$ and $\mathbf{P}(\bar{\Omega}_2) = 0$ since $\mathbf{P}(\Omega_1) = 1$ and $\mathbf{P}(\Omega_2) = 1$. Therefore,

$$\mathbf{P}(\bar{\Omega}_0) = \mathbf{P}(\bar{\Omega}_1 \cup \bar{\Omega}_2) \leq \mathbf{P}(\bar{\Omega}_1) + \mathbf{P}(\bar{\Omega}_2) = 0.$$

This proves that $\mathbf{P}(\Omega_0) = 1$, and thus Theorem (2.2.5) as well. ■

CHAPTER 3

EMPIRICAL ATKINSON INDEX: ONE POPULATION

3.1 Asymptotic Normality

3.1.1 Nonparametric Case

Let us define again the Atkinson index $A_F = 1 - \frac{1}{\mu}(E(X^a))^{\frac{1}{a}}$ and denote $\mu_a = E(X^a)$. Hence, A_F can be expressed as

$$A_F = 1 - \frac{1}{\mu}(\mu_a)^{\frac{1}{a}},$$

where $\mu_a = E(X^a)$.

We construct the nonparametric Atkinson estimator by replacing F^{-1} by F_n^{-1}. Using Eq. (2.5) and the representation of $F_n^{-1}(t) = X_{i:n}$, $\frac{i-1}{n} < t < \frac{i}{n}$, we obtain the following expression:

$$A_n[X] = 1 - \frac{1}{\overline{X}}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^a\right)^{\frac{1}{a}}, \tag{3.1}$$

with $X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$ denoting the order statistics of $X_1, X_2, \ldots, X_n$. With $\overline{X}_a = \frac{1}{n}\sum_{i=1}^{n} X_i^a$, Eq. (3.1) becomes

$$A_n[X] = 1 - \frac{1}{\overline{X}}(\overline{X}_a)^{\frac{1}{a}}.$$

The asymptotic normality of the appropriately centered and normalized $A_n[X]$ is obtained in the following theorem.

Theorem 3.1.1 *Under the assumptions (2.2), (2.4) and also assuming that*

$$\mathbf{E}(X^2) < \infty, \ \mathbf{E}(X^{2a}) < \infty, \tag{3.2}$$

the asymptotic distribution of $\sqrt{n}(A_n[X] - A_F)$ *is normal with mean zero and variance*

$$\sigma^2_{F,a} = \frac{1}{a^2\mu^2}\mu_a^{\frac{2}{a}-2}(\mu_{2a} - \mu_a^2) - \frac{2}{a\mu^3}\mu_a^{\frac{2}{a}-1}(\mu_{a+1} - \mu\mu_a) + \frac{\mu_a^{\frac{2}{a}}}{\mu^4}(\mu_2 - \mu^2). \tag{3.3}$$

Proof:

To prove Theorem (3.1.1) one needs to make use of the Taylor expansion along with the Delta Method. These are stated below.

Theorem 3.1.2 (Taylor expansion) *If a function f has continuous derivatives up to the $(n+1)^{th}$ order, then this function can be expanded in the following manner:*

$$f(x+h) = f(x) + hf^{(1)}(x) + h^2\frac{f^{(2)}(x)}{2!} + \cdots + h^k\frac{f^{(k)}(x)}{k!} + R$$

where the remainder term is

$$R = \int_x^{x+h} \frac{f^{(k+1)}(t)(x+h-t)^k}{k!}dt.$$

Theorem 3.1.3 (Delta Method) *Let Y_n be a sequence of random variables that satisfy $\sqrt{n}(Y_n - \theta) \to n(0, \sigma^2)$ in distribution. For a given function g of Y_n and a specific value of θ, suppose that $g'(\theta)$ exists and is not 0. Then*

$$\sqrt{n}[g(Y_n) - g(\theta)] \to n(0, \sigma^2[g'(\theta)]^2) \text{ in distribution.}$$

Consider the function $h(x,y) = 1 - \frac{1}{x}y^{1/a}$. From Eqs. (2.3) and (2.7) we have that $A_F = h(\mu, \mu_a)$ and $A_n = h(\overline{X}, \overline{X}_a)$.

Applying the Delta Method, we want to show that

$$\sqrt{n}\,(A_n - A_F) \to^d n(0, \sigma^2_{F,a}).$$

But $\sqrt{n}\,(A_n - A_F) = \sqrt{n}\,\big(h(\mu,\mu_a) - h(\overline{X},\overline{X}_a)\big)$, therefore we have

$$\begin{aligned}\sqrt{n}\,(h(\mu,\mu_a)) - h(\overline{X},\overline{X}_a) &= \sqrt{n}\,\big(h'_x(\mu,\mu_a)(\overline{X}-\mu) + h'_y(\overline{X}_a - \mu_a) + \dots\big)\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[h'_x(\mu,\mu_a)(X_i-\mu) + h'_y(\mu_i,\mu_a)(X_i^a - \mu_a)].\end{aligned}$$

Therefore one obtains

$$\sqrt{n}\,\big(h(\overline{X},\overline{X}_a) - h(\mu,\mu_a)\big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[-\frac{1}{\mu a}(\mu_a)^{\frac{1}{a}-1}(X_i^a - \mu_a) + \frac{1}{\mu^2}(\mu_a)^{\frac{1}{a}}(X_i - \mu)].$$

Let

$$Y_i = -\frac{1}{\mu a}(\mu_a)^{\frac{1}{a}-1}(X_i^a - \mu_a) + \frac{1}{\mu^2}(\mu_a)^{\frac{1}{a}}(X_i - \mu). \tag{3.4}$$

Hence the random variables $Y_1, Y_2, \dots, Y_n$ are independent, identically distributed, with means zero and variance defined in Eq. (3.3). Therefore $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i$ is asymptotically normal with mean 0 and finite variance $E(Y_i^2)$, by the Central Limit Theorem. We can easily see that

$$E[Y_i^2] = \frac{\mu_a^{\frac{2}{a}-2}}{\mu^2 a^2}E[(X_i^a - \mu_a)^2] + \frac{\mu_a^{\frac{2}{a}}}{\mu^4}E[(X_i - \mu)^2] - \frac{2}{a\mu^3}\mu_a^{\frac{2}{a}-1}E[(X_i^a - \mu)(X_i - \mu)],$$

which gives us the formula from Theorem (3.1.1). ■

3.1.2 Parametric Case

For the parametric discussion we consider that income follows one of the three parametric families: Pareto, Exponential and Lognormal. We construct confidence intervals by applying standard asymptotic theory for the maximum likelihood estimators.

For the **Pareto distribution**, the cumulative distribution function is

$$F(x) = 1 - \left(\frac{x_0}{x}\right)^{\lambda}.$$

The maximum likelihood estimator for λ is

$$\widehat{\lambda} = \left(\frac{1}{n} \sum_{i=1}^{n} log \left(\frac{x_i}{x_0} \right) \right)^{-1}. \tag{3.5}$$

From Appendix A, the Atkinson index for the Pareto distribution is

$$A_F = 1 - (1 - \frac{1}{\lambda})(\frac{\lambda}{\lambda - a})^{\frac{1}{a}}.$$

Hence, the empirical Atkinson index for the Pareto distribution is

$$\widehat{A}_F = 1 - (1 - \frac{1}{\widehat{\lambda}})(\frac{\widehat{\lambda}}{\widehat{\lambda} - a})^{\frac{1}{a}}.$$

Theorem 3.1.4 *Assuming that $E(X^2) < \infty$ and $E(X^{2a}) < \infty$, the asymptotic distribution of $\sqrt{n}(\widehat{A}_F - A_F)$ is normal with mean zero and variance*

$$\widehat{\sigma}_F^2 = \frac{(1-a)^2 \lambda^{\frac{2}{a}-2}}{(\lambda - a)^{\frac{2}{a}+2}}.$$

Proof:

Define the theoretical and the empirical Atkinson indices as a function h of λ and $\widehat{\lambda}$, respectively. Hence $A_F = h\left(\frac{1}{\lambda}\right)$ and $\widehat{A}_F = h\left(\frac{1}{\widehat{\lambda}}\right)$.

From Eq. (3.5), we have $\frac{1}{\widehat{\lambda}} = \frac{1}{n} \sum_{i=1}^{n} log \left(\frac{x_i}{x_0} \right)$. Let Z as follows, $Z = log \left(\frac{X}{x_0} \right)$, therefore $\frac{1}{\widehat{\lambda}} = \frac{1}{n} \sum_{i=1}^{n} Z_i$, and

$$\sqrt{n}(\widehat{A}_F - A_F) = \sqrt{n} \left[h\left(\frac{1}{\widehat{\lambda}}\right) - h\left(\frac{1}{\lambda}\right) \right] = \sqrt{n} \left[h\left(\frac{1}{n} \sum_{i=1}^{n} Z_i \right) - h\left(\frac{1}{\lambda}\right) \right]. \tag{3.6}$$

Since $F(x) = 1 - \left(\frac{x_0}{x}\right)^{\lambda} = 1 - \left[e^{log\left(\frac{x_0}{x}\right)} \right]^{\lambda} = 1 - \left[e^{log\left(\frac{x}{x_0}\right)^{-1}} \right]^{\lambda}$, it is seen that for $z = log\left(\frac{x}{x_0}\right)$ the distribution function for z is $F(z) = 1 - e^{-\lambda z}$.

Hence, the expected value and the variance for Z are $\frac{1}{\lambda}$ and $\frac{1}{\lambda^2}$, respectively.

By the Law of Large Numbers, we have $\frac{1}{\lambda} = \frac{1}{n}\sum_{i=1}^{n} Z_i \to^p \frac{1}{\lambda} = E[Z]$.

Also,

$$\sqrt{n}\left[h\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\right) - h\left(\frac{1}{\lambda}\right)\right] = \sqrt{n}\left[h\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\right) - \frac{1}{\lambda}\right] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(Z_i - \frac{1}{\lambda}\right),$$

which by the Central Limit Theorem has a limiting normal distribution.

Expanding Eq. (3.6) using the Taylor series around $\frac{1}{\lambda}$ we obtain

$$\sqrt{n}(\widehat{A}_F - A_F) = \sqrt{n}\left[h'\left(\frac{1}{\lambda}\right)\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \frac{1}{\lambda}\right) + \frac{1}{2}h''\left(\frac{1}{\lambda}\right)\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \frac{1}{\lambda}\right)^2\right], \tag{3.7}$$

where the first and the second derivative for the h function are

$$h'\left(\frac{1}{\lambda}\right) = \frac{\left[1 + \left(\frac{1-\lambda}{\lambda-a}\right)\right]\lambda^{1/a}}{(\lambda - a)^{1/a}} \tag{3.8}$$

and

$$h''\left(\frac{1}{\lambda}\right) = \frac{(1-a)(\lambda+1)(\lambda)^{1+1/a}}{(\lambda-a)^{2+1/a}}. \tag{3.9}$$

As shown in Chapter 2, the second term in Eq. (3.7) (the remainder) converges in probability to 0. Therefore $\sqrt{n}(\widehat{A}_F - A_F)$ converges to a normal distribution with mean 0 and variance

$$\widehat{\sigma}_F^2 = \left(h'\left(\frac{1}{\lambda}\right)\right)^2 Var(Z) = \frac{(1-a)^2\lambda^{\frac{2}{a}-2}}{(\lambda-a)^{\frac{2}{a}+2}}. \tag{3.10}$$

Hence,

$$\frac{\sqrt{n}(\widehat{A}_F - A_F)}{\sqrt{\left(h'\left(\frac{1}{\lambda}\right)\right)^2 Var(z)}} \to^d N(0,1).$$

■

For the **Exponential distribution**, the cumulative distribution function is

$$F(x) = 1 - e^{-\frac{x-x_0}{\theta}},$$

with $x > x_0$ and $\theta > 0$. Here, the parameter x_0 is known, and can be equated to the basic income, or social support income. The parameter θ is unknown.

The maximum likelihood estimator for θ is

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n}(x_i - x_0). \tag{3.11}$$

From Appendix A, we have that the Atkinson index for the Exponential distribution is

$$A_{F,a} = 1 - \Gamma(a+1)^{\frac{1}{a}} = c. \tag{3.12}$$

It is interesting to observe that for the exponential distribution, the Atkinson index does not depend on the parameter λ.

Since the Atkinson index for the exponential distribution is a constant, one is interested in determining if the Exponential distribution is valid.

This can be tested by the following null and alternative hypothesis:

$$H_0 : A_F = c;$$

vs

$$H_a : A_F \neq c.$$

For the **Lognormal distribution**, the cumulative distribution function is

$$F(x) = \Phi\left(log(x - x_0) - \mu\right),$$

with $x > x_0$ and $-\infty < \mu < \infty$. The parameter σ is considered to be 1. We consider the parameter x_0 to be known and we can interpret this parameter as basic income, or social support income. The parameter μ is unknown. In the case of Lognormal distribution without the location parameter x_0, the a^{th} moment about x_0 is $E[(X - X_0)^a] = e^{a\mu + \frac{a^2}{2}}$. In the case of Lognormal distribution with location parameters, the moments are calculated using the the moments presented above and Newton's Binomial theorem.

Therefore, the expected value in the Lognormal with x_0 case is $E[X] = x_0 + e^{\mu+\frac{1}{2}}$. The second moment is $E[X^2] = 2x_0e^{\mu+\frac{1}{2}} + e^{2\mu+2} - x_0^2$. The third moment is $E[X^3] = x_0^3 - 3x_0^2e^{\mu+\frac{1}{2}} + 3x_0e^{2\mu+2} + e^{3\mu+\frac{9}{2}}$. The fourth moment is $E[X^4] = 4x_0^3e^{\mu+\frac{1}{2}} - 6x_0^2e^{2\mu+2} + 4x_0e^{3\mu+\frac{9}{2}} + e^{4\mu+8} - x_0^4$.

The maximum likelihood estimator for μ is

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} log\left(x_i - x_0\right). \tag{3.13}$$

From Appendix A, the Atkinson index for the Lognormal distribution is

$$A_F = 1 - \frac{(x_0 + e^{a\mu+\frac{a^2}{2}})^{\frac{1}{a}}}{x_0 + e^{\mu+\frac{1}{2}}}.$$

Hence, the empirical Atkinson index for the Lognormal distribution is

$$\widehat{A}_F = 1 - \frac{(x_0 + e^{a\widehat{\mu}+\frac{a^2}{2}})^{\frac{1}{a}}}{x_0 + e^{\widehat{\mu}+\frac{1}{2}}}.$$

Theorem 3.1.5 *Assuming that the* $E(X^2) < \infty, E(X^{2a}) < \infty$, *the asymptotic distribution of* $\sqrt{n}(\widehat{A}_F - A_F)$ *is normal with mean zero and variance*

$$\widehat{\sigma}_F^2 = \left[\left(1 - \frac{1}{x_0 + e^{a\mu+\frac{a^2}{2}}}\right)\frac{\left(x_0 + e^{a\mu+\frac{a^2}{2}}\right)^{\frac{1}{a}}}{x_0 + e^{\mu+\frac{1}{2}}}\right]^2.$$

Proof:

We define the theoretical and the empirical Atkinson indices as a function h of μ and $\widehat{\mu}$, respectively. Therefore $A_F = h(\mu)$ and $\widehat{A}_F = h(\widehat{\mu})$.

We also define the random variable Z such that $z_i = log(x_i - x_0)$. Therefore, from Eq. (3.13) $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} z_i$. The expected value is $E[Z] = \mu$ and the variance is $Var(Z) = \sigma^2$.

Hence,

$$\sqrt{n}(\widehat{A}_F - A_F) = \sqrt{n}\left[h(\widehat{\mu}) - h(\mu)\right] = \sqrt{n}\left[h\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\right) - h\left(\mu\right)\right]. \tag{3.14}$$

By the Law of Large Numbers, we have $\widehat{\mu} = \frac{1}{n}\sum\limits_{i=1}^{n} Z_i \to^{p} \mu = E[z_i]$.

Also,

$$\sqrt{n}\left[h\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\right) - h\left(\mu\right)\right] = \sqrt{n}\left[h\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\right) - \mu\right] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(Z_i - \mu\right),$$

which by the Central Limit Theorem converges to $N(0, Var(Z))$.

Expanding Eq. (3.14) using the Taylor series, around μ, one obtains

$$\sqrt{n}(\widehat{A}_F - A_F) = \sqrt{n}\left[h'\left(\mu\right)\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mu\right) + \frac{1}{2}h''\left(\mu\right)\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mu\right)^2\right]. \tag{3.15}$$

By calculating the first and second derivatives for the h function we obtain

$$h'\left(\mu\right) = \left(1 - \frac{1}{x_0 + e^{a\mu + \frac{a^2}{2}}}\right)\frac{\left(x_0 + e^{a\mu + \frac{a^2}{2}}\right)^{\frac{1}{a}}}{x_0 + e^{\mu + \frac{1}{2}}} \tag{3.16}$$

and

$$h''\left(\mu\right) = \frac{(a-1)\left(x_0 + e^{a\mu + \frac{a^2}{2}}\right)^{\frac{1}{a}-2} - \left(x_0 + e^{a\mu + \frac{a^2}{2}}\right)}{x_0 + e^{\mu + \frac{1}{2}}}. \tag{3.17}$$

The second term in Eq. (3.15) converges in probability to 0. Therefore, we can conclude that $\sqrt{n}(\widehat{A}_F - A_F)$ converges to a normal distribution with mean 0 and variance

$$\sigma_F^2 = \left(h'\left(\mu\right)\right)^2 Var(z) = \left[\left(1 - \frac{1}{x_0 + e^{a\mu + \frac{a^2}{2}}}\right)\frac{\left(x_0 + e^{a\mu + \frac{a^2}{2}}\right)^{\frac{1}{a}}}{x_0 + e^{\mu + \frac{1}{2}}}\right]^2. \tag{3.18}$$

Hence,

$$\frac{\sqrt{n}(\widehat{A}_F - A_F)}{\sqrt{\left(h'\left(\mu\right)\right)^2 Var(z)}} \rightarrow^d N(0, 1).$$

■

3.2 Asymptotic Variance and its Estimation

To obtain an estimator for the asymptotic variance we replace $\mu, \mu_a, \mu_{2a},$ and μ_{a+1} in Theorem (3.1.1) by their corresponding estimates.

Therefore, the estimator for the asymptotic variance is

$$\sigma_n^2 = \frac{1}{a^2\widehat{\mu}^2}\widehat{\mu}_a^{\frac{2}{a}-2}(\widehat{\mu}_{2a} - \widehat{\mu}_a^2) - \frac{2}{a\widehat{\mu}^3}\widehat{\mu}_a^{\frac{2}{a}-1}(\widehat{\mu}_{a+1} - \widehat{\mu}\widehat{\mu}_a) + \frac{\widehat{\mu}_a^{\frac{2}{a}}}{\widehat{\mu}^4}(\widehat{\mu}_2 - \widehat{\mu}^2). \tag{3.19}$$

The estimator σ_n^2 converges in probability to the asymptotic variance $\sigma_{F,a}^2$.

A simulation study will be undertaken in order to determine the validity and robustness of the asymptotic results for finite samples.

3.3 Simulation Studies

Simulation of the Atkinson index were be run for different distributions (Pareto, Exponential and Lognormal) with sample sizes, n = 100, n = 500, n =1,000 and n =10,000 and m =1,000 replications.

We are interested in calculating 95% confidence intervals and in determining the coverage probabilities for the different sample sizes. The coverage probability is calculated as (Number of CI that cover the theoretical index)/(m).

When m is very large the coverage probability should be approximately 0.95.

3.3.1 Confidence Intervals: Nonparametric Case

For $X_1, X_2 \ldots X_n$ independent and identically distributed random variables

$$\frac{A_n - A_F}{\sqrt{\frac{\sigma^2_{F,a}}{n}}} \rightarrow^d N(0,1).$$

Theorem 3.3.1 *(SLUTSKY'S Theorem) Let X_n and Y_n be sequences of random variables. Suppose X_n converges to X in distribution, and Y_n converges to the constant b in probability. Then*

1. *$X_n + Y_n$ converges to $X + b$ in distribution.*

2. *$X_n * Y_n$ converges to bX in distribution.*

3. *$\frac{X_n}{Y_n}$ converges to $\frac{X}{b}$ in distribution.*

Using Slutsky's argument we can conclude that the following result is true:

$$\frac{A_n - A_F}{\sqrt{\frac{\sigma^2_n}{n}}} \rightarrow^d N(0,1). \tag{3.20}$$

Equation (3.20) can be used to construct confidence intervals. A $100(1-\alpha)\%$ asymptotic confidence interval is given by

$$A_n \pm z_{\alpha/2}\sqrt{\frac{\sigma^2_n}{n}}.$$

For constructing a bootstrap confidence interval, one starts with the original sample $X_1, X_2 \ldots X_n$ and calculates the empirical Atkinson index $A_n[X]$. Next, a random sample with replacement will be drawn from the original sample to obtain a new sample $X_1^*, X_2^* \ldots X_n^*$. Then the Atkinson index $A_n[X^*]$ and the expression

$$\sqrt{n}|A_n[X^*] - A_n[X]| \tag{3.21}$$

are calculated. This procedure is repeated m times to obtain m values of expression (3.21).

Then we define z^*_α as the smallest value of z such that at least $100(1-\alpha)\%$ values of expression (3.21) are at or below z.

As such, the bootstrap confidence interval for the nonparametric bootstrap is given by

$$A_n \pm z^*_\alpha \frac{1}{\sqrt{n}}.$$

3.3.2 Confidence Intervals: Parametric Case

To construct parametric confidence intervals based on maximum likelihood estimators, one uses the same idea as in the previous section.

As such, a $100(1-\alpha)\%$ asymptotic confidence interval is given by

$$A_n \pm z_{\alpha/2}\sqrt{\frac{\widehat{\sigma}^2_F}{n}}.$$

For constructing a bootstrap confidence interval, one starts with an original sample $X_1, X_2 \ldots X_n$ from one of the parametric distributions. The probability density function for the distribution considered is of the form $f(X|\theta)$, where θ is the parameter of the distribution. Let $\widehat{\theta}$ be the maximum likelihood estimator of θ and let $X_1, X_2 \ldots X_n$ be a random sample from $f(X|\widehat{\theta})$.

A new random sample will be drawn, with replacement, from $X_1, X_2 \ldots X_n$ to obtain a new sample $X^*_1, X^*_2 \ldots X^*_n$. Then, the Atkinson index $A_n[X^*]$ and the expression

$$\sqrt{n}|A_n[X^*] - A_n[X]| \tag{3.22}$$

are calculated. This procedure is repeated m times and m values of the expression

(3.22) are obtained. We define z^*_α as the smallest z such that at least $100(1-\alpha)\%$ values of the quantity $\sqrt{n}[A_n[X^*] - A_n[X]]$ are at or below z.

As such, a $100(1-\alpha)\%$ bootstrap confidence interval for the parametric bootstrap is given by

$$A_n \pm z^*_\alpha \frac{1}{\sqrt{n}}.$$

CHAPTER 4

SIMULATION STUDIES I

4.1 Pareto Distribution

In this section we generated 1,000 Pareto samples of different sizes, $n = 100, n = 500, n = 1{,}000$ and $n = 10{,}000$ and calculated the Atkinson index for each sample in order to compare the empirical estimators and the 95% confidence intervals or coverage probabilities to the theoretical Atkinson index and asymptotic confidence interval, respectively. The Pareto random variable may be considered as representing salary distribution. The parameter of the Pareto distribution is $\lambda = 3$ and the Atkinson parameter is $a = 0.2$. Of interest is calculating the 95% confidence interval or coverage probability for the Atkinson index. In this case the expected α is 0.05.

Results for the nonparametric case show that for sample size $n = 100$ with $m = 1{,}000$ replicates there were 25 intervals not covering the theoretical Atkinson (14 smaller that the lower limit and 11 greater than the upper limit). For the parametric case there were 59 intervals not covering the theoretical Atkinson (2 smaller that the lower limit and 57 greater than the upper limit).

For the sample size $n = 500$, with $m = 1{,}000$ replicates there were 34 intervals not covering the theoretical Atkinson (25 smaller that the lower limit and 9 greater than the upper limit) in the nonparametric case, and there were 56 intervals not covering the theoretical Atkinson (12 smaller that the lower limit and 44 greater than the upper limit) for the parametric case.

For the sample size n = 1,000, with m = 1,000 replicates there were 46 intervals not covering the theoretical Atkinson (31 smaller that the lower limit and 15 greater than the upper limit) in the nonparametric case, and there were 57 intervals not covering the theoretical Atkinson (20 smaller that the lower limit and 37 greater than the upper limit) for the parametric case.

For the sample size n = 10,000, with m = 1,000 replicates there were 50 intervals not covering the theoretical Atkinson (35 smaller that the left lower and 15 greater than the upper limit) in the nonparametric case, and there were 48 intervals not covering the theoretical Atkinson (18 smaller that the lower limit and 30 greater than the upper limit) for the parametric case. We summarize these results in Table 4.1.

Table 4.1 Coverage probability of the Atkinson index from simulation using the Pareto distribution with 1,000 replicates

Sample size	Proportion Nonparametric	Proportion Parametric
n=100	0.975	0.941
n=500	0.966	0.944
n=1,000	0.954	0.943
n=10,000	0.95	0.952

From results in Table 4.1 for α = 0.05, one concludes that as the sample size increases, the coverage $(1 - \alpha)$ converges to 0.95 or the type I error (α) converges to 0.05.

In practice, some populations, such as salaries at institutions, may be relatively small, in which case results for $n = 100$ or $n = 500$ would be relevant. In such case, results in Table 4.1 show the $1 - \alpha$ level to be somewhat larger than 0.95 for the nonparametric case, but close to 0.95 in the parametric case. These deviations from expected $(1 - \alpha = 0.05)$ are deemed acceptable for application purposes.

Appendix B, presents the program code for the simulation, in the case of $n =$ 10,000, using R. The random variables generated represent incomes at a scale of 1:30,000. Therefore, the minimum income in the income vector was Income$_{min}$ = \$30,000.26 and the maximum income Income$_{max}$ = \$659,758.90 with the mean income being \$44,979.13. In what follows, the theoretical Atkinson index is $A_{F,a} =$ 0.05870694, the vector $Atkinsonn[k]$ represents the nonparametric Atkinson estimates for m = 1,000 replicates, and the vector $Atkinsonpe[k]$ represents the parametric Atkinson estimates for m = 1,000 replicates. The parametric variance is $\sigma_F^2 = 0.0180823$, and the vector $\overline{\sigma}_F^2$ represents the parametric estimates for m = 10,000 replicates. The nonparametric variance is $\sigma_n^2 = 0.0807779$, and the vector $\overline{\sigma}_n^2$ represents the nonparametric estimates for m = 1,000 replicates.

The plot in Fig. (4.1) represents a histogram of the incomes for sample size 10,000 from the Pareto distribution. It is seen, as expected, that the empirical distribution agrees with the theoretical Pareto distribution.

Figure (4.2) represents a scatter plot of the 1,000 nonparametric estimates of the Atkinson index for n = 10,000 incomes following a Pareto distribution.

Figure (4.3) represents a plot of the 1,000 parametric Atkinson indices for n = 10,000.

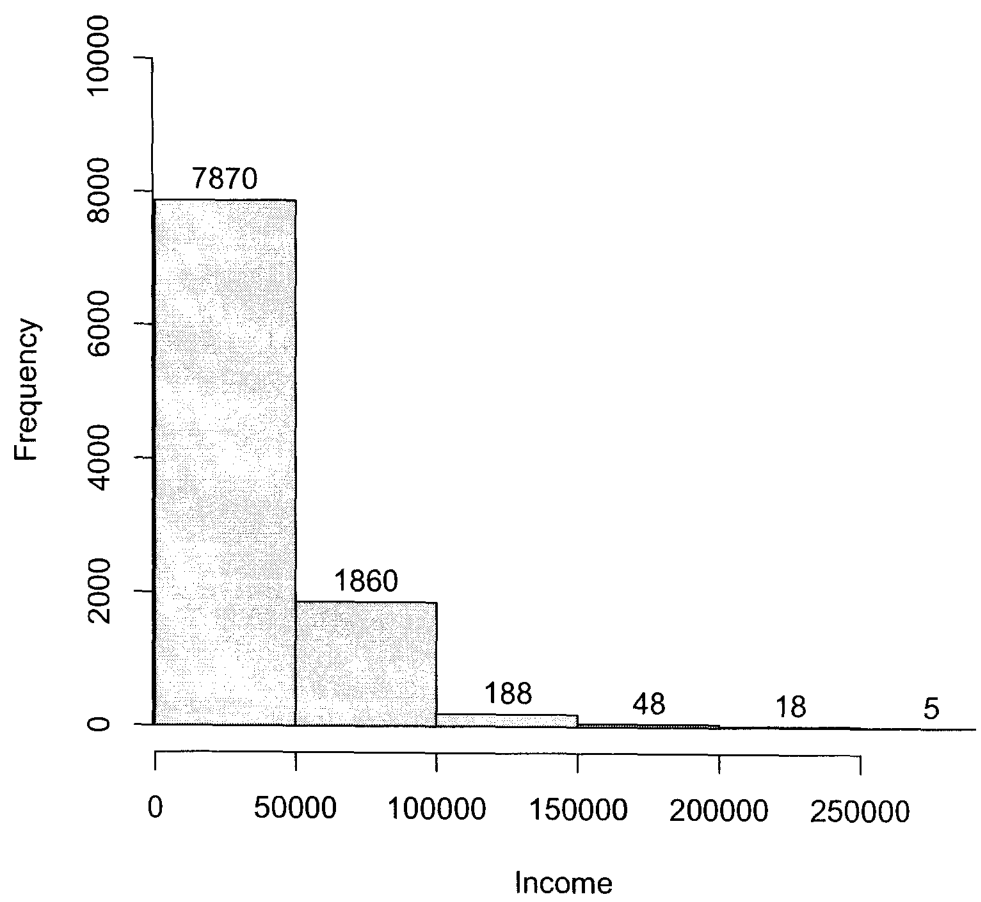

Figure 4.1 Income distribution over 1,000 replications for a sample size of 10,000 from the Pareto distribution

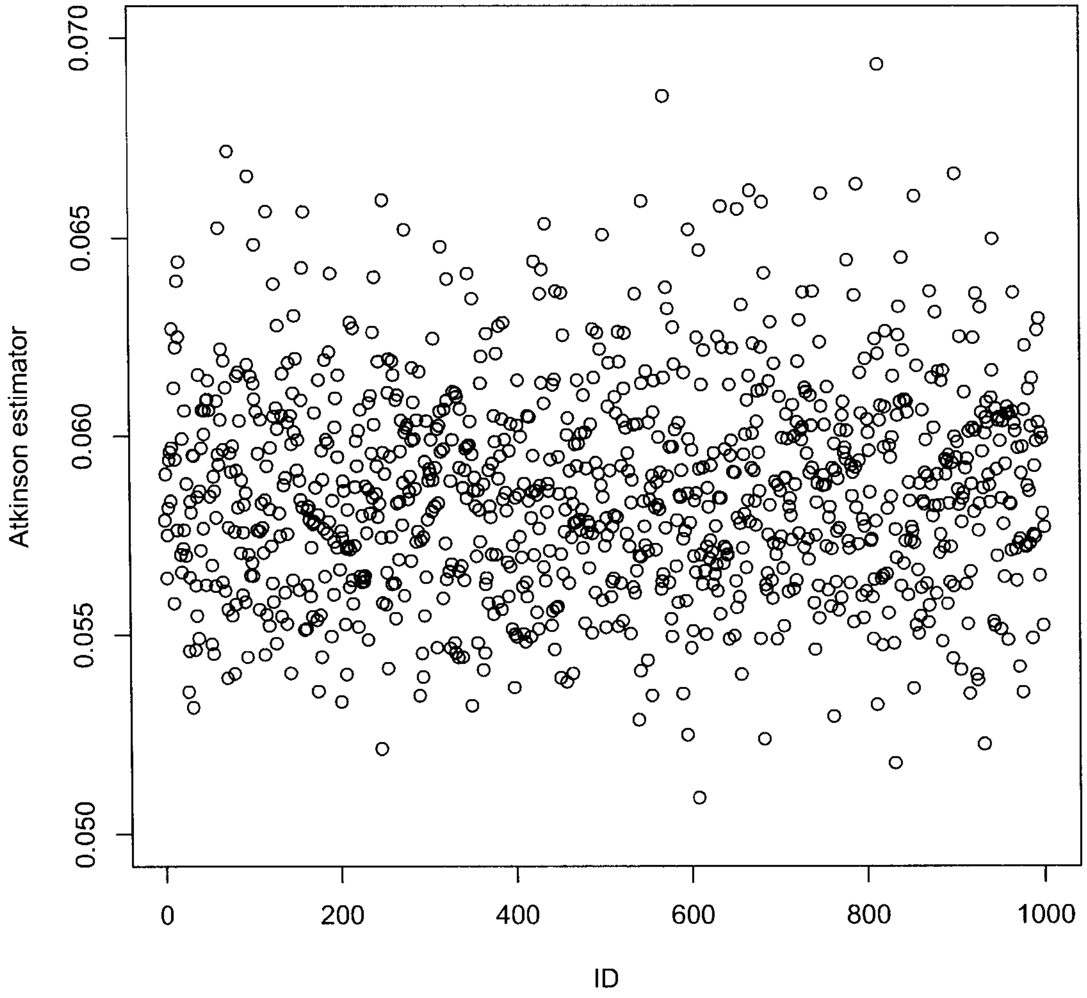

Figure 4.2 Scatter plot of nonparametric Atkinson estimates for sample size 10,000 from the Pareto distribution

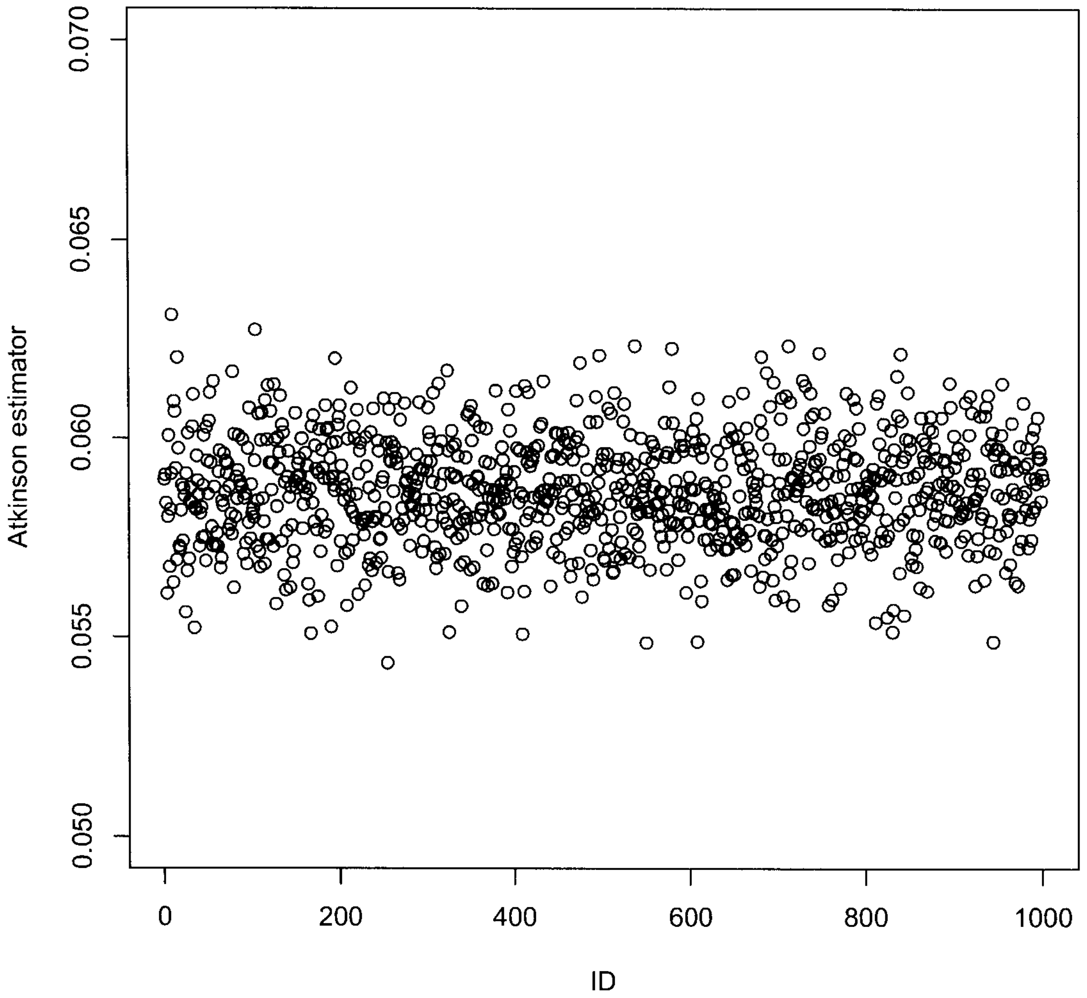

Figure 4.3 Scatter plot of the parametric Atkinson estimates for sample size 10,000 from the Pareto distribution

It is seen from the Figs. (4.2) and (4.3) that the spread of variance is considerably smaller for the parametric than for the nonparametric estimator. This result implies that the parametric estimator of the Atkinson index is more efficient than the nonparametric estimator. The empirical mean of the nonparametric estimates is 0.05856107, and the parametric mean is 0.0586206. The two means are close to the theoretical Atkinson index $A_{F,a} = 0.05870694$, implying that the parametric as well as the nonparametric estimates are unbiased.

The plot in Fig. (4.4) represents a histogram of the 1,000 nonparametric variance estimates.

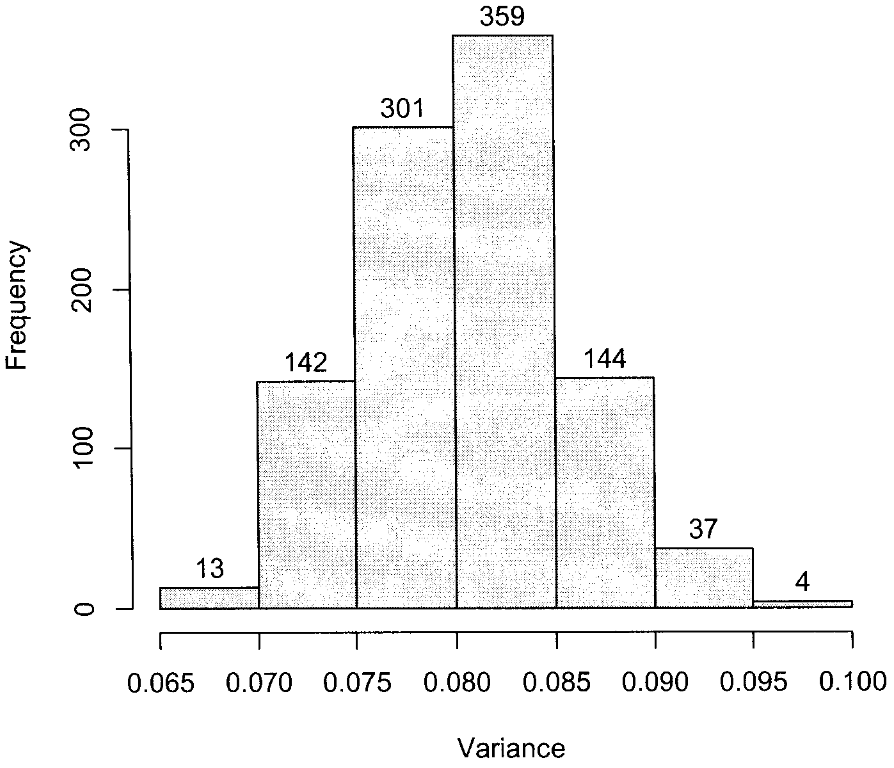

Figure 4.4 Histogram of nonparametric variance estimates for sample size 10,000 from the Pareto distribution

The plot in Fig. (4.5) represents a histogram of the 1,000 parametric variance estimates.

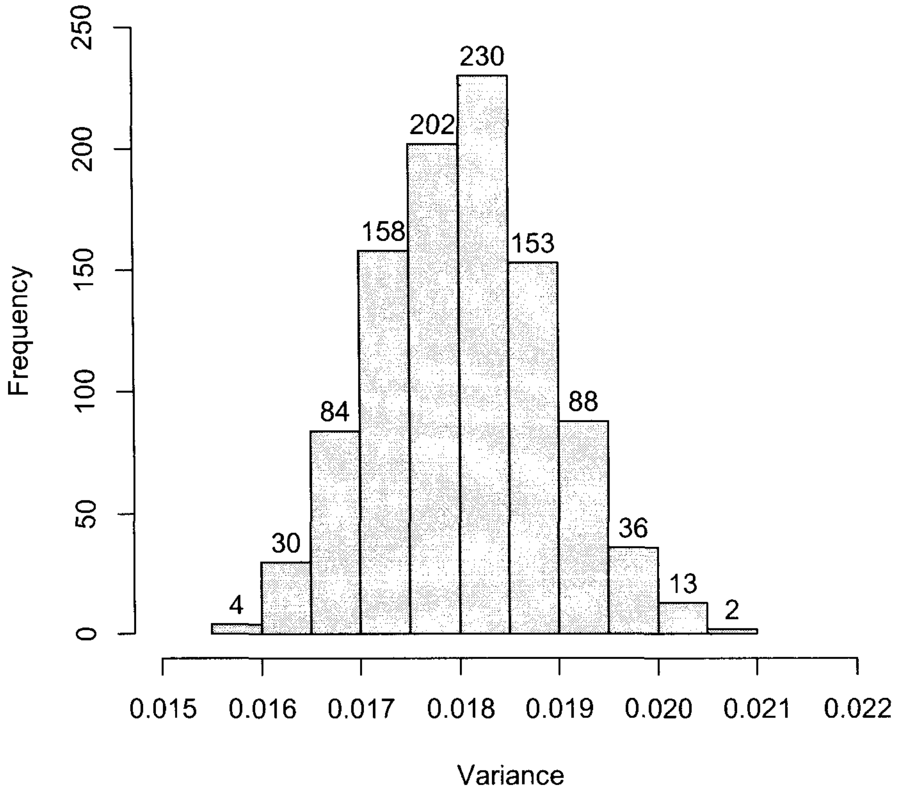

Figure 4.5 Histogram of parametric variance estimates for sample size 10,000 from the Pareto distribution

Both distributions are alike. The means over 1,000 replications for the nonparametric and parametric estimates is 0.0805688 and 0.018037, respectively. Compared to the nonparametric variance $\sigma_n^2 = 0.0807779$ and parametric variance $\sigma_F^2 = 0.0180823$, both nonparametric and parametric estimates are unbiased.

4.2 Exponential Distribution

In the Exponential case, we have simulated 1,000 samples of size $n = 1{,}000$ following an Exponential distribution with parameter $\theta = 2$.

For each sample, the theoretical Atkinson index and its nonparametric estimates were calculated.

As we have seen in Eq. (3.12), the Atkinson index does not depend on the parameter of the distribution. Therefore, one cannot calculate the parametric asymptotic variance and its estimates. For this situation, a bootstrap method should be used, since it does not require the practitioner to know the form of the variance.

The mean of the nonparametric Atkinson estimates is 0.3468753, which is very close to the theoretical Atkinson index of $A_F = 0.347459$, implying that the nonparametric estimates are unbiased.

A scatter plot of the 1,000 nonparametric Atkinson estimates is presented in Fig. (4.6).

Figure (4.7) represents a histogram of the nonparametric Atkinson estimates, for 1,000 incomes that follow an Exponential distribution.

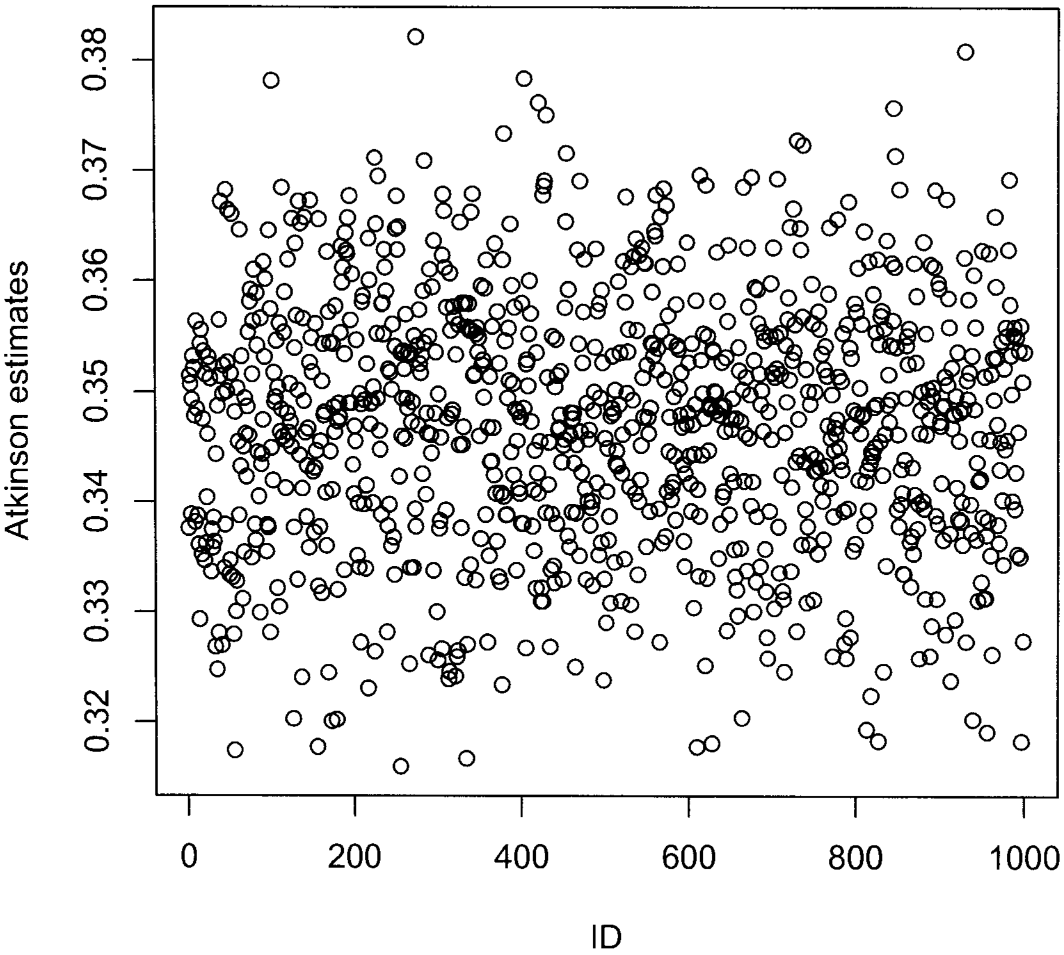

Figure 4.6 Scatter plot of the Atkinson index estimates for sample size 1,000 in the exponential case

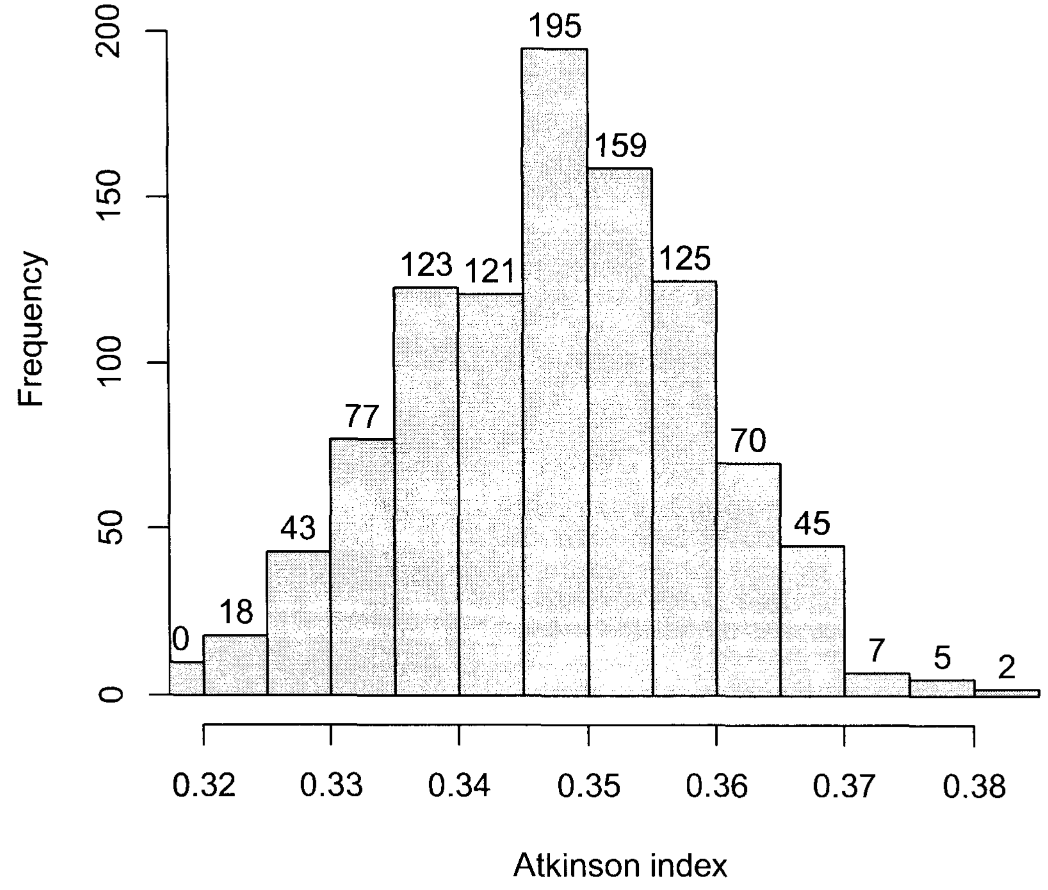

Figure 4.7 Histogram of the Atkinson index estimates for sample size 1,000 in the exponential case

4.3 Lognormal Distribution

In what follows we present the simulations for a data set following a Lognormal distribution with mean $\mu = 0.5$ variance $\sigma^2 = 1$ and location parameter $x_0 = 0.01$. The Atkinson parameter considered is $a = 0.2$.

We generated 1,000 samples of different sizes, $n = 100$, $n = 500$, $n = 1{,}000$ and $n = 10{,}000$ and calculated the Atkinson index for each sample in order to compare the empirical 95% confidence interval or coverage probability to the asymptotic confidence interval for the Atkinson index.

Results for the nonparametric case, for the four sample sizes mentioned above are summarized in the Table (4.2).

Table 4.2 Coverage probability of the Atkinson index from simulation using the Lognormal distribution with 1,000 replicates

Sample size	Proportion Nonparametric
n=100	0.898
n=500	0.916
n=1,000	0.927
n=10,000	0.943

One can conclude that as the sample size increases, the coverage $(1-\alpha)$ converges to 0.95 or the type I error (α) converges to 0.05. However, the convergence is slower for the Lognormal distribution than it is for the the Pareto distribution.

In practice, some populations, such as salaries at institutions, may be relatively small, in which case results for $n = 100$ or $n = 500$ would be relevant. In such a case, results in Table 4.2 show the $1-\alpha$ level to be somewhat less than 0.95 for the nonparametric case, but close to 0.95 in the parametric case. These derivations from expected $(1-\alpha = 0.95)$ are deemed acceptable for applications purposes.

The random variables generated represent incomes at a scale of 1:20,000. Therefore, the minimum income in the income vector was Income_{min} = \$1,212.68 and the maximum income Income_{max} = \$965,479.93 with the mean income being \$ 34,256.56.

The theoretical Atkinson index is $A_{F,a} = 0.3019898$, the vector $Atkinsonn[k]$ represents the nonparametric Atkinson estimates for $m = 1,000$ replicates, and the vector $Atkinsonpe[k]$ represents the parametric Atkinson estimates the m = 1,000 replicates. The parametric variance is $\sigma_F^2 = 0.007118831$, and the vector $\overline{\sigma}_F^2$ represents the parametric estimates for m = 10,000 replicates.

The plot in Fig. (4.8) represents a histogram of the incomes for sample size 1,000 from the Lognormal distribution. It is seen from the histograms that about 87% of the incomes are around the mean income.

Figure (4.9) represents a plot of the 1,000 nonparametric Atkinson indices for n = 1,000 and Fig. (4.10) represents a plot of the 1,000 parametric Atkinson indices for n = 1,000. It is seen from these two figures that the spread of variance is considerably smaller for the parametric than for the nonparametric estimator. This result implies that parametric estimator of the Atkinson index is more efficient than the nonparametric estimator. The empirical mean of the nonparametric estimates is 0.3257333, the parametric mean is 0.3019916, and the theoretical Atkinson index is 0.3019898. The two means are close to the theoretical Atkinson index, implying that the parametric as well as the nonparametric estimators are close to being unbiased.

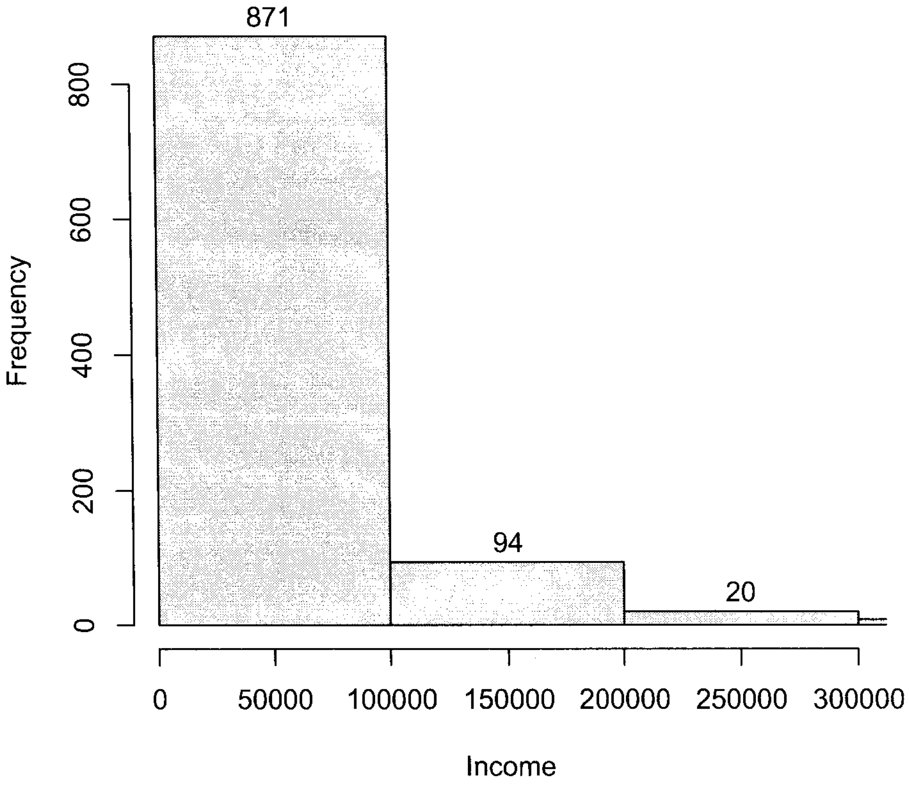

Figure 4.8 Income distribution over 1,000 replications for a sample size of 1,000 from the Lognormal distribution

The plot in Fig. (4.11) represents a histogram of the 1,000 parametric variance estimates. The mean of the parametric estimates is 0.00715778. The parametric variance $\sigma_F^2 = 0.007118831$ implies that the parametric estimates are unbiased.

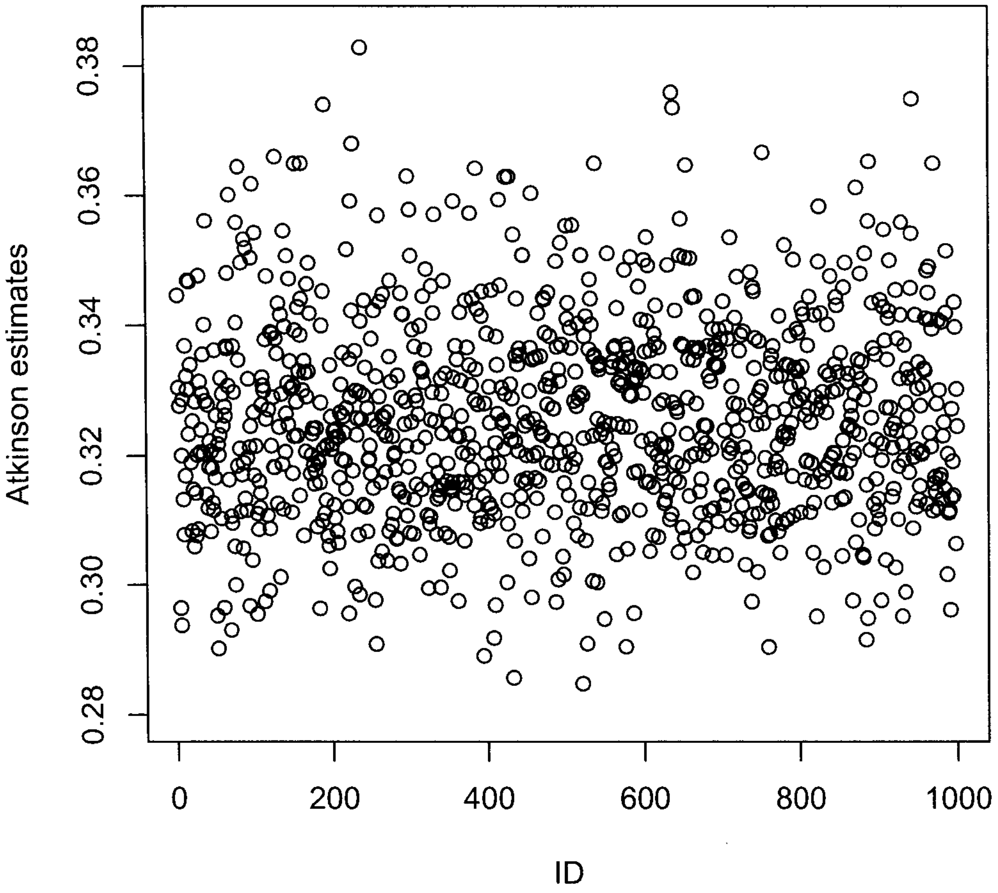

Figure 4.9 Scatter plot of nonparametric Atkinson index estimates for sample size 1,000 from the Lognormal distribution

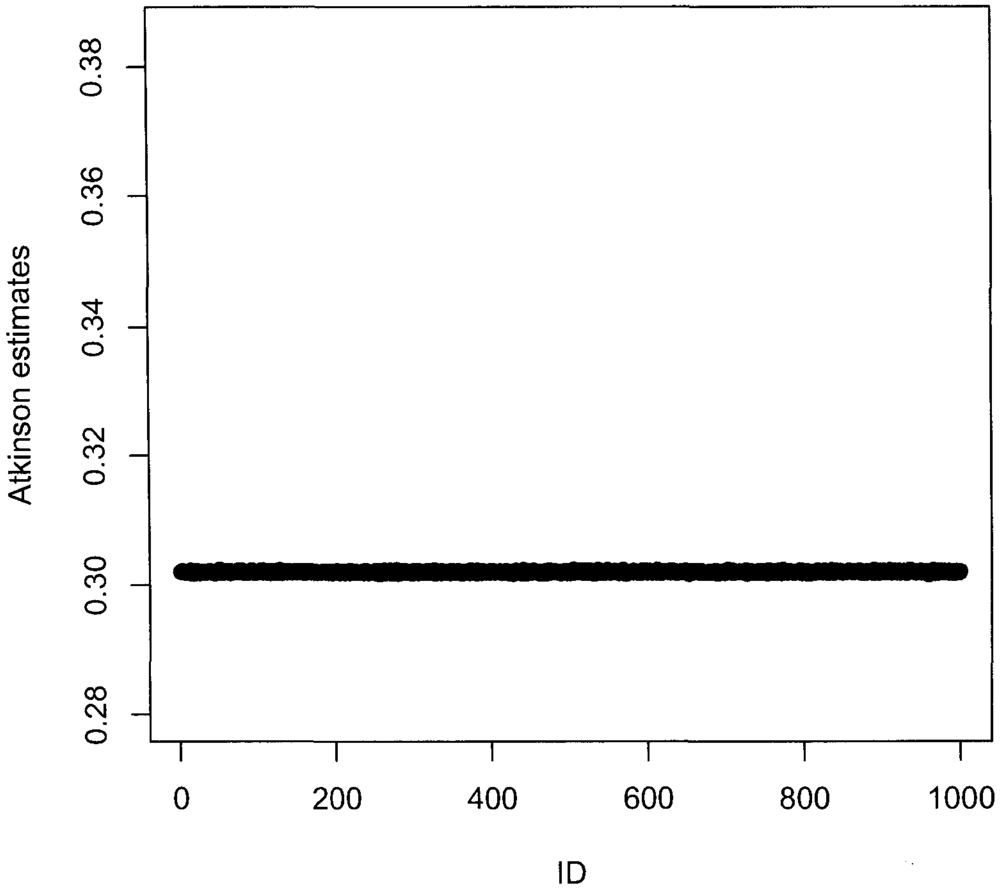

Figure 4.10 Scatter plot of parametric Atkinson index estimates for sample size 1,000 from the Lognormal distribution

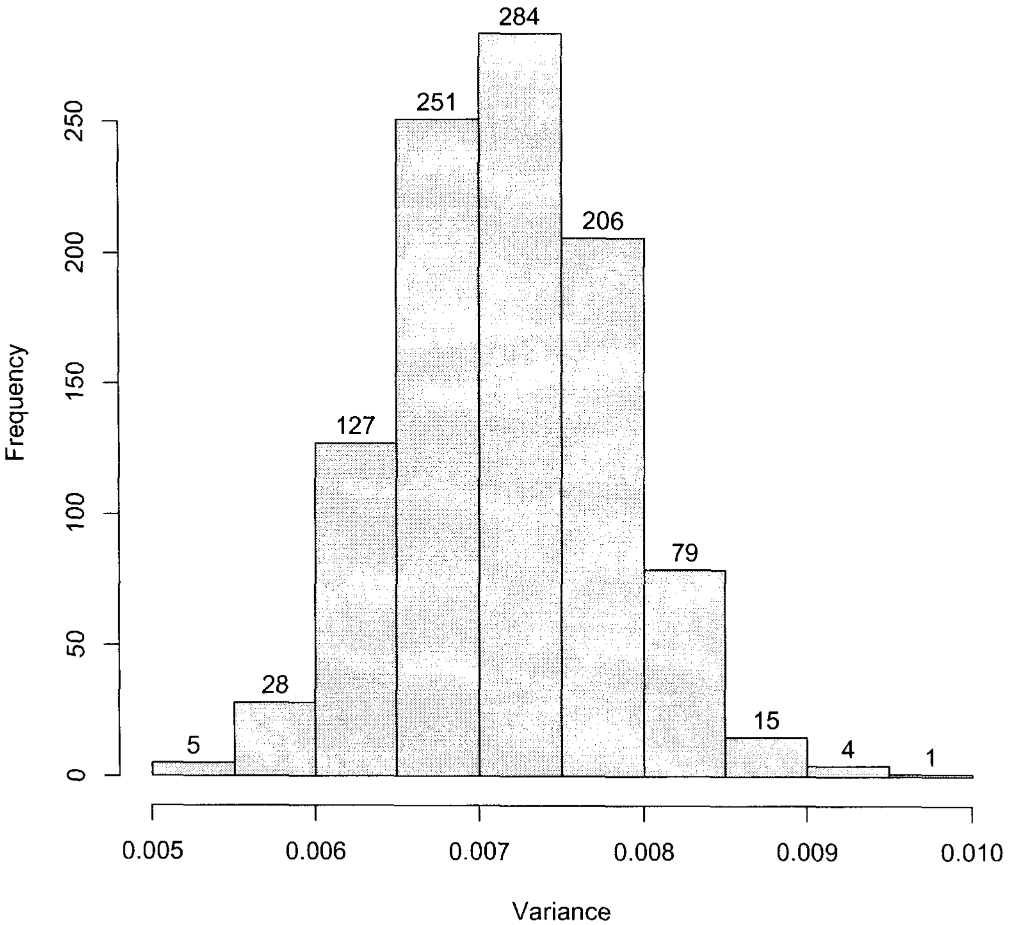

Figure 4.11 Histogram of the parametric variance estimates for sample size 1,000 from the Lognormal distribution

4.4 Bootstrap Method

In this section, we generated 1,000 Pareto samples of different sizes, $n = 100, n = 500, n =$ 1,000 and $n =$ 10,000. We then calculated the Atkinson index for each sample in order to compare the empirical estimators and the 95% confidence intervals, or coverage probabilities, to the theoretical Atkinson index and the bootstrap confidence interval for the Atkinson index, respectively. In order to generate the bootstrap Atkinson index, we sampled with replacement from the original sample, and for each sample the bootstrap Atkinson index was calculated.

The parameter of the Pareto distribution is $\lambda = 3$ and the Atkinson parameter is $a = 0.2$. Of interest is calculating 95% bootstrap confidence interval or coverage probability for the Atkinson index. In this case the expected α is 0.05. The algorithm used for calculating the bootstrap confidence intervals is described in Subsections 3.3.1 and 3.3.2.

The coverage probability was replicated ten times in order to determine the magnitude of fluctuations in the simulation results.

Results for the nonparametric case with $m =$ 1,000, $k = 10$ gave the following probabilities summarized in Table 4.3.

Results for the parametric case with $m =$ 1,000, $k = 10$ gave the following probabilities summarized in Table 4.4.

In Table 4.5 we summarize the results of the mean of the 10 probabilities presented in Table 4.3 and Table 4.4 for the nonparametric and parametric case. We have taken $\alpha = 0.05$, and we observe that as n increases, the values converge to $1 - \alpha = 0.95$ as expected.

Table 4.3 Pareto nonparametric bootstrap coverage probabilities

Sample size	Proportion Nonparametric
n=100	0.953 0.941 0.962 0.947 0.953 0.944 0.951 0.946 0.945 0.932
n=500	0.956 0.937 0.956 0.956 0.942 0.944 0.942 0.947 0.956 0.937
n=1,000	0.937 0.957 0.939 0.961 0.934 0.950 0.933 0.937 0.957 0.970
n=10,000	0.951 0.930 0.954 0.952 0.953 0.947 0.948 0.961 0.951 0.945

Table 4.4 Pareto parametric bootstrap coverage probabilities

Sample size	Proportion Parametric
n=100	0.963 0.960 0.949 0.951 0.948 0.963 0.965 0.971 0.963 0.949
n=500	0.950 0.961 0.956 0.934 0.963 0.966 0.962 0.951 0.945 0.940
n=1,000	0.942 0.932 0.948 0.949 0.941 0.942 0.970 0.956 0.943 0.949
n=10,000	0.950 0.925 0.950 0.958 0.953 0.968 0.949 0.957 0.952 0.963

One can conclude from Table 4.5 that as the sample size increases, the coverage $(1-\alpha)$ converges to 0.95 or the type I error (α) converges to 0.05 for the nonparametric case. In the parametric case the convergence is about 0.95 for all the sample sizes considered.

In practice, some populations, such as salaries at institutions, may be relatively small, in which case results for $n = 100$ or $n = 500$ would be relevant. In such a case, results in Table 4.5 show that the bootstrap method gives better convergence than the asymptotic method, as shown in Table 4.1.

Table 4.5 Coverage probability of the bootstrap Atkinson index from simulation using the Pareto distribution with 1,000 replicates

Sample size	Proportion Nonparametric	Proportion Parametric
n=100	0.9474	0.9582
n=500	0.9473	0.9528
n=1,000	0.9475	0.9472
n=10,000	0.9492	0.9525

The theoretical Atkinson index is $A_{F,a} = 0.05870694$. The vector $Atkinsonn[k]$ represents the nonparametric Atkinson estimates for $m = 1{,}000$ replicates and has the mean 0.05877483. The vector $Atkinsonnboot[k]$ represents the nonparametric bootstrap Atkinson estimates for $m = 1{,}000$ replicates and has the mean 0.05876312. The vector $Atkinsonpe[k]$ represents the parametric Atkinson estimates for m =1,000 replicates and has the mean 0.05869812 and the vector $Atkinsonpeboot[k]$ represents the parametric bootstrap Atkinson estimates for $m = 1{,}000$ replicates and has the mean 0.05868551.

Note 4.4.1 *For the nonparametric case, the bootstrap estimates have the mean closer to the theoretical Atkinson index than the asymptotic estimates have. For the parametric case, the asymptotic estimates have the mean closer to the theoretical Atkinson index than the bootstrap estimates.*

From the scatter plots in Figs. (4.12) – (4.15), it is seen that the bootstrap estimates have a larger variance than the asymptotic estimates. Also, the nonparametric estimates have a smaller variance than the parametric estimates.

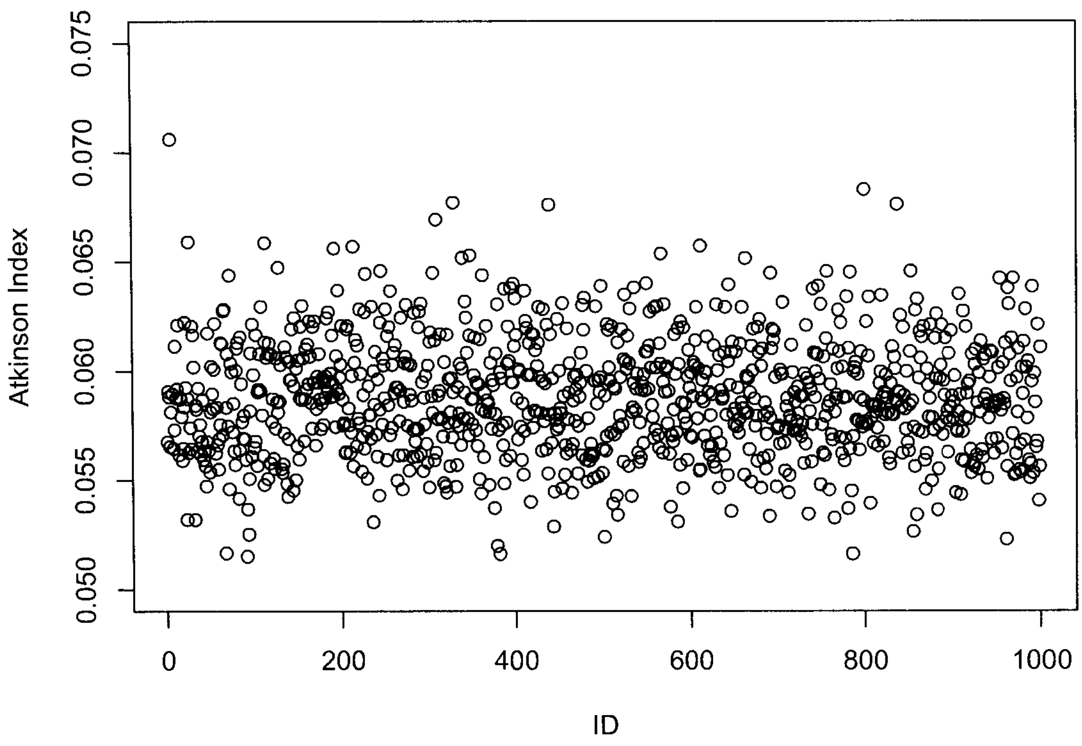

Figure 4.12 Scatter plot of the nonparametric asymptotic Atkinson estimates for sample size 10,000

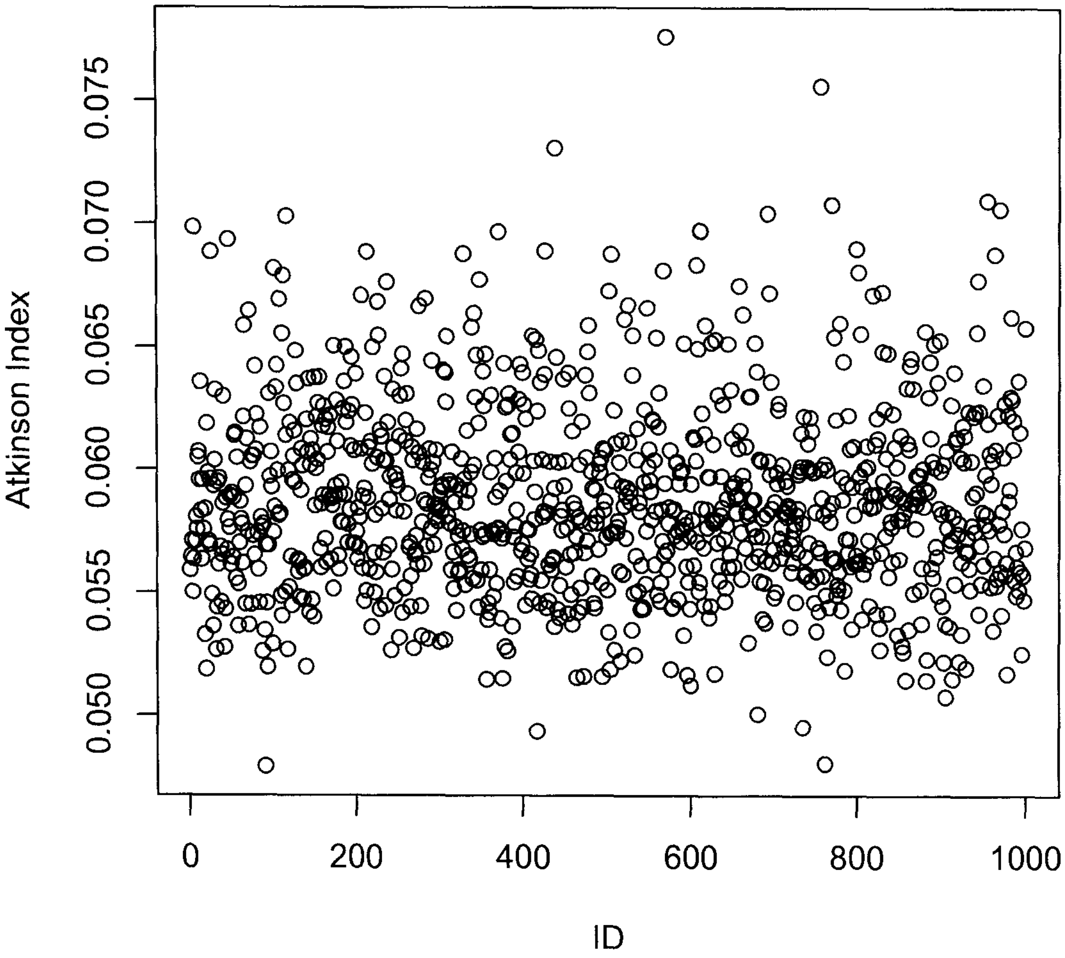

Figure 4.13 Scatter plot of the nonparametric bootstrap Atkinson estimates for sample size 10,000

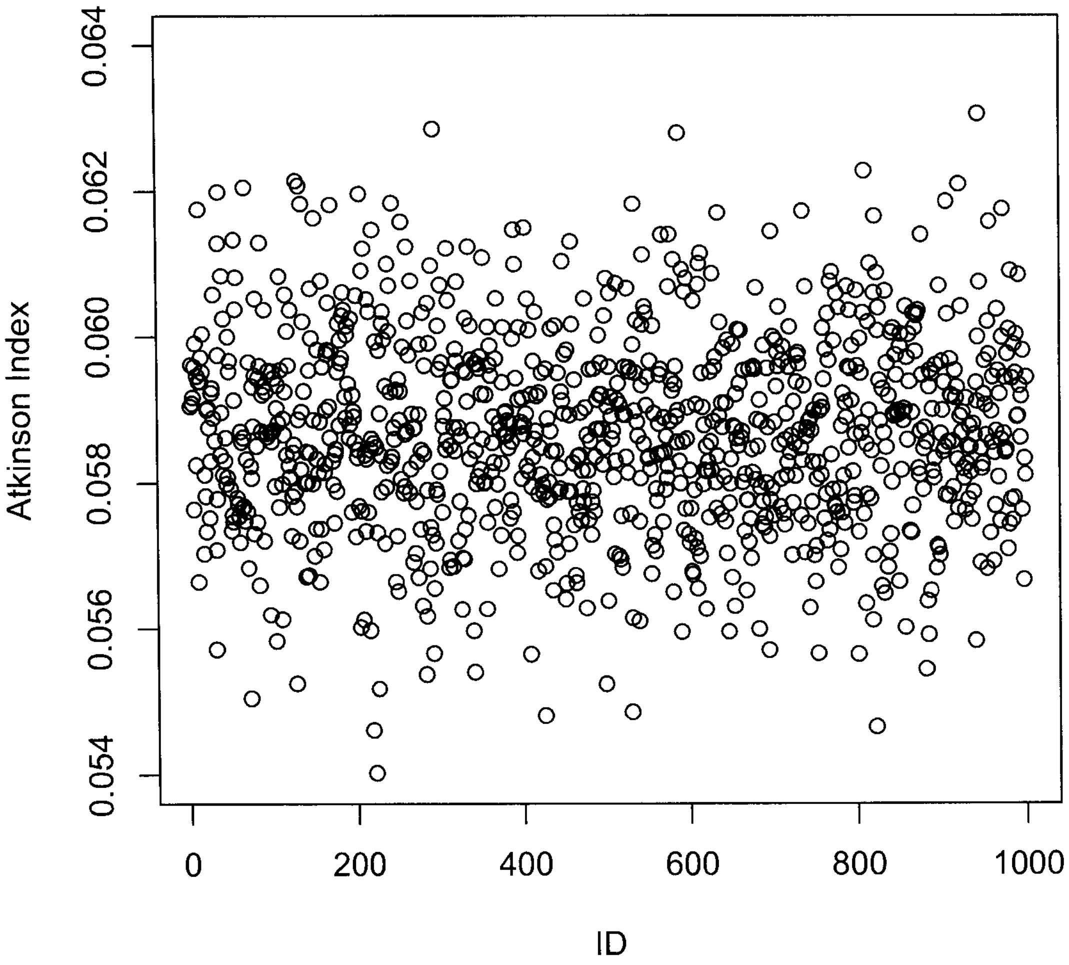

Figure 4.14 Scatter plot of the parametric asymptotic Atkinson estimates for sample size 10,000

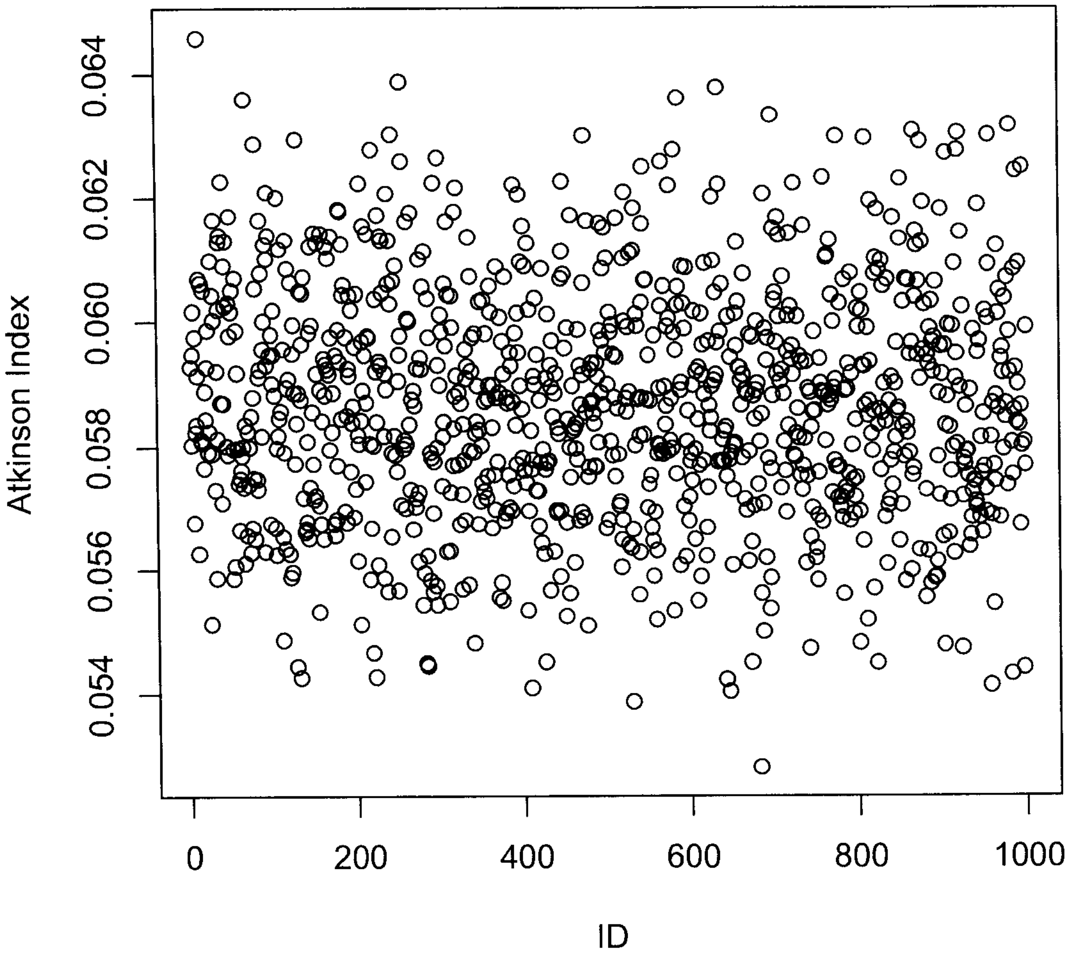

Figure 4.15 Scatter plot of the parametric bootstrap Atkinson estimates for sample size 10,000

In the Exponential case, we simulated 1,000 samples of size n = 1,000, following an Exponential distribution with parameter $\theta = 2$. For each sample, the parametric Atkinson index was calculated. As we have seen in Eq. (3.12), the Atkinson index does not depend on the parameter of the distribution.

We are calculating the coverage probability for the nonparametric case and we replicate the calculation of the coverage probability $k = 10$ times in order to see how the results fluctuate in our simulation.

Results for the nonparametric case with m = 1,000, $k = 10$ gave the following probabilities summarized in Table 4.6.

Table 4.6 Exponential nonparametric bootstrap coverage probabilities

Sample size	Proportion Nonparametric
n=100	0.938 0.939 0.966 0.952 0.944 0.941 0.966 0.961 0.939 0.921
n=500	0.938 0.962 0.945 0.954 0.940 0.946 0.937 0.943 0.943 0.958
n=1,000	0.935 0.958 0.956 0.935 0.957 0.942 0.966 0.956 0.948 0.937
n=10,000	0.961 0.946 0.945 0.959 0.946 0.928 0.963 0.956 0.947 0.953

In Table 4.7 we summarize the results of the mean of the 10 probabilities presented in Table 4.6 for the nonparametric. We have taken $\alpha = 0.05$, and we observe that as n increases, the values converge to $1 - \alpha = 0.95$ as expected. The theoretical Atkinson index is $A_{F,a} = 0.3474519$. The vector $Atkinsonn[k]$ represents the nonparametric Atkinson estimates for m = 1,000 replicates and has the mean 0.3473711. The vector $Atkinsonnboot[k]$ represents the nonparametric bootstrap Atkinson estimates for m =

Table 4.7 Coverage probability of the Atkinson index from simulation using the Exponential distribution with 1,000 replicates

Sample size	Proportion Nonparametric
n=100	0.9467
n=500	0.9466
n=1,000	0.9490
n=10,000	0.9504

1,000 replicates and has the mean 0.3471849.

From the scatter plots in Figs. (4.16) and (4.17) it is seen that the nonparametric bootstrap estimator has a larger spread (less efficient) than the asymptotic nonparametric estimator. The mean of the asymptotic estimators is closest to the theoretical Atkinson index, than the mean of the bootstrap estimators.

Note 4.4.2 *We have seen in the previous chapter that for some distributions, like the Exponential or Lognormal it is very hard to see how the variance looks like or even to estimate it. In cases like this, the bootstrap method is more useful since one does not need to estimate the variances in order to construct confidence intervals and study the coverage probabilities.*

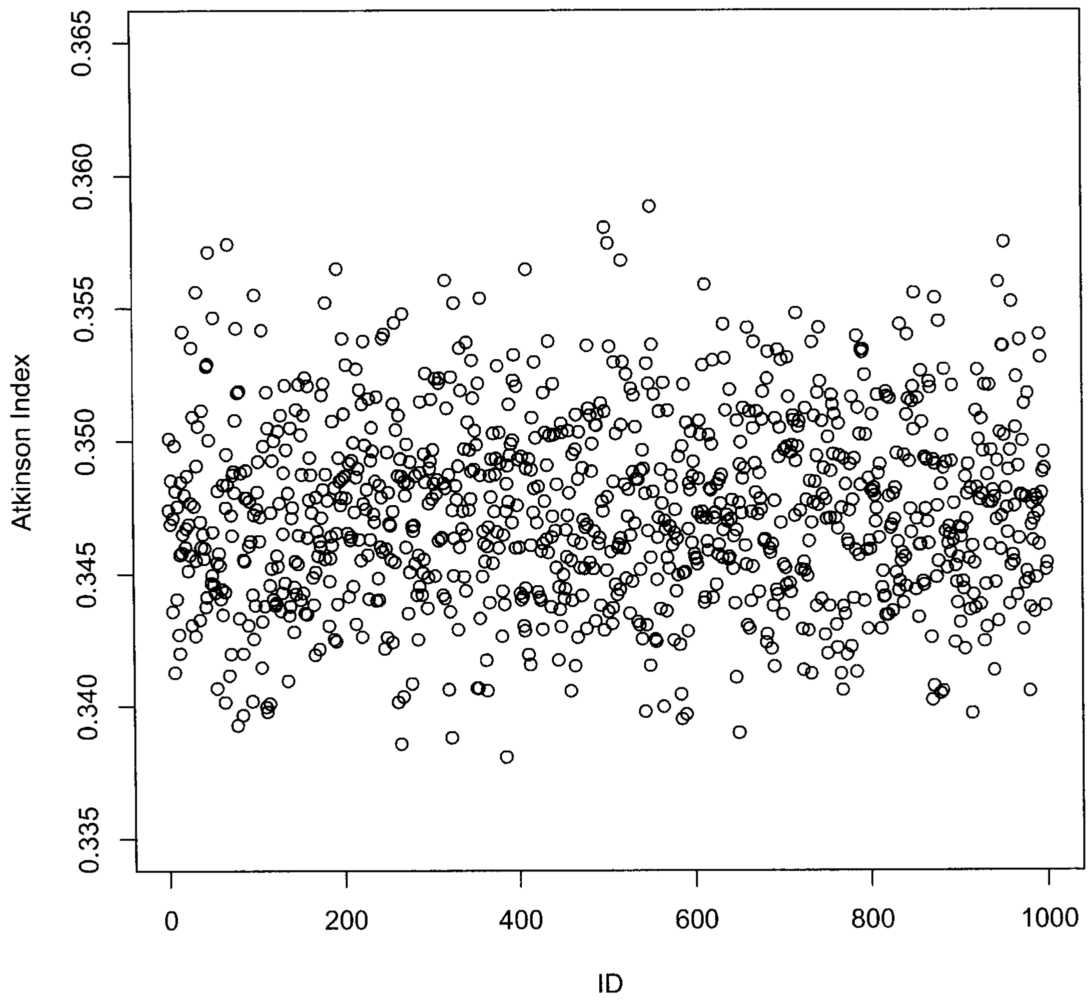

Figure 4.16 Scatter plot of the nonparametric Atkinson estimates for sample size 10,000

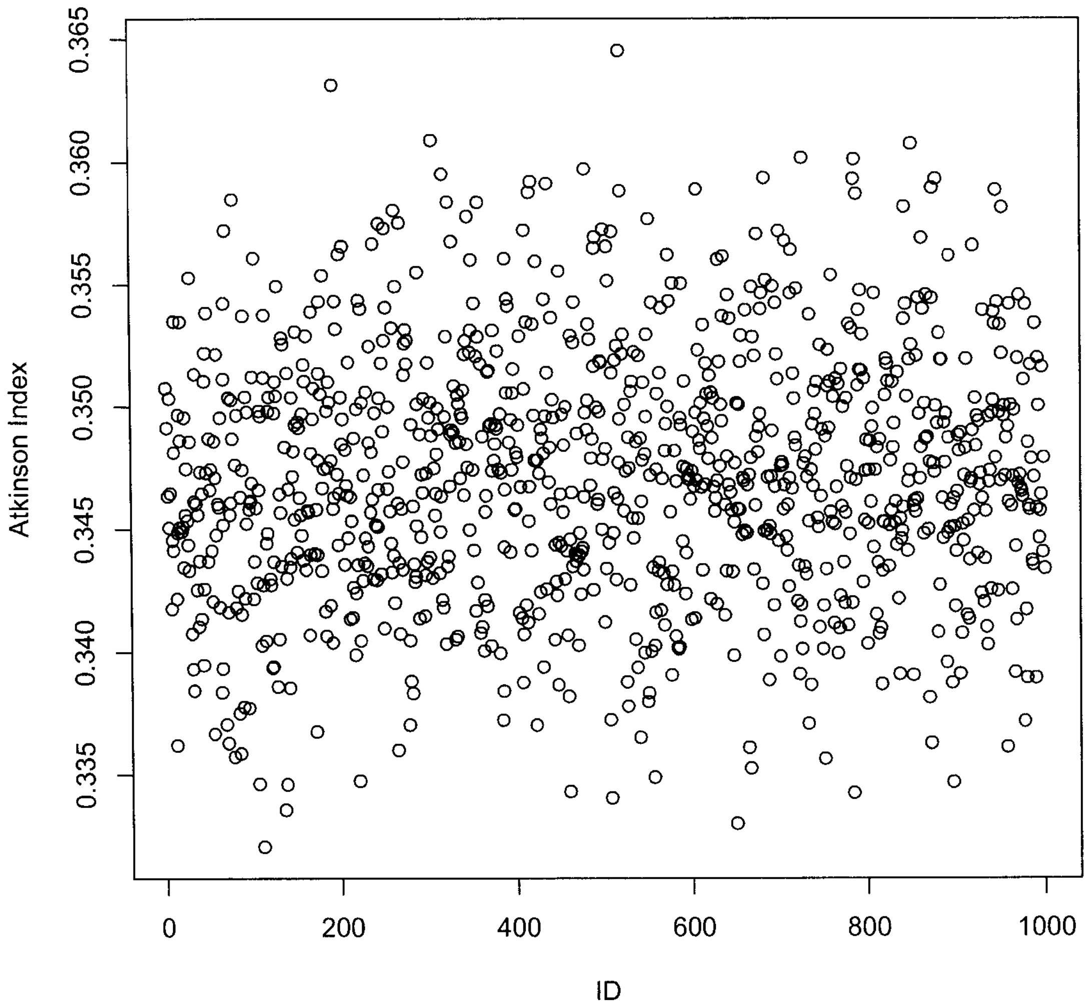

Figure 4.17 Scatter plot of the nonparametric bootstrap Atkinson estimates for sample size 10,000

4.5 Case Study

In this case study, we present incomes from the Louisiana Tech University, for the year 2005-2006. We wish to calculate the two Atkinson indices and see how the salaries differ between 12 months group and the 9 months group. An Atkinson index close to zero implies that there is little inequality in incomes. When the Atkinson index is zero, one has a state of perfect equality. When the Atkinson index is one, one has perfect inequality.

4.5.1 12 Months Incomes

The first part of the study deals with a sample of $n = 558$ salaries for the 12 months employees. A plot of these salaries is shown in Fig. (4.18). The minimum income in the sample is $1,252 per year, the maximum is $200,020 per year, and the mean salary is mean $35,283.60 per year.

To observe better how the incomes are spread in our data set, we present a histogram of the incomes in Fig. (4.19).

It is seen that 372 employees earn less than the mean in a year and 186 employees earn more than the mean. The income extremes and the quartiles are as follows:

```
     0%       25%       50%       75%      100%
1252.00  19474.75  27458.00  40875.00 200020.00
```

The annual total income for the 12 months employees is $19,688,247 and the top 25% of the incomes has an annual total of $10,091,524. Therefore, the top 25% of the incomes represent more than 50% of the total annual income.

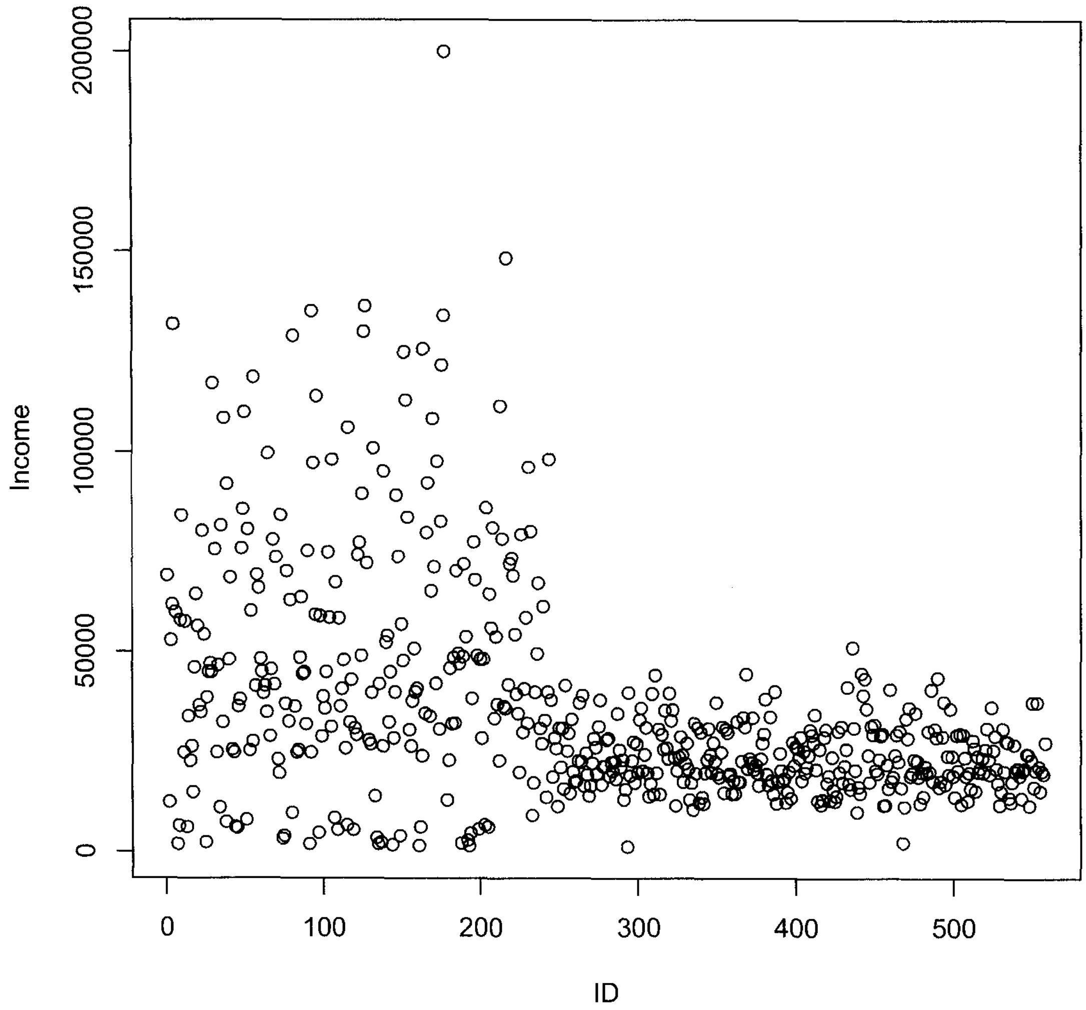

Figure 4.18 Plot of the yearly incomes for the 12 months employees at Louisiana Tech University

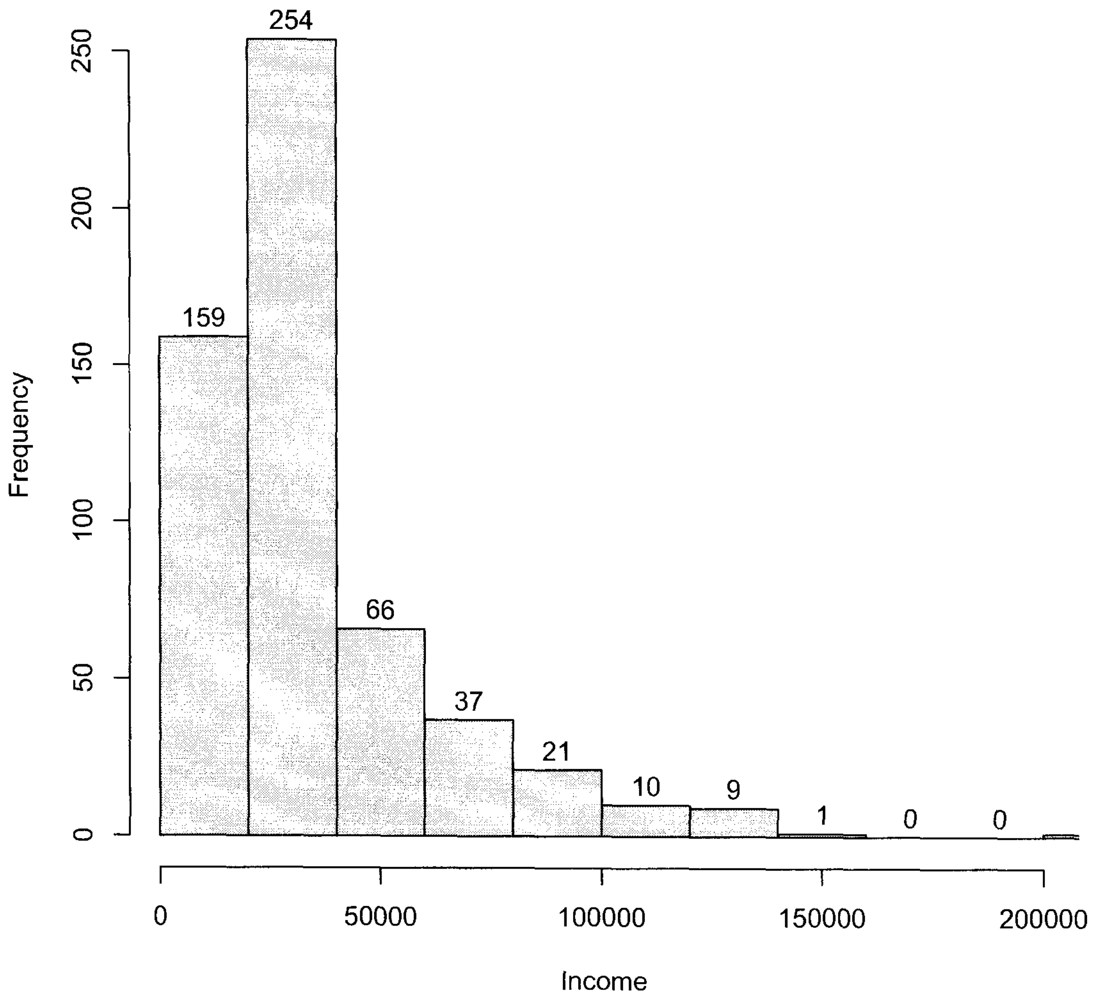

Figure 4.19 Histogram of the yearly incomes for the 12 months employees at Louisiana Tech University

Based on these incomes we assumed that the population is risk averse and calculated the empirical Atkinson index for $a = 0.1$. This was found to be 0.1952934.

The plot in Fig. (4.20) represents the plot of the empirical distribution function for the income.

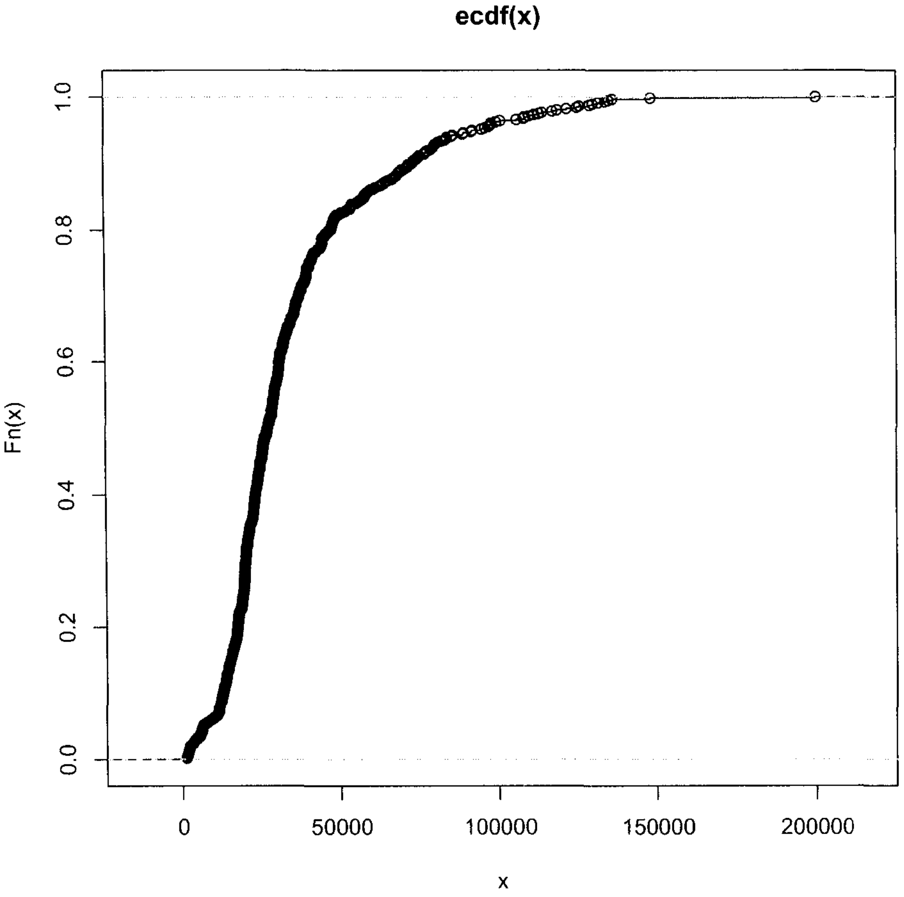

Figure 4.20 Plot of the empirical distribution function for income of 12 months employees at Louisiana Tech University

Based on this plot we conclude that the data fits a Lognormal distribution and using the maximum likelihood estimates, we determined the parameters of the distribution. In this case the mean μ, variance σ^2 and location parameter x_0 for the fitted Lognormal are 10.21529, 0.5787538 and 0.001, respectively. The plot of the fitted distribution function is given in Fig. (4.21). Based on this Lognormal parameters, the Atkinson index was calculated to be 0.2061501 in agreement with the empirical index 0.1952934 obtained from the sample.

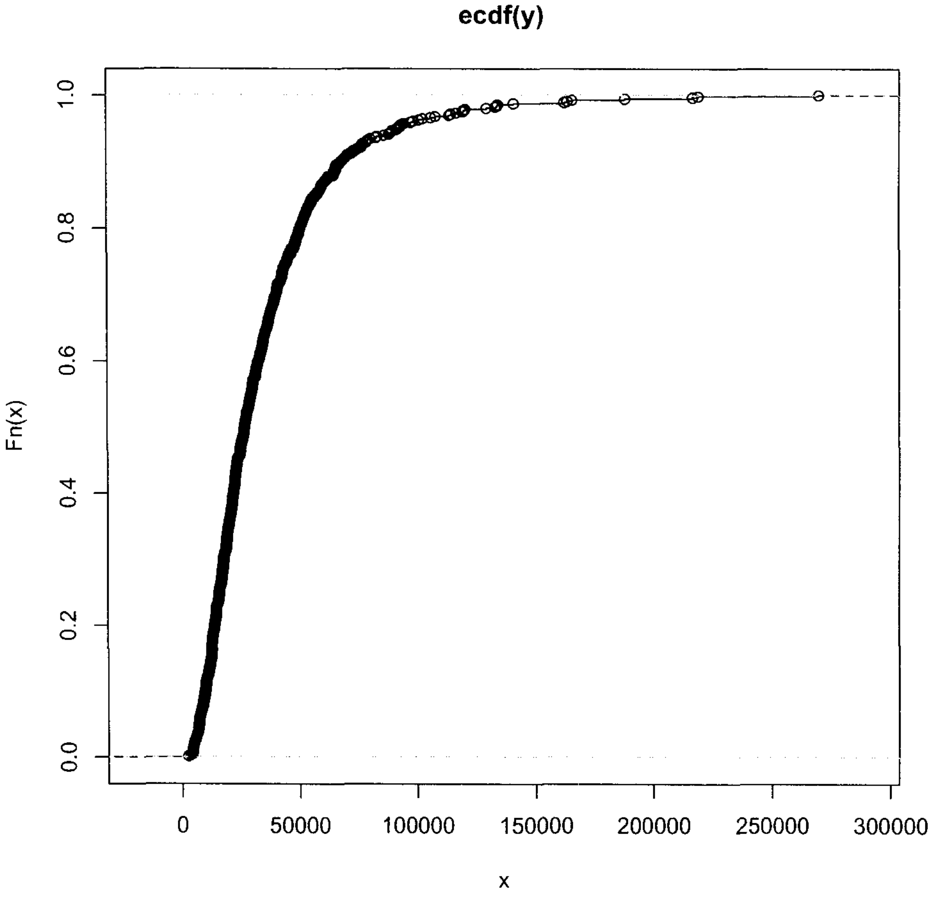

Figure 4.21 Plot of the Lognormal distribution function for μ = 10.21529, σ = 0.7607587 and $x_0 = 0.001$

4.5.2 9 Months Incomes

The second part of the study deals with a sample of n = 336 salaries for the 9 months employees. The minimum income in the sample is \$4,990 per year, the maximum is \$129,289 per year, and the mean salary is \$51,452.94 per year.

Figures (4.22) and (4.23) are plots of the yearly incomes of the 9 months employees in the sample.

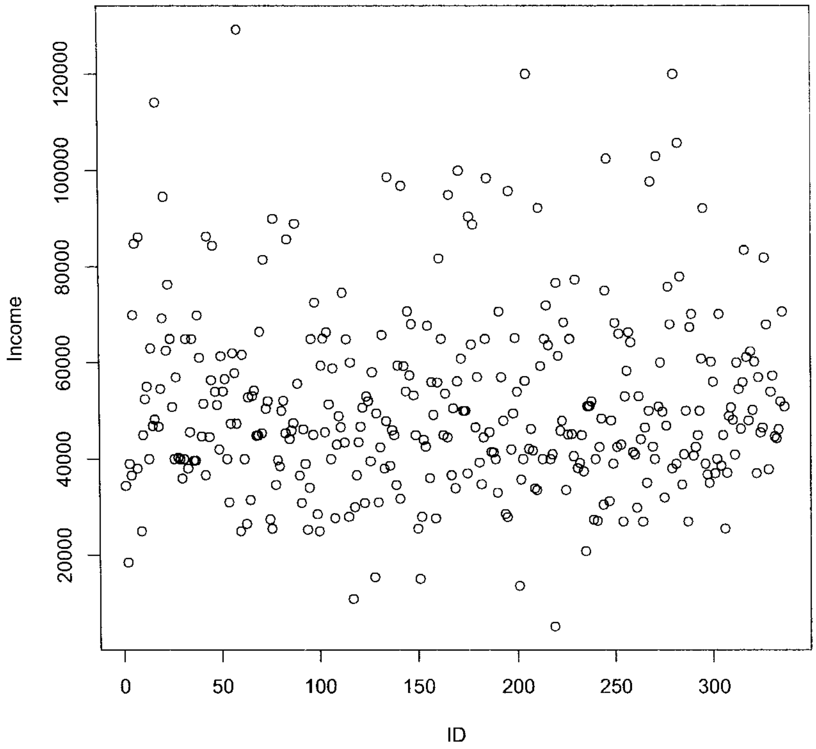

Figure 4.22 Plot of yearly incomes for the 9 months employees at Louisiana Tech University

To observe better how the incomes are spread in our data set, we present a histogram of the incomes.

It is seen from Fig. (4.23) that 135 employees earn less than the mean and 201 employees earn more than the mean. The income extremes and the quartiles are as follows:

0%	25%	50%	75%	100%
4990	39703	47419	60363	129288

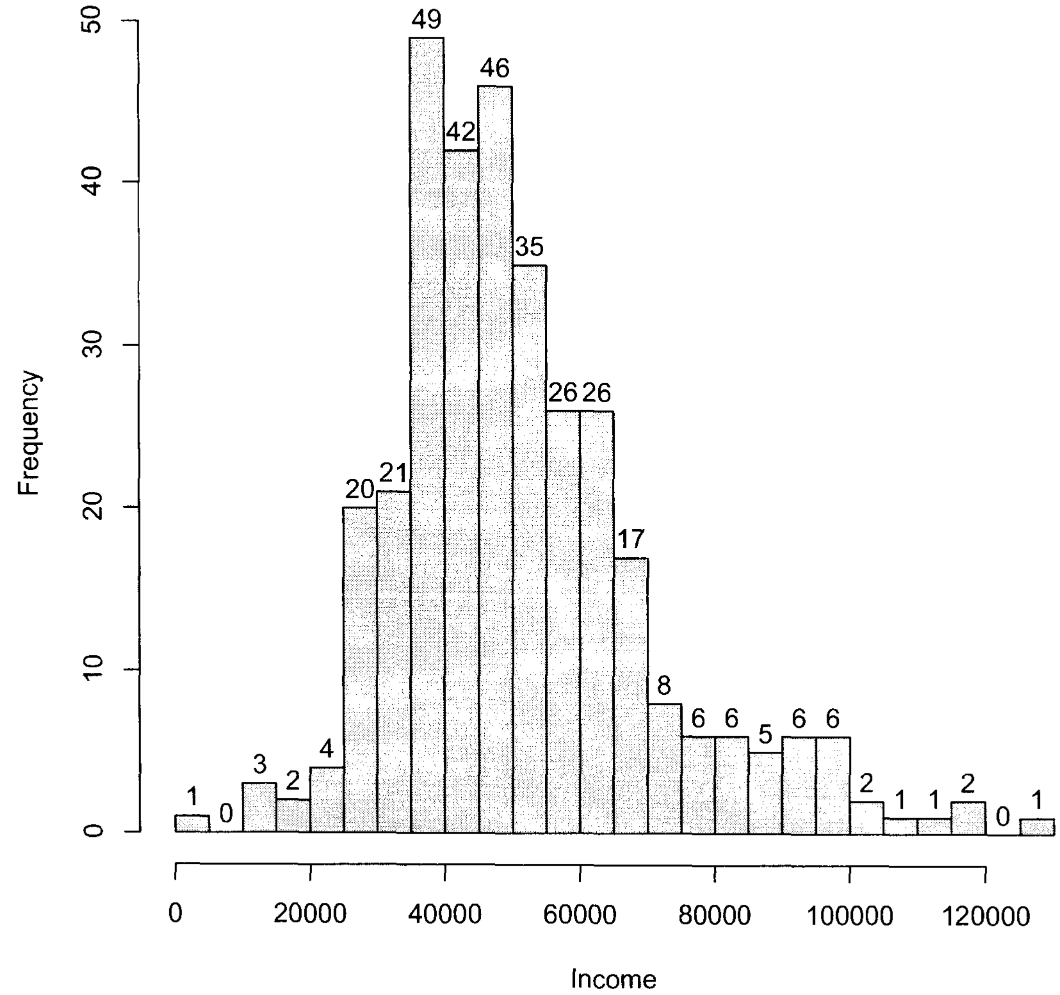

Figure 4.23 Histogram of yearly incomes for the 9 months employees at Louisiana Tech University

The annual total income for the 9 months employees is \$17,288,187 and the top 25% of the incomes has an annual total of \$6,504,277. Therefore, the top 25% of the incomes represent only about 38% of the annual income. We calculate the empirical Atkinson index for $a = 0.1$, which assumes that the population is risk averse, and obtained a value of 0.06714808.

The plot in Fig. (4.24) is the plot of the empirical distribution function for our incomes.

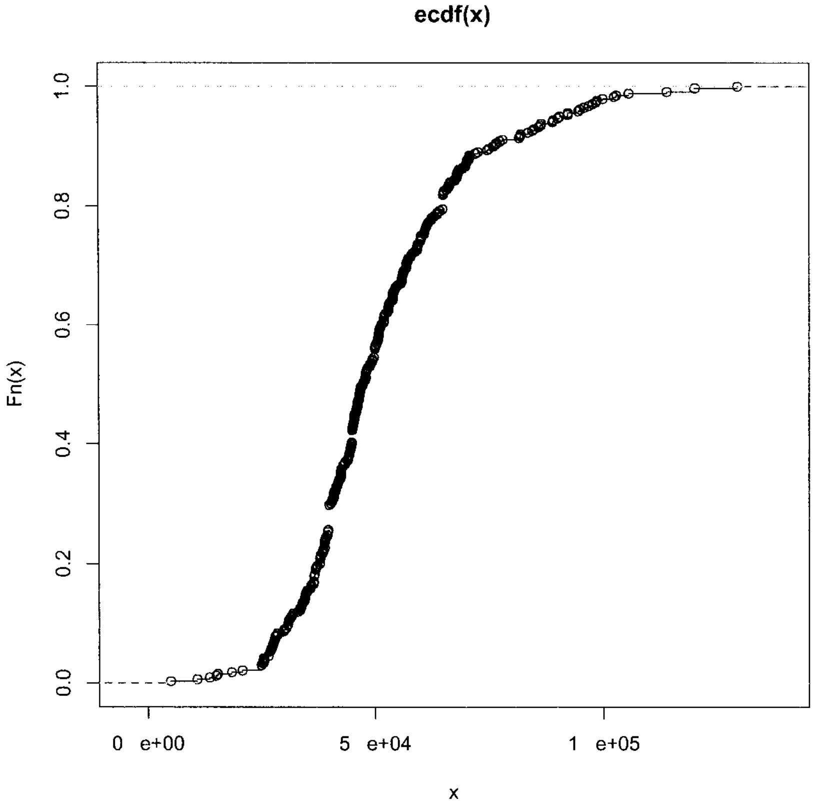

Figure 4.24 Plot of the empirical distribution function of the salaries for the 9 months employees at Louisiana Tech University

Based on this plot we conclude that the data fits a Lognormal distribution and using the maximum likelihood estimates, we determine the parameters of the distribution. In this case the mean μ, variance σ^2 and location parameter x_0 for the fitted Lognormal are 10.78023, 0.145943 and 0.001. The plot of the fitted distribution function is given in Fig. (4.25). Based on this Lognormal, the Atkinson index was calculated to be 0.06037529 in agreement with 0.05914808 obtained from the sample.

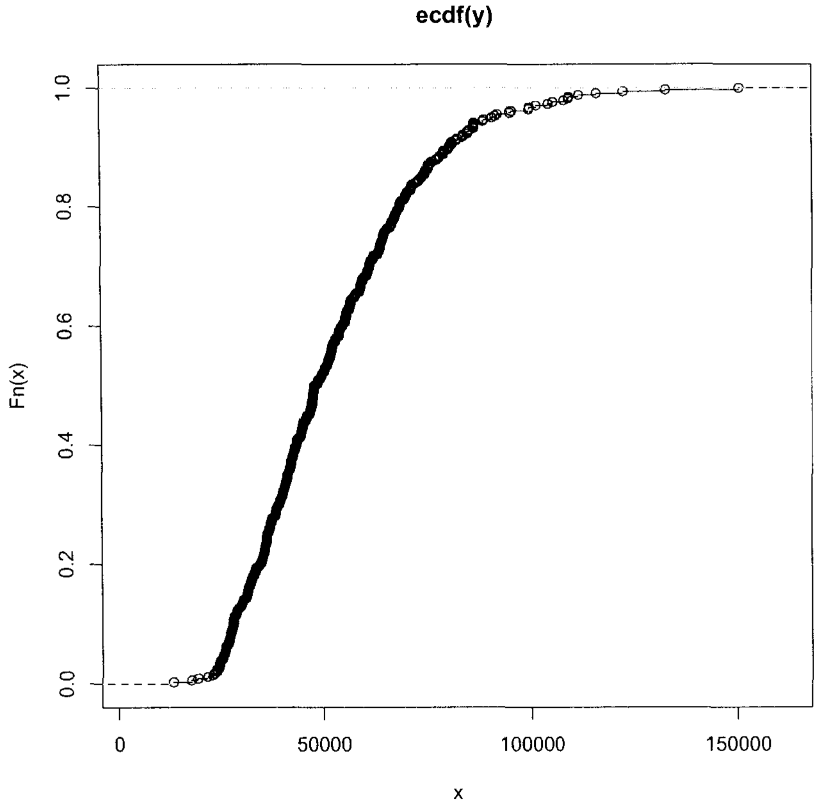

Figure 4.25 Plot of the Lognormal distribution function for $\mu = 10.78023$, $\sigma = 0.0382025$ and $x_0 = 0.001$

We can test the null hypothesis that the Atkinson index for the 12 months employees is equal to the Atkinson index for the 9 months employees, against the one-sided alternative that the Atkinson index for the 12 months employees is greater that the

Atkinson index for the 9 months employees.

$$H_0 : A_{12} - A_9 = 0$$

vs

$$H_a : A_{12} - A_9 \neq 0.$$

Note 4.5.1 *Examining only the mean of the two samples we cannot tell anything about the inequality between the salaries. The 12 months employees have a mean of 35,283.6 and the 9 months employees have a mean of 51,452.94.*

Therefore, we test the above hypothesis. Since $a = 0.1$ we are sensitive to inequality, and we want to emphasize the poor individuals (lower end of the distribution). The Z test statistics is 2.962127.

We consider different levels of significance.

At 5%, we reject the null hypothesis because $|Z| > z_{0.025} = 1.96$. There are more poor people in the 12 months employees. The bottom 5% of the incomes for the 12 months employees is in the interval (1,252, 10,000) and for the 9 months employees is in the interval (4,990, 27,000).

At 10% significance level, we reject the null hypothesis if $|Z| > z0.05 = 1.645$. The bottom 10% of the incomes for 12 months employees is in the interval (1,252, 15,000) and for 9 months employees is in the interval (4,990, 30,000).

At 20% significance level, we reject the null hypothesis if $|Z| > z0.1 = 1.282$. The bottom 20% of the incomes for 12 months employees is in the interval (1,252, 20,000) and for 9 months employees is in the interval (4,990, 35,000).

We reject the null hypothesis for all three levels of significance. After calculating the two corresponding Atkinson indices, the 12 months Atkinson index is 0.1812934 and the 9 months Atkinson index 0.05914808 one can conclude that the 9 months employees have less income inequality than the 12 months employees. The 9 months employees have their incomes around the middle of the distribution, and the Atkinson index is close to zero, a state of equality, whereas the 12 months employees have greater income inequality and with more incomes at the lower end.

CHAPTER 5

ATKINSON INDEX: TWO POPULATIONS

5.1 Asymptotic Normality Method

5.1.1 Nonparametric Case

For this chapter, we are interested in comparing two indices, say A_F and A_G, corresponding to two populations of incomes, with distribution functions F and G. In the literature, we find this comparison done in terms of comparing distribution functions F and G, or their integrals. This procedures require testing for stochastic dominance and verifying the conditions imposed, see [Horvath et al., 2005]. A different approach was chosen in the previous chapter.

The aim is to construct direct (parametric and nonparametric) tests that, given empirical evidence, would allow the practitioner to decide (at a prescribed confidence or significance level) whether A_F and A_G are equal or not under various alternatives, and to also construct confidence intervals for the difference $A_F - A_G$.

Assume we have m individuals from one society with incomes $X_1, X_2 \dots X_m$ and another n individuals from another society with incomes $Y_1, Y_2 \dots Y_n$. All random variables in the set of incomes are **independent** and distribution free.

The Atkinson index A_m is a consistent and asymptotically normally distributed estimator of A_F, and A_n is a consistent and asymptotically normally distributed estimator of A_G.

For the first population, the Atkinson index A_F measures the inequality of the incomes $X_1, X_2 \ldots X_m$ and for the second population, the Atkinson index A_G measures the inequality of the incomes $Y_1, Y_2 \ldots Y_n$. We wish to determine the magnitude of the difference between the two indices.

Through the rest of the thesis we will assume, following [Zitikis, 2003], that

Assumption 5.1.1 *There exists a function $\psi(s)$ which is continuous on (0,1) and satisfies the bound*

$$|\psi(s)| \leq cs^{\alpha-1}(1-s)^{\beta-1}, 0 < s < 1,$$

for some $\alpha, \beta > 1/2$ and $c < \infty$. Furthermore,

$$E[|X|^\gamma] + E[|Y|^\gamma] < \infty$$

for some γ such that $\gamma > 1/(\alpha - 1/2)$ and $\gamma > 1/(\beta - 1/2)$.

Also, let

$$Q_{F,G}(\psi) := \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(Pr\{X \leq x, Y \leq y\} - F(x)G(y))\psi(F(x))\psi(G(y))dxdy, \quad (5.1)$$

and

$$Q_{F,F}(\psi) := \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(F(x \wedge y) - F(x)G(y))\psi(F(x))\psi(G(y))dxdy, \quad (5.2)$$

where $x \wedge y = min(x, y)$.

Theorem 5.1.2 *Under the assumptions made above, when X_i and Y_i are independent, then*

$$\frac{(A_m[X] - A_n[Y]) - (A_F - A_G)}{\sqrt{Var(A_m[X] - A_n[Y])}} \rightarrow^d N(0,1) \quad (5.3)$$

as m and n converge to infinity in such a way that the ratio $n/(m+n)$ converges to a constant $\eta \in (0,1)$.

We have proved in Chapter 3, section 3.2 that for one population,

$$\frac{A_m[X] - A_F}{\sqrt{Var(A_m[X])}} \rightarrow^d N(0,1) \tag{5.4}$$

and

$$\frac{A_n[Y] - A_G}{\sqrt{Var(A_n[Y])}} \rightarrow^d N(0,1). \tag{5.5}$$

In this chapter, we will discuss the case when the sample sizes are equal $(m = n)$ and the case when the sample sizes are not equal $(m \neq n)$.

Before starting our proof, we would like to state the Slutsky's theorem, that will be referenced in our argument.

Theorem 5.1.3 *Let $X_n \rightarrow_d X$ and $Y_n \rightarrow_p c$, where c is a finite constant. Then*

(i) $X_n + Y_n \rightarrow^d X + c$

(ii) $X_n Y_n \rightarrow^d cX$

(iii) $\dfrac{X_n}{Y_n} \rightarrow^d \dfrac{X}{c}$, *if* $c \neq 0$.

Proof:

When $m = n$ we have

$$\frac{\sqrt{n}(A_n[X] - A_n[Y]) - \sqrt{n}(A_F - A_G)}{\sqrt{Var(A_n[X] - A_n[Y])}}.$$

Since X and Y are independent,

$$Var(A_n[x] - A_n[Y]) = Var(A_n[X]) + Var(A_n[Y]) = \frac{\sigma_F^2}{n} + \frac{\sigma_G^2}{n}.$$

Therefore, we have

$$\frac{(A_n[X]-A_n[Y])-(A_F-A_G)}{\sqrt{Var(A_n[X]-A_n[Y])}}=\frac{(A_n[X]-A_F)-(A_n[Y]-A_G)}{\sqrt{Var(A_n[X]-A_n[Y])}}$$

$$=\frac{A_n[X]-A_F}{\sqrt{Var(A_n[X])}}\times\frac{\sqrt{Var(A_n[X])}}{\sqrt{Var(A_n[X])+Var(A_n[Y])}}$$

$$-\frac{A_n[X]-A_F}{\sqrt{Var(A_n[Y])}}\times\frac{\sqrt{Var(A_n[Y])}}{\sqrt{Var(A_n[X])+Var(A_n[Y])}}.$$

Since

$$\frac{\sqrt{Var(A_n[X])}}{\sqrt{Var(A_n[X])+Var(A_n[Y])}}=\sqrt{\frac{\sigma_F^2/n}{\sigma_F^2/n+\sigma_G^2/n}}=$$

$$\sqrt{\frac{\sigma_G^2/n+\sigma_F^2/n-\sigma_G^2/n}{\sigma_F^2/n+\sigma_G^2/n}}=\sqrt{1-\frac{\sigma_G^2/n}{\sigma_G^2/n+\sigma_F^2/n}}$$

and

$$\frac{\sqrt{Var(A_n[Y])}}{\sqrt{Var(A_n[X])+Var(A_n[Y])}}=\sqrt{\frac{\sigma_G^2/n}{\sigma_G^2/n+\sigma_F^2/n}},$$

one has

$$\frac{(A_n[X]-A_F)-(A_n[Y]-A_G)}{\sqrt{Var(A_n[X])+Var(A_n[Y])}}=\frac{A_n[X]-A_F}{\sqrt{Var(A_n[X])}}\times\sqrt{1-\beta}$$

$$-\frac{A_n[Y]-A_G}{\sqrt{Var(A_n[Y])}}\times\sqrt{\beta},$$

where

$$\beta=\sqrt{\frac{\sigma_G^2/n}{\sigma_G^2/n+\sigma_F^2/n}}.$$

By Slutsky's Theorem, and from Eq. (5.4) we can see that the first term in the right–hand side of the equality converges in distribution to a standard normal, and from Eq. (5.5) the second term in the right hand side of the equality in the last expression converges in distribution to a standard normal. Let Z_1 and Z_2 be two independent standard normal variables. Therefore, by Slutsky's Theorem, the expression

$$\frac{(A_n[X] - A_F) - (A_n[Y] - A_G)}{\sqrt{Var(A_n[X]) + Var(A_n[Y])}}$$

converges in distribution to $Z_1\sqrt{1-\beta} + Z_2\sqrt{\beta}$. Let $Z = Z_1\sqrt{1-\beta} + Z_2\sqrt{\beta}$.

Therefore,

$$E[Z] = E[Z_1\sqrt{1-\beta} + Z_2\sqrt{\beta}] = \sqrt{1-\beta}E[Z_1] + \sqrt{\beta}E[Z_2] = 0.$$

We also have

$$\begin{aligned} Var(Z) &= Var(Z_1\sqrt{1-\beta} + Z_2\sqrt{\beta}) = (\sqrt{1-\beta})^2 Var(Z_1) + (\sqrt{\beta})^2 Var(Z_2) \\ &= (1-\beta) + \beta = 1. \end{aligned}$$

Hence, Z converges in distribution to a standard normal distribution, which implies that for $m = n$

$$\frac{(A_n[X] - A_n[Y]) - (A_F - A_G)}{\sqrt{Var(A_n[X] - A_n[Y])}} \tag{5.6}$$

has a standard normal distribution.

If $m \neq n$ we have to prove the asymptotic normality for

$$\frac{(A_m[X] - A_n[Y]) - (A_F - A_G)}{\sqrt{Var(A_m[X]) + Var(A_n[Y])}}. \tag{5.7}$$

We rewrite the above expression as follows:

$$\frac{(A_m[X] - A_F) - (A_n[Y] - A_G)}{\sqrt{Var(A_m[X]) + Var(A_n[Y])}} = \frac{A_m[X] - A_F}{\sqrt{Var(A_m[X])}} \times \sqrt{\frac{Var(A_m[X])}{Var(A_m[X]) + Var(A_n[Y])}}$$

$$-\frac{A_n[Y]-A_G}{\sqrt{Var(A_n[Y])}} \times \sqrt{\frac{Var(A_n[Y])}{Var(A_m[X])+Var(A_n[Y])}}.$$

Consequently,

$$\frac{\sqrt{Var(A_n[Y])}}{\sqrt{Var(A_n[Y])+Var(A_m[X])}} = \sqrt{\frac{\frac{\sigma_G^2}{n}}{\frac{\sigma_G^2}{n}+\frac{\sigma_F^2}{m}}} = \sqrt{\frac{\frac{m+n}{n}\sigma_G^2}{\frac{m+n}{n}\sigma_G^2+\frac{m+n}{m}\sigma_F^2}}.$$

We have assumed that the sizes of the two samples m and n converge to infinity in such a way that $0 \le \frac{n}{n+m} \le 1$. Let $\eta = \frac{n}{n+m}$ with $\eta \in (0,1)$.

Therefore,

$$\sqrt{\frac{\frac{\sigma_G^2}{n}}{\frac{\sigma_G^2}{n}+\frac{\sigma_F^2}{m}}} = \sqrt{\frac{\frac{1}{\eta}\sigma_G^2}{\frac{1}{\eta}\sigma_G^2+\frac{1}{1-\eta}\sigma_F^2}} = \sqrt{\delta}.$$

Similarly,

$$\frac{\sqrt{Var(A_m[X])}}{\sqrt{Var(A_n[Y])+Var(A_m[X])}} = \sqrt{\frac{\frac{\sigma_F^2}{m}}{\frac{\sigma_G^2}{n}+\frac{\sigma_F^2}{m}}} = \sqrt{\frac{\frac{m+n}{m}\sigma_F^2}{\frac{m+n}{n}\sigma_G^2+\frac{m+n}{m}\sigma_F^2}}.$$

Therefore,

$$\sqrt{\frac{\frac{\sigma_F^2}{m}}{\frac{\sigma_G^2}{n}+\frac{\sigma_F^2}{m}}} = \sqrt{\frac{\frac{1}{1-\eta}\sigma_F^2}{\frac{1}{\eta}\sigma_G^2+\frac{1}{1-\eta}\sigma_F^2}} = \sqrt{1-\delta}.$$

Since from Eq. (5.4) and Eq. (5.5), Z_1 and Z_2 are two standard normal independent variables, we have that

$$\frac{(A_m[X]-A_n[Y])-(A_F-A_G)}{\sqrt{Var(A_m[X])+Var(A_n[Y])}} \text{ converges to } Z_1\sqrt{1-\delta}+Z_2\sqrt{\delta}.$$

Let $Z = Z_1\sqrt{1-\delta}+Z_2\sqrt{\delta}$. Hence,

$$E[Z] = E[Z_1\sqrt{1-\delta}+Z_2\sqrt{\delta}] = \sqrt{1-\delta}E[Z_1]+\sqrt{\delta}E[Z_2] = 0.$$

We also have

$$Var(Z) = Var(Z_1\sqrt{1-\delta} + Z_2\sqrt{\delta}) = (\sqrt{1-\delta})^2 Var(Z_1) + (\sqrt{\delta})^2 Var(Z_2)$$

$$= (1-\delta) + \delta = 1.$$

Therefore, from Theorem (4.1.2), the asymptotic distribution of

$$\frac{(A_m[X] - A_n[Y]) - (A_F - A_G)}{\sqrt{Var(A_m[X] - A_n[Y])}} \tag{5.8}$$

is standard normal. ■

In the case of **paired random** variables, we consider two populations of incomes, the first population with incomes $X_1, X_2 \ldots X_n$ following a distribution F, and a second population with incomes $Y_1, Y_2, \ldots Y_n$ following a distribution G. It is assumed that the pairs (X_i, Y_i), $i = 1, \ldots, n$, are independent, and there is correlation between X_i and Y_i.

In practice, one is interested in comparing the Atkinson indices for the two populations at two different times. We have already proved in Chapter 3, that for one sample

$$\frac{\sqrt{n}(A_n[x] - A_F)}{\sqrt{\sigma_F^2}} \text{ and } \frac{\sqrt{n}(A_n[y] - A_G)}{\sqrt{\sigma_G^2}}$$

are asymptotically standard normal.

In this section, we will investigate the case of two samples that are not independent. Using the Taylor series expansion, we have

$$\sqrt{n}(A_n[X] - A_F) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h(X_i) + R_{n,x}$$

and

$$\sqrt{n}(A_n[Y] - A_G) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} h(Y_i) + R_{n,y},$$

where $h(X_i) = [h'_x(\mu, \mu_a)(X_i - \mu) + h'_y(\mu_i, \mu_a)(X_i^a - \mu_a)]$ and $h(Y_i)$ is defined similarly for the second sample. The two remainders $R_{n,x}$ and $R_{n,y}$ converge to 0.

In order to obtain the desired asymptotic results for the difference of measures concerning X and Y, we shall now look at the the following:

$$\sqrt{n}(A_n[X] - A_n[Y]) - \sqrt{n}(A_F - A_G)$$

$$= \sqrt{n}(A_n[X] - A_F) - \sqrt{n}(A_n[Y] - A_G) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(h(X_i) - h(Y_i)) + (R_{n,x} - R_{n,y}).$$

Let $Z_i = h(X_i) - h(Y_i)$ and $R_n = R_{n,x} - R_{n,y}$. Therefore, we have

$$\sqrt{n}(A_n[X] - A_n[Y]) - \sqrt{n}(A_F - A_G) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i + R_n \rightarrow^d N(0, \sigma^2).$$

From Chapter 2, using Slutsky's Theorem we can conclude that R_n converges in probability to 0 and we know that $E[Z_i] = E[h(X_i) - h(Y_i)] = E[h(X_i)] - E[h(Y_i)] = 0$. Therefore $\sqrt{n}(A_n[X] - A_n[Y]) - \sqrt{n}(A_F - A_G)$ converges to a normal distribution, with mean zero and variance σ^2, where

$$\sigma^2 = Var(Z_i) = Var(h(X_i) - h(Y_i))$$

$$= E[(h(X_i) - h(Y_i))^2] = E[h(X_i)]^2 - 2E[h(X_i)h(Y_i)] + E[h(Y_i)]^2.$$

Hence,

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} h(X_i)^2 - 2\frac{1}{n}\sum_{i=1}^{n} h(X_i)h(Y_i) + \frac{1}{n}\sum_{i=1}^{n} h(Y_i)^2, \tag{5.9}$$

which by the Law of Large Numbers converges to σ^2.

In summary, using the Slutsky's argument, we have the asymptotic result:

$$\frac{\sqrt{n}\left((A_n[X] - A_n[Y]) - (A_F - A_G)\right)}{\sqrt{\widehat{\sigma}^2}} \rightarrow^d N(0, 1). \tag{5.10}$$

5.1.2 Parametric Case

In the previous chapter, we discussed the one-sample parametric Atkinson index. Here, we consider that all the random variables in the set $\{X_1, \ldots, X_n, Y_1, \ldots, Y_n\}$ are **independent**.

The parametric families under consideration are the Pareto, Exponential, and Lognormal distributions:

$$F_1(x) = 1 - \left(\frac{x_0}{x}\right)^{\lambda}, x > x_0, \lambda > 0, \tag{5.11}$$

$$F_2(x) = 1 - e^{-(x-x_0)/\theta}, x > x_0, \theta > 0, \tag{5.12}$$

and

$$F_3(x) = \Phi(log(x - x_0) - \mu, x > x_0, -\infty < \mu < \infty. \tag{5.13}$$

In this study, we are interested the following pairs of distributions: Pareto-Pareto, Lognormal-Lognormal, and Pareto-Lognormal.

We assume first that the random variables X_i and Y_i follow a Pareto distribution with cumulative distribution functions F and G and parameters λ and θ, respectively. The X_i and Y_i are independent variables. From Chapter 2, we know that

$$\frac{\sqrt{n}(\widehat{A}_F - A_F)}{\sqrt{Var(A_F)}} \rightarrow^d N(0, 1) \tag{5.14}$$

and

$$\frac{\sqrt{m}(\widehat{A}_G - A_G)}{\sqrt{Var(A_G)}} \rightarrow^d N(0, 1). \tag{5.15}$$

Hence, we want to show that

$$\frac{(\widehat{A}_F - \widehat{A}_G) - (A_F - A_G)}{\sqrt{\dfrac{Var(\widehat{A}_F)}{n} + \dfrac{Var(\widehat{A}_G)}{m}}} \rightarrow^d N(0, 1) \tag{5.16}$$

as n and m converge to infinity in such a way that the ratio $n/(n+m)$ converges to a constant $\alpha \in (0,1)$.

It is seen that,

$$\frac{(\widehat{A}_F - \widehat{A}_G) - (A_F - A_G)}{\sqrt{\frac{Var(\widehat{A}_F)}{n} + \frac{Var(\widehat{A}_G)}{m}}} = \frac{\widehat{A}_F - A_F}{\sqrt{\frac{Var(\widehat{A}_F)}{n}}} \times \sqrt{\frac{\frac{Var(\widehat{A}_F)}{n}}{\frac{Var(\widehat{A}_F)}{n} + \frac{Var(\widehat{A}_G)}{m}}}$$

$$+ \frac{\widehat{A}_G - A_G}{\sqrt{\frac{Var(\widehat{A}_G)}{m}}} \times \sqrt{\frac{\frac{Var(\widehat{A}_G)}{m}}{\frac{Var(\widehat{A}_F)}{n} + \frac{Var(\widehat{A}_G)}{m}}}.$$

Since we already know from Eq. (4.14) and Eq. (4.15) that each of the first terms in the sum on the right side of the above formula are asymptotically normal, we are now interested in the remaining two terms.

It follows that

$$\sqrt{\frac{\frac{Var(\widehat{A}_F)}{n}}{\frac{Var(\widehat{A}_F)}{n} + \frac{Var(\widehat{A}_G)}{m}}} = \sqrt{\frac{\frac{1}{n}\widehat{\sigma}_F^2}{\frac{1}{n}\widehat{\sigma}_F^2 + \frac{1}{m}\widehat{\sigma}_G^2}} = \sqrt{\frac{\frac{m+n}{n}\widehat{\sigma}_F^2}{\frac{m+n}{n}\widehat{\sigma}_F^2 + \frac{m+n}{m}\widehat{\sigma}_G^2}},$$

where $\widehat{\sigma}_F^2 = \frac{1}{\theta^2}\left(h'\left(\frac{1}{\theta}\right)\right)^2$ and $\widehat{\sigma}_G^2 = \frac{1}{\lambda^2}\left(h'\left(\frac{1}{\lambda}\right)\right)^2$ as in the previous chapter.

We let $\eta = n/(n+m)$, therefore, we have $\frac{1}{\eta} = \frac{m+n}{n}$ and $\frac{1}{1-\eta} = \frac{m+n}{m}$. Hence, the previous equality becomes

$$\sqrt{\frac{\frac{Var(\widehat{A}_F)}{n}}{\frac{Var(\widehat{A}_F)}{n} + \frac{Var(\widehat{A}_G)}{m}}} = \sqrt{\frac{\frac{1}{\eta}\sigma_F^2}{\frac{1}{\eta}\sigma_F^2 + \frac{1}{1-\eta}\sigma_G^2}} = \sqrt{\delta} \qquad (5.17)$$

and similarly

$$\sqrt{\frac{\frac{Var(\widehat{A}_G)}{m}}{\frac{Var(\widehat{A}_F)}{n}+\frac{Var(\widehat{A}_G)}{m}}}=\sqrt{\frac{\frac{1}{1-\eta}\sigma_F^2}{\frac{1}{\eta}\sigma_F^2+\frac{1}{1-\eta}\sigma_G^2}}=\sqrt{1-\delta}. \tag{5.18}$$

Since, from Eq. (4.4) and Eq. (4.5) Z_1 and Z_2 are two standard normal independent variables, we have that

$$\frac{(\widehat{A}_F-\widehat{A}_G)-(A_F-A_G)}{\sqrt{\frac{Var(\widehat{A}_F)}{n}+\frac{Var(\widehat{A}_G)}{m}}} \text{ converges to } Z_1\sqrt{1-\delta}+Z_2\sqrt{\delta}.$$

Let $Z = Z_1\sqrt{1-\delta}+Z_2\sqrt{\delta}$, and $E(Z) = 0, Var(Z) = 1$, respectively. Therefore, from Theorem (5.1.2) the asymptotic distribution of

$$\frac{(\widehat{A}_F-\widehat{A}_G)-(A_F-A_G)}{\sqrt{\frac{Var(\widehat{A}_F)}{n}+\frac{Var(\widehat{A}_G)}{m}}} \tag{5.19}$$

is standard normal, $N(0,1)$.

The same asymptotic results are obtained for F denoting a Pareto distribution and G denoting a Lognormal distribution or for F and G denoting Lognormal distributions.

In the parametric case for the **paired** random variables we have to show, analogous, to Eq. (5.10), the following asymptotic result:

$$\frac{\sqrt{n}\Big((\widehat{A}_F-\widehat{A}_G)-(A_F-A_G)\Big)}{\sqrt{\widehat{\sigma}^2}} \rightarrow^d N(0,1), \tag{5.20}$$

where the denominator on the left-hand side of the above formula is the asymptotic standard deviation of $\sqrt{n}(\widehat{A}_F-\widehat{A}_G)$.

Depending on the distribution of F and G, the formulae for $\widehat{A}_F$ and $\widehat{A}_G$ are of the form

$$\begin{aligned} A_{F_1} &= 1 - \left(1 - \frac{1}{\lambda}\right)\left(\frac{\lambda}{\lambda - a}\right)^{\frac{1}{a}}, && (5.21)\\ A_{F_2} &= 1 - \Gamma(a+1)^{\frac{1}{a}}, && (5.22)\\ A_{F_3} &= 1 - \frac{(x_0 + e^{a\mu + \frac{a^2}{2}})^{\frac{1}{a}}}{x_0 + e^{\mu + \frac{1}{2}}}, && (5.23) \end{aligned}$$

with the corresponding parametric empirical estimators

$$\begin{aligned} A_{F_1} &= 1 - \left(1 - \frac{1}{\widehat{\lambda}}\right)\left(\frac{\widehat{\lambda}}{\widehat{\lambda} - a}\right)^{\frac{1}{a}}, && (5.24)\\ A_{F_2} &= 1 - \Gamma(a+1)^{\frac{1}{a}}, && (5.25)\\ A_{F_3} &= 1 - \frac{(x_0 + e^{a\widehat{\mu} + \frac{a^2}{2}})^{\frac{1}{a}}}{x_0 + e^{\widehat{\mu} + \frac{1}{2}}}. && (5.26) \end{aligned}$$

5.2 Asymptotic Variance and its Estimation

Having proved in Section 5.1 the asymptotic normality for the nonparametric case and independent samples, we can continue by constructing consistent empirical estimators for the variances.

Using the aforementioned indices, we have that

$$Var(A_m[X] - A_n[Y]) = \frac{Q_{F,F}(\psi)}{m} + \frac{Q_{G,G}(\psi)}{n}, \quad (5.27)$$

where $Q_{F,F}(\psi)$ and $Q_{G,G}(\psi)$ are given by Eq. (5.2).

[Jones and Zitikis, 2003] provided the following estimator for $Q_{F,F}(a,b)$, with that for $Q_{G,G}(a,b)$ defined analogously:

$$Q_n[X](\psi) := \sum_{j=1}^{n-1}\sum_{k=1}^{n-1}\left(\frac{j}{n}\wedge\frac{k}{n}-\frac{j}{n}\frac{k}{n}\right)\psi\left(\frac{j}{n}\right)\psi\left(\frac{k}{n}\right)(X_{j+1:n}-X_{j:n})(X_{k+1:n}-X_{k:n}). \tag{5.28}$$

It is proved in [Zitikis, 2002] that under Assumption (5.1.1), $Q_n[X](\psi)$ is a consistent estimator of $Q_{F,F}(\psi)$. Again, using Slutsky's argument we have the following result

$$\frac{(A_m[X]-A_n[Y])-(A_F-A_G)}{\sqrt{\dfrac{Q_m[X](\psi)}{m}+\dfrac{Q_n[Y](\psi)}{n}}}\to^d N(0,1). \tag{5.29}$$

One can use this result to test the null hypothesis $H_0 : A_F = A_G$ against any (one-sided or two-sided) alternative and construct asymptotic confidence intervals, as is seen in the following section.

In the parametric case, we have the following asymptotic result:

$$\frac{(\widehat{A}_F-\widehat{A}_G)-(A_F-A_G)}{\sqrt{\dfrac{\widehat{Q}_{F,F}(\psi)}{m}+\dfrac{\widehat{Q}_{G,G}(\psi)}{n}}}\to^d N(0,1). \tag{5.30}$$

Parametric estimators for the quantities used in the above result are obtained by replacing the parameters of the distributions by their corresponding maximum likelihood estimators.

Therefore, for the Exponential and Pareto distributions

$$\widehat{Q}_{F_1,F_1}(\psi) = \frac{\widehat{\theta}^2}{(1-a)^2} \tag{5.31}$$

and

$$\widehat{Q}_{F_2,F_2}(\psi) = \frac{(1-a)^2\widehat{\lambda}^2}{((1-a)\widehat{\lambda}-1)^4}. \tag{5.32}$$

In the case of a paired sample, we have the following nonparametric asymptotic result:

$$\frac{(A_n[X] - A_n[Y]) - (A_F - A_G)}{\sqrt{Q_n[X](\psi) + Q_n[Y](\psi) - 2Q_n[X,Y](\psi)}} \to^d N(0,1) \tag{5.33}$$

The estimator $Q_n[X,Y]$ is defined as follows:

$$\begin{aligned} Q_n[X,Y](\psi) &= \sum_{j=1}^{n-1}\sum_{k=1}^{n-1}\left(\frac{\epsilon_n(j,k)}{n} - \frac{j}{n}\frac{k}{n}\right)\psi\left(\frac{j}{n}\right)\psi\left(\frac{k}{n}\right) \\ &\times\ (X_{j+1:n} - X_{j:n})(Y_{k+1:n} - Y_{k:n}), \end{aligned}$$

where $\epsilon_n(j,k) \equiv \sum_{i=1}^{j} 1_{\{Y_{i:n}^{ind} \le Y_{k:n}\}}$ and $Y_{i:n}^{ind}$ are the induced order statistics, as shown in [Zitikis, 2002].

5.3 Confidence Intervals and Simulation Studies

5.3.1 Nonparametric Case

For two independent samples, a nonparametric asymptotic $100(1-\alpha)\%$ confidence interval is given by

$$A_m[X] - A_n[Y] \pm z_{\alpha/2}\sqrt{\frac{Q_m[X](\psi)}{m} + \frac{Q_n[Y](\psi)}{n}},$$

where z_α is the $100(1-\alpha)\%$ quantile of the standard normal distribution.

To construct nonparametric bootstrap approximations we reformulate the asymptotic result in Eq. (5.29) as follows:

$$\sqrt{\frac{mn}{m+n}}((A_m[X]-A_n[Y])-(A_F-A_G)) \to^d N(0,(1-\eta)Q_{F,F}(\psi)+\eta Q_{G,G}(\psi)), \tag{5.34}$$

where η is the limit of $m/(m+n)$ when both m and n approach infinity. From the left-hand side of Eq. (5.34) we see that the bootstrap critical value z_α^* in the following $100(1-\alpha)\%$ confidence interval

$$(A_m[X] - A_n[Y]) \pm z_\alpha^*\sqrt{\frac{m+n}{mn}} \tag{5.35}$$

for the difference $A_F - A_G$, which can be constructed as follows: First, we sample with replacement m values $X_1^*, X_2^*, \ldots X_n^*$ from the original sample $X_1, X_2 \ldots X_n$ and we calculate the corresponding income measure, which we denote by $A_m[X^*]$. In the same manner, we obtain $A_n[Y^*]$ from the sample $Y_1, Y_2, \ldots, Y_n$.

Then, we calculate the values of

$$\sqrt{\frac{mn}{m+n}}|(A_m[X^*] - A_n[Y^*]) - (A_m[X] - A_n[Y])|. \tag{5.36}$$

The above procedure is repeated k times to obtain k values for the expression in Eq. (5.36). Now one can define z_α^* in Eq. (5.35) as the smallest value of z such that at least $100(1-\alpha)\%$ of the values of the expression in Eq. (5.36) are at or below z.

In the case of paired samples, with equal sample size, the nonparametric asymptotic confidence interval for the difference $A_F - A_G$ is given by the following expression.

$$A_n[X] - A_n[Y] \pm z_{\alpha/2}\frac{1}{\sqrt{n}}\sqrt{Q_n[X](\psi) + Q_n[Y](\psi) - 2Q_n[X,Y](\psi)}. \tag{5.37}$$

For a bootstrap nonparametric confidence interval, for the difference $A_F - A_G$ in the case of the paired samples, we start with the original pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ and calculate the empirical measures $A_n[X]$ and $A_n[Y]$.

We will continue our procedure by sampling with replacement from the original sample to obtain a new sample $(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$ and calculate $A_n[X^*]$ and $A_n[Y^*]$ as well as

$$\sqrt{n}|(A_n[X^*] - A_n[Y^*]) - (A_n[X] - A_n[Y])|. \tag{5.38}$$

The above procedure is repeated k times to obtain k values of the expression in Eq. (5.38). Now, one can define z_α^* as the smallest value of z such that at least $100(1-\alpha)\%$ of the values of the expression in Eq. (5.38) are at or below z.

Hence, the desired confidence interval for $A_F - A_G$ is

$$A_n[X] - A_n[Y] \pm z^*_{\alpha/2}\frac{1}{\sqrt{n}}. \tag{5.39}$$

5.3.2 Parametric Case

To construct asymptotic parametric confidence intervals and tests of hypotheses for the case of independent samples, we apply the following asymptotic result:

$$\frac{(\widehat{A}_F - \widehat{A}_G) - (A_F - A_G)}{\sqrt{\dfrac{\widehat{Q}_{F,F}(\psi)}{m} + \dfrac{\widehat{Q}_{G,G}(\psi)}{n}}} \to^d N(0,1). \tag{5.40}$$

Therefore, a $100(1-\alpha)\%$ asymptotic confidence interval is given by

$$\widehat{A}_F - \widehat{A}_G \pm z_{\alpha/2}\sqrt{\frac{\widehat{Q}_{F,F}(\psi)}{m} + \frac{\widehat{Q}_{G,G}(\psi)}{n}}. \tag{5.41}$$

To construct parametric bootstrap confidence intervals, we assume parametric families for X and Y. The statistics $\widehat{A}_F$ and $\widehat{A}_G$ can be estimated as discussed earlier, when F and G are distribution functions from the three distributions considered in this thesis, Pareto, Exponential, and Lognormal. The quantities $\widehat{Q}_{F,F}(\psi)$ and $\widehat{Q}_{G,G}(\psi)$ are defined in Eqs. (5.31) and (5.32), respectively.

Since we have the maximum likelihood estimators of the parameters, one can draw simple random samples from the three distributions using the bootstrap technique and obtain new samples $(X^*_1, \ldots, X^*_m)$ and $(Y^*_1, \ldots, Y^*_n)$ using the bootstrap technique.

Using the new samples, we calculate again the maximum likelihood estimators for the three distributions and formulate the expression

$$\sqrt{\frac{mn}{m+n}}|(\widehat{A}^*_F - \widehat{A}^*_G) - (\widehat{A}_F - \widehat{A}_G)|. \tag{5.42}$$

The above procedure is repeated k times to obtain k values for the expression in Eq. (5.42). We define z_α^* as the smallest value of z such that at least $100(1-\alpha)\%$ of the values of Eq. (5.42) are at or below z.

Therefore, the corresponding parametric bootstrap based $100(1-\alpha)\%$ interval for $A_F - A_G$ is

$$(\widehat{A}_F - \widehat{A}_G) \pm z_\alpha^* \sqrt{\frac{m+n}{mn}}. \tag{5.43}$$

Also, the asymptotic parametric confidence interval for the difference $A_F - A_G$ in the paired sample case is given by

$$A_n[X] - A_n[Y] \pm z_{\alpha/2} \frac{1}{\sqrt{n}} \sqrt{Q_{F,F}(\psi) + Q_{G,G}(\psi) - 2Q_{F,G}(\psi)}. \tag{5.44}$$

CHAPTER 6

SIMULATION STUDIES II

In this chapter we are interested in verifying the consistency of the estimators for the variance of the difference of two indices and studying the coverage probability.

The first study is the asymptotic parametric case for two independent populations of the same size, and we consider two sets of incomes, following the Pareto distribution, with $\lambda_m = 8$ and $\lambda_n = 10$. The two populations are assumed to be independent and the samples have the same size $m = n$. We construct confidence intervals based on Eq. (5.41) with the quantities $\widehat{Q}_{F,F}(\psi)$, $\widehat{Q}_{G,G}(\psi)$ defined in Eqs. (5.31) and (5.32), respectively.

We have used the function $\psi(s) = g'(1-s)$, where $g(s) = s^r$ and $r = 1-a$. For the asymptotic conditions to be satisfied, we take $r > 0.5$, which means that we are only interested in the cases where the Atkinson parameter 'a' is $0 < a < 0.5$. Appendix B presents the program code in R for the simulation for the case m =10,000 and n =10,000.

The sample in the first population has a theoretical Atkinson index $Atkinson_m =$ 0.03073342 and the sample in the second population has a theoretical Atkinson index $Atkinson_n = 0.01885358$. The horizontal line in Figs. (6.1) and (6.2) represents the value of the theoretical Atkinson index.

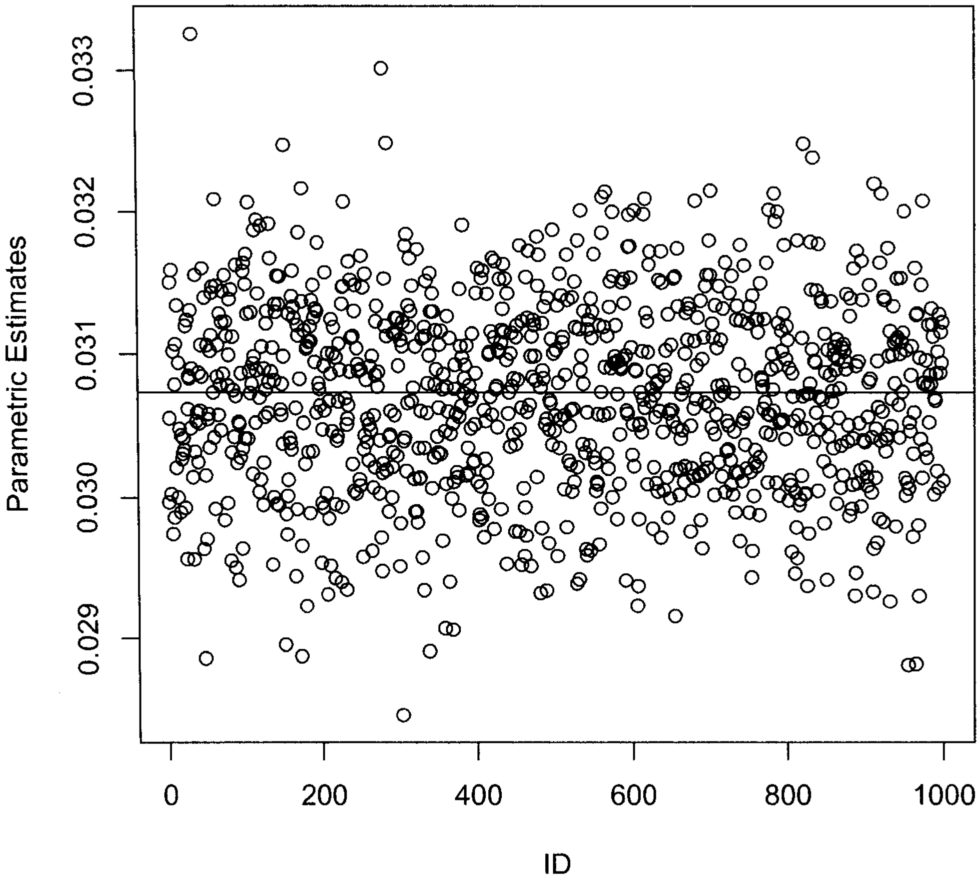

Figure 6.1 Scatter plot of parametric Atkinson estimates for sample size 10,000 in the first population

The mean of the parametric Atkinson estimates for the first sample is 0.03070846, and the mean of the parametric Atkinson estimates for the second sample is 0.01884488. These means are very close to the theoretical means. Figures (6.1) and (6.2) show that the scatter of the estimates is larger in the first population than it is in the second population.

The plot in Fig. (6.3) represents a scatter plot of the 1,000 parametric variance estimates for the first sample. The parametric variance for the first population is

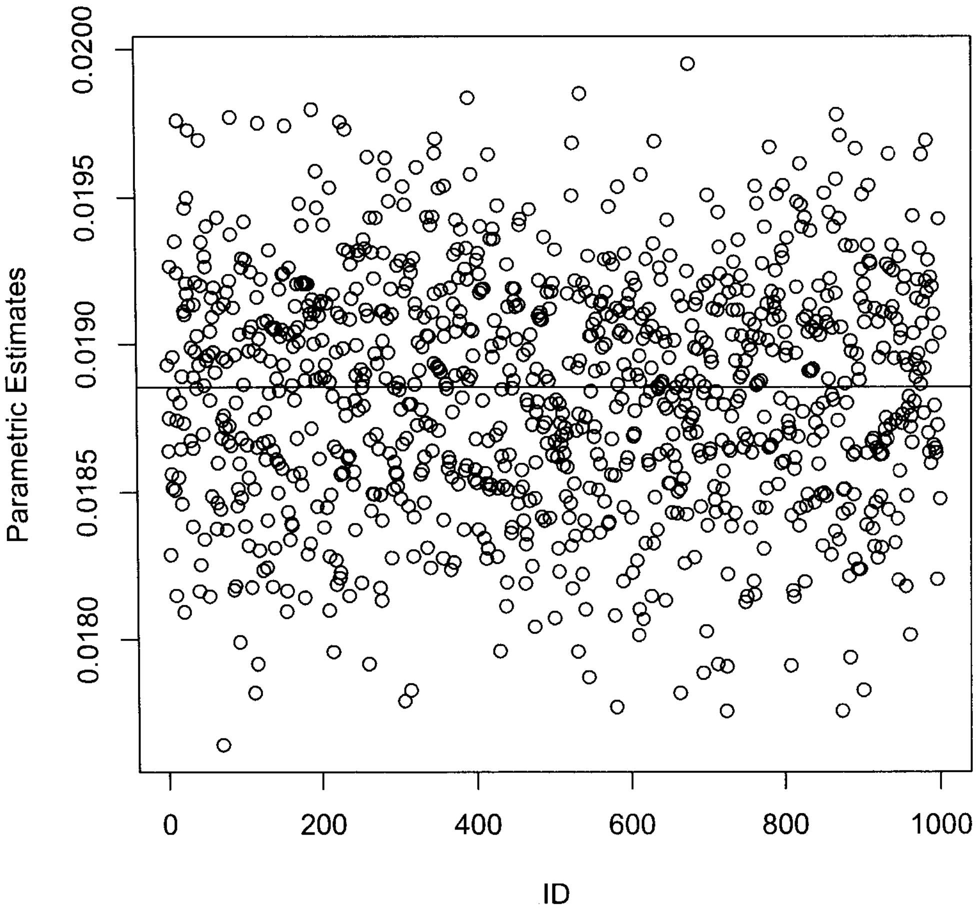

Figure 6.2 Scatter plot of parametric Atkinson estimates for sample size 10,000 in the second population

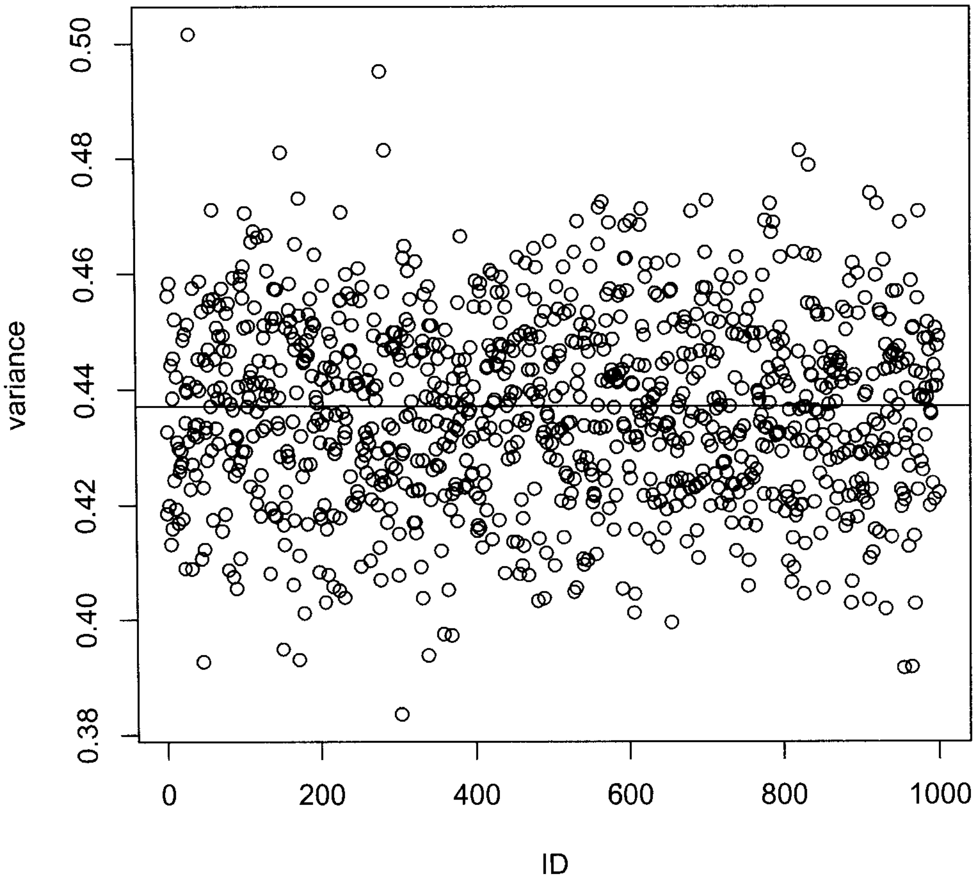

Figure 6.3 Scatter plot of parametric variance estimates for sample size 10,000 in the first population

0.4371286, and the mean of the variance estimates is 0.4366905.

The plot in Fig. (6.4) represents a scatter plot of the 1,000 parametric variance estimates for the second population. The parametric variance for the second population is 0.1975309 and the mean of the variance estimates is 0.1974422.

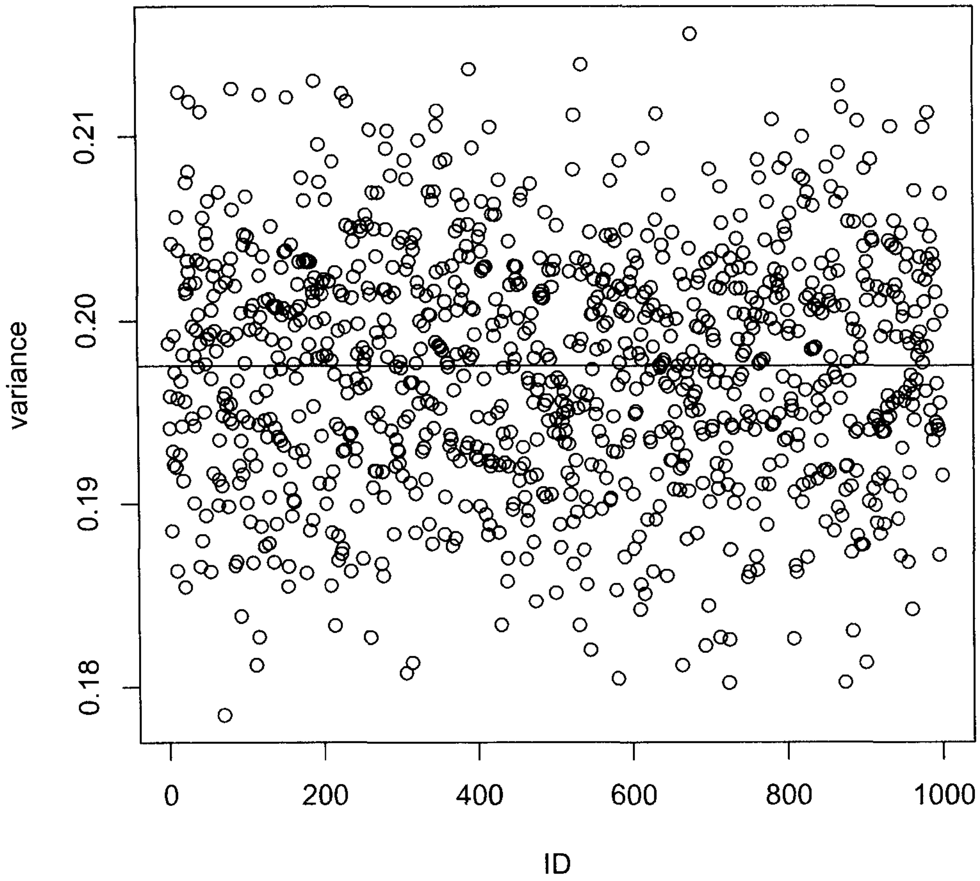

Figure 6.4 Scatter plot of parametric variance estimates for sample size 10,000 in the second population

The previous two figures indicate that the spread of variance estimates in the second sample is considerably smaller than the spread of the variance estimates in the first sample.

Also, from the plots of the parametric Atkinson estimates, we observe that the second sample has more equality than the first sample and the theoretical Atkinson index for the second sample is smaller that the theoretical Atkinson index for the first sample.

The approximate 95 percent confidence intervals for the difference of two Atkinson indices using the parametric estimation with asymptotic intervals are presented in Table 6.1.

Table 6.1 Coverage probability of the difference of two Atkinson indices from simulation using the Pareto distribution with 1,000 replicates.

Sample size	Proportion Parametric
n=100	0.951
n=500	0.948
n=1,000	0.943
n=10,000	0.939

We observe that the parametric approach gives very good coverage for small sample size, mainly due to the fact that the correct distributions were fitted.

The second study is the nonparametric bootstrap case for two independent populations with different sample sizes. We consider two sets of incomes, following a Pareto and an Exponential distribution, with $\lambda = 3$ and $\gamma = 5$. The two populations are independent, and the samples have different sizes $(m \neq n)$. We construct confidence intervals based on Eqs. (5.35) and (5.36).

In some cases one doesn't know how the variance looks like or how to estimate the variance, as we have seen in Chapter 3 for the Exponential distribution.

In cases like this, we use the sampling distribution of the statistic to calculate the critical values. We consider that the sample is the new population and we resample from this new population. The variance of the new population is very close to the variance of the old population and the critical values that we obtain for the test are similar to the critical values of the initial population.

The Pareto sample has an Atkinson index of 0.05870694 and the mean of the nonparametric estimates is 0.05867414. The Exponential sample has an Atkinson index of 0.3474519 and the mean of the nonparametric estimates is 0.3468081. As expected, the means from simulation are very close in value to the theoretical means.

We are interested in testing the null hypothesis $H_0 : A_F = A_G$ against the the alternative alternative $H_a : A_F < A_G$ and in constructing confidence intervals for the difference $A_F - A_G$.

The difference between the two Atkinson indices to be $A_F - A_G = -0.288745$. Figure (6.5) is a scatter plot of the differences between the nonparametric estimates of the two Atkinson indices.

We reject the null hypothesis in favor of the alternative hypothesis. The approximate 95 percent confidence intervals for the difference of two Atkinson indices using the nonparametric estimation with bootstrap intervals are presented in Table 6.2.

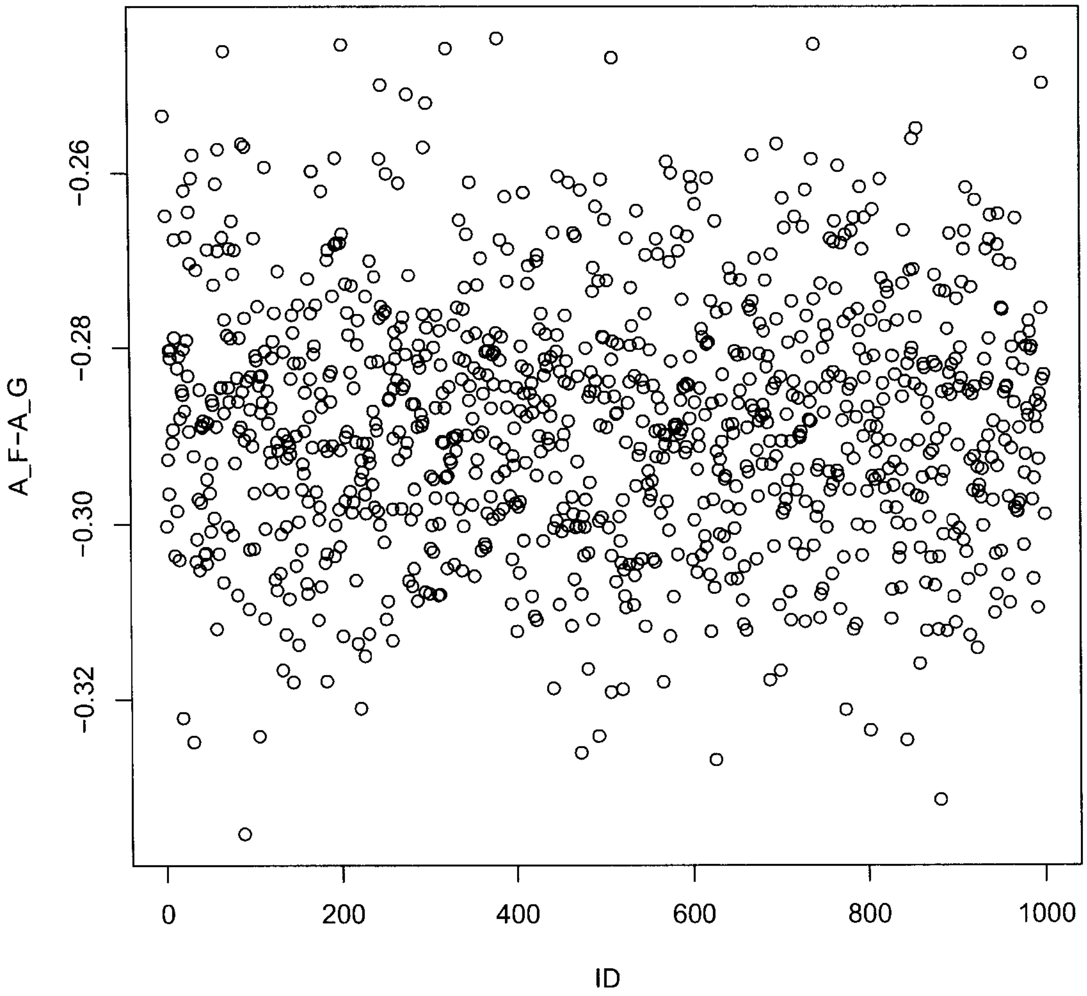

Figure 6.5 Scatter plot of the differences of two Atkinson index estimates over 1,000 replicates.

Table 6.2 Coverage probability of the difference of two Atkinson indices from simulation using the nonparametric bootstrap method.

Sample size	Proportion Nonparametric Bootstrap
n=100, m=120	0.978
n=500, m=510	0.971
n=1,000, m=1,050	0.964

It is seen that these intervals are close to 95%, the expected confidence interval. Also, as the sample size increases, the intervals approaches 95% consistently.

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

In this dissertation, we obtained parametric and nonparametric empirical estimators of the Atkinson index and developed the asymptotic and bootstrap inference for these estimators. Specifically, A_n denotes the classical empirical estimator of the theoretical Atkinson index $A_{F,a}$ and $\widehat{A}_F$ is the parametric estimator of the parametric Atkinson index A_F, see Eqs. (2.3) and (2.5).

An advantage of using nonparametric empirical estimation is that it allows the practitioner to estimate the index value from a random sample without first fitting a parametric model to the data. However, we do not want to imply that the nonparametric methods should be preferred to the parametric ones: The methods complement each other and experimentation with simulated data could be used to determine when it is preferable to use one method or the other in practice.

In some cases, one may be interested in comparing the Atkinson index for two or more unknown distributions. Chapter 4 of this dissertation addresses the two distribution case.

For example, the Atkinson index may differ for different periods of time. It is of interest to understand how the Atkinson index values differ. One way to understand the relationship between the Atkinson index values is to perform a test of hypothesis where the values are equal versus various alternative hypotheses. Here, we examined empirical tests in the case where only two Atkinson indices are being compared.

We plan to pursue this idea further and develop hypothesis tests for equality of more than two Atkinson indices. It should be mentioned that constructing tests for the equality of three or more Atkinson indices values is not a straightforward generalization of the case of two Atkinson indices discussed in Chapter 4. It is a very interesting problem from the theoretical point of view and is very important for practical purposes.

Several papers in the literature, ([Puri, 1965], [Shorack, 1967], to mention only a few) address testing hypotheses about the equality of k distribution functions, $F_1, F_2, ..., F_k$ against various alternatives.

We assume k populations with independent samples from each population. For each i^{th} population construct the corresponding empirical distribution function and the empirical Atkinson index.

One can test the null hypothesis that the Atkinson index values are equal to some known value A_0 against the one-sided alternative in which the Atkinson index values for several groups are no less than the specified A_0 with at least one group having an Atkinson index value that is strictly greater.

As such,

$$H_0 : A_i = A_0 \text{ for all } i = 1, 2...k;$$

vs

$$H_a : A_i \geq A_0 \text{ for all } i = 1, 2...k-1;$$

and there exists an i_0 such that $A_{i_0} > A_0$, with $1 \leq i_0 < k$.

We can also test the following hypothesis:

$$H_0 : A_i = A_0 \text{ for all } i = 1, 2...k;$$

vs

$$H_a : \text{ there exists an } i_0 \text{ such that } A_{i_o} \neq A_0.$$

In this case the Atkinson index values associated with a number of groups are equal to some known value against the alternative hypothesis that at least one of the Atkinson index values differs from the specified value.

One other possible scenario may suggest some natural ordering of several groups of incomes and the economist may be interested in whether the Atkinson index values corresponding to these groups should be ordered in a manner consistent with this natural ordering. For example, incomes in highly industrialized areas (high-tech companies, etc.) have greater income than other areas. Another scenario is where an economist may be interested in whether or not the Atkinson index value for some group of incomes has increased over time as a result of inflation. In this case, on may wish to test the following hypothesis:

$$H_0 : A_i = A_0 = \cdots = A_k;$$

vs

$$H_a : A_1 \leq A_1 \leq \cdots \leq A_k;$$

with at least one strict inequality.

Several other possible scenarios can be tested following real life situations. Many statistical methods can be used to make inferences about populations and their various parameters. Since no single method is best for every situation, all these methods are interesting and would help an economist in understanding the problem on hand.

APPENDIX A

FORMULAS

In this dissertation we consider three distributions: Pareto, Exponential and Lognormal distributions. In what follows we will present the formulas for the Gini and the Atkinson indices for each of the above distributions.

For the Pareto distribution, the probability density function is of the form

$$f(x) = x^{-\lambda-1}\lambda {x_0}^{\lambda},$$

where $\lambda > 0$ and $x_0 > 0$. Therefore, the cumulative distribution function is of the form

$$F(x) = 1 - \left(\frac{x_0}{x}\right)^{\lambda}.$$

The Lorenz curve for the Pareto distribution is of the form

$$L(p) = 1 - (1-p)^{(\lambda-1)/\lambda}.$$

The S-Gini index has the form

$$I_{F,\nu} = 1 - \frac{(\lambda-1)\nu}{(\nu\lambda-1)},$$

and the Atkinson index is

$$A_{F,a} = 1 - \left(1 - \frac{1}{\lambda}\right)\left(\frac{\lambda}{\lambda-a}\right)^{\frac{1}{a}}.$$

For the exponential distribution the probability density function

$$f(x) = \frac{e^{-\frac{x-x_0}{\lambda}}}{\lambda},$$

where $x > 0$ and $\lambda > 0$. The cumulative distribution function is

$$F(x) = 1 - e^{-\frac{x-x_0}{\lambda}}.$$

Lorenz curve for the exponential distribution is of the following form

$$L(p) = p + (1 + \lambda x_0)^{-1}(1-p)ln(1-p).$$

The S-Gini index has the form

$$I_{F,\nu} = \frac{\nu - 1}{\nu(1 + \lambda x_0)},$$

and the Atkinson index has the form

$$A_{F,a} = 1 - \Gamma(a+1)^{\frac{1}{a}}.$$

The standard Lognormal distribution has the probability density function

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{(ln(x)-x_0)^2}{2\sigma^2}},$$

where $x \geq x_0 > 0$ and $\sigma > 0$. The cumulative distribution function is of the form

$$F(x) = \Phi\left(\frac{ln(x - x_0) - \mu}{\sigma}\right).$$

The Lorenz curve for the Lognormal distribution is

$$L(p) = \frac{1}{e^{\mu+\sigma^2/2}} \int_0^z xf(x)dx = \Phi\left(\frac{lnz - \mu - \sigma^2}{\sigma}\right).$$

The S-Gini index has the form

$$I_{F,\nu} = 2\Phi(\sigma/\sqrt{2}) - 1,$$

and the Atkinson index has the form

$$A_{F,a} = 1 - \frac{(x_0 + e^{a\mu + \frac{a^2}{2}})^{\frac{1}{a}}}{x_0 + e^{\mu + \frac{1}{2}}}.$$

APPENDIX B

SIMULATIONS CODE USING R

The simulations for this dissertation are done using R. R is a language and environment for statistical computing and graphics. It is similar to the S language and environment which was developed at Bell Laboratories (formerly *AT&T*, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. R has some important differences, but much of the code written for S runs unaltered under R. One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulas where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control. It is a free software and runs on a wide variety of UNIX platforms and similar systems, including FreeBSD and Linux, Windows and MacOS. The current R is the result of a collaborative effort with contributions from all over the world. R was initially written by Robert Gentleman and Ross Ihakaalso known as "*R&R*" of the Statistics Department of the University of Auckland.

```
###1. PARETO DISTRIBUTION - ASYMPTOTIC THEORY -ONE POPULATION
 #Initialize the array's
Atkinsonn<-array(0,1000); Atkinsonpe<-array(0,1000);
sigma2ne<-array(0,1000); sigma2Fhat<-array(0,1000);
nlimitl<-array(0,1000); nlimitr<-array(0,1000);
plimitl<-array(0,1000); plimitr<-array(0,1000);

#Initialize the variables
ncv=0; pcv=0; ncvl=0; ncvr=0; pcvl=0;
pcvr=0; m=1000; zalpha=1.959964; n=10000; lambda=3;

for(k in 1:1000) {

X<-array((1/(1-runif(10000)))^(1/lambda), 10000);
a=0.2; med=0;
for(i in 1:n){med=med+X[i]}; med=med/n; mede=0;

for(i in 1:n){mede=mede+X[i]^a}; mede=mede/n;

#Atkinson Empirical
Atkinsonn[k]=1-(1/med)*(mede)^(1/a);

#AtkinsonParametric Pareto Atkinsonp=0;
Atkinsonp=1-(1-(1/lambda))*(lambda/(lambda-a))^(1/a);
sum=0;

for(i in 1:n){sum=sum+log(X[i])}; lambdahat=n/sum;
#Atkinson,estimator of the parametric
Atkinsonpe[k]=1-(1-(1/lambdahat))*
(lambdahat/(lambdahat-a))^(1/a);

muhat=(lambdahat)/(lambdahat-1);
muhata=(lambdahat)/(lambdahat-a);
muhat2a=(lambdahat)/(lambdahat-2*a);
muhata1=(lambdahat)/(lambdahat-a-1);
muhat2=(lambdahat)/(lambdahat-2); mu=(lambda)/(lambda-1);
mua=(lambda)/(lambda-a); mu2a=(lambda)/(lambda-2*a);
mua1=(lambda)/(lambda-a-1); mu2=(lambda)/(lambda-2);

#Variance for Nonparametric CI
sigma2n=(1/(a^2*mu^2))*(mua^((2-2*a)/a))* (mu2a-mua^2) -
2*((mua^((2-a)/a))/(a*mu^3))*(mua1-mu*mua)
+((mua^(2/a))/(mu^4))*(mu2-mu^2);
#Variance for Nonparametric CI
sigma2ne[k]=(1/(a^2*muhat^2))*(muhata^((2-2*a)/a))
*(muhat2a-muhata^2)-2*((muhata^((2-a)/a))/(a*muhat^3))
*(muhata1-muhat*muhata)
+((muhata^(2/a))/(muhat^4))*(muhat2-muhat^2);

#variance of the parametric CI sigma2F=0;
sigma2F=((1-a)^2*lambda^((2-2*a)/a))
/((lambda-a)**((2*a+2)/a));

sigma2Fhat[k]=((1-a)^2*lambdahat^((2-2*a)/a))
/((lambdahat-a)^((2*a+2)/a));

#confidence intervals nonparametric

nlimitl[k]=Atkinsonn[k] -zalpha*sqrt(sigma2ne[k]/n);
nlimitr[k]=Atkinsonn[k] +zalpha*sqrt(sigma2ne[k]/n);

#confidence intervals parametric

plimitl[k]=Atkinsonpe[k] -zalpha*sqrt(sigma2Fhat[k]/n);
plimitr[k]=Atkinsonpe[k] +zalpha*sqrt(sigma2Fhat[k]/n);

if(nlimitl[k]<Atkinsonp)
{ if(nlimitr[k]>Atkinsonp) ncv=ncv+1 else
 ncvr=ncvr+1} else ncvl=ncvl+1;

if(plimitl[k]<Atkinsonp)
{ if(plimitr[k]>Atkinsonp) pcv=pcv+1
 else pcvr=pcvr+1} else pcvl=pcvl+1; }

c(mean(Atkinsonn),var(Atkinsonn))
c(mean(Atkinsonpe),var(Atkinsonpe))

ncv=ncv/m; pcv=pcv/m; }
```

```
###2. EXPONENTIAL DISTRIBUTION - ASYMPTOTIC THEORY


Atkinsonn<-array(0,1000);

m=1000; alpha<-0.05; zalpha<-qnorm(1-alpha/2);

a=0.2; n=1000; b=0; theta=2; for(k in 1:1000) {
y<-rep(n,0);
y<-rexp(n,theta);
med=0; for(i in 1:n){med=med+y[i]};
med=med/n; mede=0;
for(i in 1:n){mede=mede+y[i]^a}; mede=mede/n;

#Atkinson Empirical
Atkinsonn[k]=1-(1/med)*(mede)^(1/a);
#AtkinsonParametric Exponential Atkinsonp=0; b=a+1;
Atkinsonp=1-(gamma(b))^(1/a); }
```

```
#3.LOGNORMAL DISTRIBUTION-ASYMPTOTIC THEORY-ONE POPULATION
 #Initialize the array's

Atkinsonn<-array(0,1000); AtkinsonpF<-array(0,1000);
sigma2ne<-array(0,1000); sigma2Fhat<-array(0,1000);
nlimitl<-array(0,1000); nlimitr<-array(0,1000);
plimitl<-array(0,1000); plimitr<-array(0,1000);

ncv=0; m=1000; alpha<-0.05; zalpha<-qnorm(1-alpha/2);


a=0.2; n=1000;
mun<-0.5; sigma<-1 locat=0.01;

for(k in 1:1000) {
y<-rep(n,0); y<-rnorm(n,mun,sigma); Z<-exp(y);
X<-locat+exp(y);

med=0; for(i in 1:n){med=med+X[i]}; med=med/n; mede=0;

for(i in :n){mede=mede+X[i]^a}; mede=mede/n;

#Atkinson Empirical
Atkinsonn[k]=1-(1/med)*(mede)^(1/a); sum1=0;
for(i in 1:n){sum1=sum1+log(Z[i])}; munhat=sum1/n;


#Atkinson, estimator of the parametric
AtkinsonpF[k]=1-((locat+exp(a*munhat + (a^2/2)))^(1/a))
/(locat+exp(munhat+1/2));

#Atkinson Parametric LOGNORMAL
Atkinsonp=0;

Atkinsonp=1-((locat+exp(a*mun+(a^2/2)))^(1/a))
/(locat+exp(mun+1/2));


m1=0; m1=med; m2=0;
for(i in 1:n){m2=m2+X[i]^2}; m2=m2/n; ma=0;
for(i in 1:n){ma=ma+X[i]^a}; ma=ma/n; ma1=0;

for(i in 1:n){ma1=ma1+X[i]^{a+1}}; ma1=ma1/n;

m2a=0; for(i in 1:n){m2a=m2a+X[i]^{2*a}}; m2a=m2a/n;
#Variance for Nonparametric CI

sigma2ne[k]=(1/(a^2*m1^2))* (ma^((2-2*a)/a))*
 (m2a-ma^2) -2*((ma^((2-a)/a))/(a*m1^3))*
 (ma1-m1*ma)+((ma^(2/a))/(m1^4))*(m2-m1^2);

#variance of the parametric CI
sigma2F=0;

sigma2F=((1- (1)/(locat+ exp(a*mun+ (a^2)/2)))^2)
*  (((locat + exp(a*mun+ (a^2)/2)) ^ (1/a))/
(locat+exp(mun+1/2)))^2;


sigma2Fhat[k]=((1- (1)/(locat+ exp(a*munhat+ (a^2)/2)))^2)*
(((locat + exp(a*munhat+
(a^2)/2))^(1/a))/(locat+exp(munhat+1/2)))^2;


#confidence intervals nonparametric

 nlimitl[k]=Atkinsonn[k]-zalpha*sqrt(sigma2ne[k]/n);

 nlimitr[k]=Atkinsonn[k]+zalpha*sqrt(sigma2ne[k]/n);

if((nlimitl[k]<Atkinsonp)&(Atkinsonp < nlimitr[k]))

ncv=ncv+1 else ncv=ncv; ncv=ncv/m; }
```

```
###4. PARETO DISTRIBUTION - NONPARAMETRIC BOOTSTRAP METHOD
 alpha<-0.05
 N<-10000
 R<-1000
 T<-10
 lambda<-3
 a<-0.2
prop1<-rep(0,R)


Atkinsonn<-rep(0,R)
Atkinsonnboot<-rep(0,R)
Quant<-rep(0,R)


CILL<-rep(0,R)
CIRR<-rep(0,R)

prop<-0
pp<-rep(0,T)

for(t in 1:T)  ################ Begin T loop
{

for(r in 1:R)  ################ Begin R loop
{
X<-(1/(1-runif(N)))^(1/lambda)

med=0;
for(i in 1:N){med=med+X[i]};
med=med/N;
mede=0;
 for(i in 1:N){mede=mede+X[i]^a}; mede=mede/N;

#Atkinson  Pareto
Atkinsonp=0;
Atkinsonp=1-(1-(1/lambda))*(lambda/(lambda-a))^(1/a);

#Atkinson index
Atkinsonn[r]=1-(1/med)*(mede)^(1/a);

index<-sample(1:N,N,replace=TRUE) X_boot<-X[index]

medb=0;
for(i in 1:N){medb=medb+X_boot[i]};
medb=medb/N;
medeb=0;
for(i in 1:N){medeb=medeb+X_boot[i]^a};
medeb=medeb/N;

#Atkinson index bootstrap
Atkinsonnboot[r]=1-(1/medb)*(medeb)^(1/a);

Quant[r]<-sqrt(N)* abs(Atkinsonnboot[r]-Atkinsonn[r])
}
############################ End R loop

Ord_Quant<-sort(Quant)
Ord_alpha<-trunc((1-alpha)*R)

criticalval<-Ord_Quant[Ord_alpha]
Q<-R

for(q in 1:Q)  ################ Begin Q loop
{

CILL[q]<-Atkinsonn[q] - criticalval/sqrt(N)

CIRR[q]<-Atkinsonn[q] + criticalval/sqrt(N)


if((Atkinsonp>CILL[q])& (Atkinsonp<CIRR[q])) prop=prop+1
}################ End Q loop
pp[t]<-prop/Q prop<-0
}################ End T loop
mu<-mean(pp)
```

```
###5. PARETO DISTRIBUTION - NONPARAMETRIC BOOTSTRAP METHOD

 alpha<-0.05
 N<-10000
 T<-10
 R<-1000
 lambda<-3
 a<-0.2
 prop1<-rep(0,R)
 Atkinsonpe<-rep(0,R)
 Atkinsonpeboot<-rep(0,R)
 Quant<-rep(0,R)

CILL<-rep(0,R)
CIRR<-rep(0,R)
prop<-0
pp<-rep(0,T)

for(t in 1:T)  ################ Begin T loop
{

for(r in 1:R) ################ Begin R loop {
X<-(1/(1-runif(N)))^(1/lambda)

#Atkinson Parametric Pareto
Atkinsonp=0;
Atkinsonp=1-(1-(1/lambda))*(lambda/(lambda-a))^(1/a);

sum=0;
for(i in 1:N){sum=sum+log(X[i])};
lambdahat=N/sum;

#Atkinson, estimator of the parametric
Atkinsonpe[r]=1-(1-(1/lambdahat))*
(lambdahat/(lambdahat-a))^(1/a);

index<-sample(1:N,N,replace=TRUE) X_boot<-X[index]
sum=0;
for(i in 1:N){sum=sum+log(X_boot[i])};
lambdahatboot=N/sum;

#Atkinson index bootstrap
Atkinsonpeboot[r]=1-(1-(1/lambdahatboot))*
(lambdahatboot/(lambdahatboot-a))^(1/a);
Quant[r]<-sqrt(N)* abs(Atkinsonpeboot[r]-Atkinsonpe[r])

} ############################ End R loop

Ord_Quant<-sort(Quant)
Ord_alpha<-trunc((1-alpha)*R)

criticalval<-Ord_Quant[Ord_alpha]
Q<-R
for(q in 1:Q)
################ Begin Q loop
{

CILL[q]<-Atkinsonpe[q] - criticalval/sqrt(N)
CIRR[q]<-Atkinsonpe[q] + criticalval/sqrt(N)

if((Atkinsonp>CILL[q])& (Atkinsonp<CIRR[q])) prop=prop+1
}############################ End Q loop
pp[t]<-prop/Q prop<-0

} ################ End T loop
mu<-mean(pp)
```

```
###6. EXPONENTIAL DISTRIBUTION-NONPARAMETRIC BOOTSTRAP METHOD
alpha<-0.05 N<-10000 R<-1000 T<-10 theta<-2 a<-0.2 b<-0
prop1<-rep(0,R)

Atkinsonn<-rep(0,R)
Atkinsonnboot<-rep(0,R)
Quant<-rep(0,R)

CILL<-rep(0,R)
CIRR<-rep(0,R)

prop<-0
pp<-rep(0,T)

for(t in 1:T)  ################ Begin T loop
{

for(r in 1:R)  ################ Begin R loop
 {
 X<-rexp(N,theta);

med=0;
for(i in 1:N){med=med+X[i]};
med=med/N;
mede=0;
for(i in1:N){mede=mede+X[i]^a};
mede=mede/N;

#Atkinson Parametric Pareto
Atkinsonp=0;
b=a+1;
 Atkinsonp= 1 -(gamma(b))^(1/a);

#Atkinson index
Atkinsonn[r]=1-(1/med)*(mede)^(1/a);

index<-sample(1:N,N,replace=TRUE) X_boot<-X[index]

medb=0; for(i in 1:N){medb=medb+X_boot[i]};
medb=medb/N;
medeb=0;
for(i in 1:N){medeb=medeb+X_boot[i]^a};
medeb=medeb/N;
#Atkinson index bootstrap
Atkinsonnboot[r]=1-(1/medb)*(medeb)^(1/a);

Quant[r]<-sqrt(N)* abs(Atkinsonnboot[r]-Atkinsonn[r])

} ############################ End R loop

Ord_Quant<-sort(Quant)
Ord_alpha<-trunc((1-alpha)*R)
criticalval<-Ord_Quant[Ord_alpha]

Q<-R
for(q in 1:Q)  ################ Begin Q loop
{

CILL[q]<-Atkinsonn[q] - criticalval/sqrt(N)

CIRR[q]<-Atkinsonn[q] + criticalval/sqrt(N)


if((Atkinsonp>CILL[q])& (Atkinsonp<CIRR[q])) prop=prop+1
}################ End Q loop

pp[t]<-prop/Q prop<-0 } ################ End T loop

mu<-mean(pp)
```

```
###7. Two independent samples, same size,
 asymptotic parametric, Pareto and Pareto

#Initialize the variables
P<-1000 Q<-1000
T<-5 m<-P n<-P mu<-0

alpha<-0.05
lambda_m<-8 lambda_n<-10

a<-0.3 zalpha<-qnorm(1-alpha/2)

#Initialize the indices
Atknpe_n <-rep(0,Q)
Atkpe_n <-rep(0,Q)

Atknpe_m <-rep(0,Q)
Atkpe_m <-rep(0,Q)

#Initialize the confidence limits
nlimitl<-rep(0,Q)
nlimitr<-rep(0,Q)
plimitl<-rep(0,Q)
plimitr<-rep(0,Q)


#Initialize the variances
sigma2n_m<-0 sigma2n_n<-0
sigma2ne_m<-rep(0,Q)
sigma2ne_n <-rep(0,Q)

sigma2F_m<-0 sigma2F_n<-0
sigma2Fm<-rep(0,Q) sigma2Fn<-rep(0,Q)

prop<-0 pp<-rep(0,T)

for(t in 1:T)  ################ Begin T loop {

for(q in 1:Q)################# Begin Q loop {

############Population X1, X2, ...Xm
X<-array((1/(1-runif(P)))^(1/lambda_m), P)

medm=0; for(i in 1:P){medm=medm+X[i]};
medm=medm/P; medem=0;
```

```
for(i in 1:P){medem=medem+X[i]^a}; medem=medem/P;

############Population Y1, Y2, ...Yn
Y<-array((1/(1-runif(P)))^(1/lambda_n), P)

medn=0; for(i in 1:P){medn=medn+Y[i]}; medn=medn/P;

meden=0; for(i in 1:P){meden=meden+Y[i]^a}; meden=meden/P;

#Nonparametric Atkinson Estimator
Atknpe_m[q]=1-(1/medm)*(medem)^(1/a);
Atknpe_n[q]=1-(1/medn)*(meden)^(1/a);

#Parametric Atkinson Index and MLE
 Atkinsonpm=0;
Atkinsonpm=1-(1-(1/lambda_m))*(lambda_m/(lambda_m-a))^(1/a);
summ=0;
for(i in 1:P){summ=summ+log(X[i])}; lambdahatm=P/summ;

Atkinsonpn=0;
Atkinsonpn=1-(1-(1/lambda_n))*(lambda_n/(lambda_n-a))^(1/a);
sumn=0;
for(i in 1:P){sumn=sumn+log(Y[i])}; lambdahatn=P/sumn;

#Parametric Atkinson estimator

Atkpe_m[q]=1-(1-(1/lambdahatm))*
(lambdahatm/(lambdahatm-a))^(1/a);

Atkpe_n[q]=1-(1-(1/lambdahatn))*
(lambdahatn/(lambdahatn-a))^(1/a);

####################### Population X and Y Asymptotic

#variance of the parametric CI sigma2F

sigma2F_m= ((1-a)^2 * lambda_m)  /
(((1+ lambda_m*a - lambda_m)^2)
*(lambda_m-2*lambda_m*a-2));

sigma2Fm[q]=((1-a)^2 * lambdahatm)  /
 (((1+ lambdahatm*a - lambdahatm)^2) *
 (lambdahatm-2*lambdahatm*a-2));
```

```
sigma2F_n=((1-a)^2 * lambda_n)  /
 ((1+ lambda_n*a - lambda_n)^2)
*(lambda_n-2*lambda_n*a-2));

sigma2Fn[q]=((1-a)^2 * lambdahatn)  /
 (((1+ lambdahatn*a - lambdahatn)^2) *
 (lambdahatn-2*lambdahatn*a-2));


##### Parametric Asymptotic Confidence Intervals
 for independent populations

nlimitl[q]= Atkpe_m[q] - Atkpe_n[q] -
 zalpha* sqrt (sigma2Fm[q]/m + sigma2Fn[q]/n);

nlimitr[q]= Atkpe_m[q] - Atkpe_n[q] +
 zalpha* sqrt (sigma2Fm[q]/m + sigma2Fn[q]/n);


if(((Atkinsonpm-Atkinsonpn) >nlimitl[q])  &
((Atkinsonpm-Atkinsonpn) < nlimitr[q])) prop=prop+1


}###################End loop Q
pp[t]<-prop/Q prop<-0

} ################ End T loop
mu<-mean(pp)
```

```
###8. Two independent samples of different size,
 nonparametric bootstrap, Pareto and Exponential

alpha<-0.05 N<-1000 M<-1050 T<-10 R<-1000

lambda<-3 theta<-5 a<-0.2

Atkinsonn1<-rep(0,R) Atkinsonnboot1<-rep(0,R)
Quant<-rep(0,R)

Atkinsonn2<-rep(0,R) Atkinsonnboot2<-rep(0,R)
Quant2<-rep(0,R)

prop<-0 pp<-rep(0,T)

CILL<-rep(0,R) CIRR<-rep(0,R)

for(t in 1:T)  ################ Begin T loop {

for(r in 1:R)  ################ Begin R loop {

X<-(1/(1-runif(N)))^(1/lambda);
 ### X follows a Pareto(3)
Y<-rexp(M,theta);
### Y follows an Exponential(5)

#Theoretic Atkinson Pareto Atkinsonp1=0;
Atkinsonp1=1-(1-(1/lambda))*(lambda/(lambda-a))^(1/a);

#Theoretic Atkinson Exponential

Atkinsonp2=0; b=a+1; Atkinsonp2= 1 -(gamma(b))^(1/a);

####Nonparametric Atkinson estimates, for population X
med=0;

for(i in 1:N){med=med+X[i]}; med=med/N; mede=0;

for(i in 1:N){mede=mede+X[i]^a}; mede=mede/N;

Atkinsonn1[r]=1-(1/med)*(mede)^(1/a);

####Nonparametric Atkinson estimates, for population Y
med=0;
for(i in 1:M){med=med+Y[i]}; med=med/M; mede=0;
for(i in 1:M){mede=mede+Y[i]^a}; mede=mede/M;

Atkinsonn2[r]=1-(1/med)*(mede)^(1/a);

index1<-sample(1:N,N,replace=TRUE) X_boot<-X[index1]
index2<-sample(1:M,M,replace=TRUE) Y_boot<-Y[index2]

##Nonparametric Atkinson Estimates, bootstrap, population X
med=0;
for(i in 1:N){med=med+X_boot[i]}; med=med/N; mede=0;

for(i in 1:N){mede=mede+X_boot[i]^a}; mede=mede/N;

Atkinsonnboot1[r]=1-(1/med)*(mede)^(1/a);

##Nonparametric Atkinson Estimates, bootstrap, population Y
med=0;
for(i in 1:M){med=med+Y_boot[i]}; med=med/M; mede=0;

for(i in 1:M){mede=mede+Y_boot[i]^a}; mede=mede/M;

Atkinsonnboot2[r]=1-(1/med)*(mede)^(1/a);

Quant[r]<-sqrt((N+M)/(N*M))*
abs((Atkinsonnboot1[r]-Atkinsonnboot2[r])-
(Atkinsonn1[r]-Atkinsonn2[r]))

} ############################ End R loop

Ord_Quant<-sort(Quant)
Ord_alpha<-trunc((1-alpha)*R)
 criticalval<-Ord_Quant[Ord_alpha]
Q<-R

for(q in 1:Q)  ############### Begin Q loop {

CILL[q]<-(Atkinsonn1[q] - Atkinsonn2[q])-
criticalval*sqrt((N+M)/(N*M))

CIRR[q]<-(Atkinsonn1[q] - Atkinsonn2[q]) +
criticalval*sqrt((N+M)/(N*M))

if(  ((Atkinsonp1 -Atkinsonp2) >CILL[q])  &
((Atkinsonp1-Atkinsonp2) < CIRR[q]) ) prop=prop+1
```

```
}############################ End Q loop

pp[t]<-prop/Q prop<-0

} ################ End T loop

mu<-mean(pp)
```

11

BIBLIOGRAPHY

[Atkinson, 1970] A. B. Atkinson, On the Measurement of Inequality, *Journal of Economic Theory* 2, (1970), 244-263.

[Barrett and Donald, 2001] G.F. Barrett and S.G. Donald, Statistical inference with generalized Gini indices of inequality and poverty, *Discussion Paper 2002/01.* School of Economics. University of New South Wales, (2000).

[Chakravarty, 1988] S. R. Chakravarty, Extended Gini indices of inequality, *International Economic Review* 29, (1988), 147156.

[Cowell, 1998] F.Cowell, C. Schluter, Income Mobility: A Robust Approach, *Income Inequality Measurement: From Theory to Practice*, (1998), 37-71.

[Dalton, 1920] H. Dalton, The Measurement of the Inequality of Incomes, *Economic Journal* 30, (1920), 348-361.

[Dasgupta et al., 1973] P. Dasgupta, A. Sen and D.A. Starrett, Notes on the Measurement of Inequality, *Journal of Economic Theory* 6(2), (1973), 180-187.

[Donaldson, 1980] D. Donaldson and J.A. Weymark, A single-parameter generalization of the Gini indices of inequality, *Journal of Economic Theory* 22, (1980), 67-86.

[Gastwirth, 2002] J.L. Gastwirth and R. Zitikis, Asymptotic distribution of the S-Gini index, *Australian and New Zealand Journal of Statistics* 44, (2002), 439-446.

[Golan et al., 2001] A. Golan, J. M. Perloff, X. Wu, Welfare Effects of Minimum Wage and Other Government Policies, Department of Agricultural and Resource Economics, (University of California, Berkeley) Working Paper 957, (2001), 1-47.

[Gusenleitner et al., 1998] M. Gusenleitner, R. Winter-Ebmer, J. Zweimller, The Distribution of Earnings in Austria 1972-1991, *Allgemeines Statistisches Archiv* 82(3), (1998), 275-290.

[Hoeffding, 1948] W. Hoeffding, A class of statistics with assymptotically normal distribution, *Annals of Mathematical Statistics* 19, (1948), 293-325.

[Horvath et al., 2005] L. Horvath, P. Kokoszka and R. Zitikis, Testing for Stochastic Dominance Using the Weighted McFadden Type Statistics, *Journal of Econometrics*, (2005), 1-19.

[Jones and Zitikis, 2003] B.L. Jones, R. Zitikis, Empirical estimation of risk measures and related quantities, *North American Actuarial Journal* 7(4), (2003), 44-54.

[Kaplow, 2005] L. Kaplow, Why Measure Inequality?, *Journal of Economic Inequality* 3(1), (2005), 65-79.

[Londoo, 2000] J. Londoo, M. Szkely, Persistent Poverty and Excess Inequality: Latin America, 1970-1995, *Journal of Applied Economics, Vol. III* 1, (2000), 93-134.

[Lovell, 1998] M. C. Lovell, Inequality within and among nations, *Journal of Income Distribution* 8(1), (1998), 5-44.

[Mayer, 2000] S. Mayer, How Did the Increase in Economic Inequality between 1970 and 1990 Affect American Children's Educational Attainment?, *American Journal of Sociology* 107(1), (2000), 1-32.

[Puri, 1965] M.L. Puri, Some distribution-free k-sample rank tests of homogeneity against ordered alternatives, *Communications on Pure and Applied Mathematics* XVIII, (1965), 51-63.

[Ravallion, 1997] M. Ravallion, M. Heil, J. Jalan, A less poor world, but a hotter one? Carbon emissions, economic growth and income inequality, *World Bank*October, (1997), 1-31.

[Shorack, 1967] G.R. Shorack, Testing against ordered alternatives in model I analysis of variance: Normal Theory and nonparametric, *Ann. Math. Statistics* 38, (1967), 1740-1752.

[Weymark, 1981] J.A. Weymark, Generalized Gini inequality indices, *Mathematical Social Sciences* 1, (1981), 409-430.

[Zitikis, 2002] R. Zitikis, Large sample estimation of a family of economic inequality indices, *Pakistan Journal of Statistics (Special Issue in Honour of Dr. S. Ejaz Ahmed)* 18, (2002), 225-248.

[Zitikis, 2003] R. Zitikis, Asymptotic estimation of the E-Gini index, *Econometric Theory* 19, (2003), 587-601.