Springer Series in Statistics

The Gini Methodology

A Primer on a Statistical Methodology

Springer

Springer Series in Statistics

For further volumes:
http://www.springer.com/series/692

Shlomo Yitzhaki • Edna Schechtman

The Gini Methodology

A Primer on a Statistical Methodology

Springer

Shlomo Yitzhaki
Department of Economics
The Hebrew University
Mount Scopus, Jerusalem
Israel

Edna Schechtman
Department of Industrial Engineering
and Management
Ben-Gurion University of the Negev
Beer-Sheva, Israel

ISSN 0172-7397
ISBN 978-1-4614-4719-1 ISBN 978-1-4614-4720-7 (eBook)
DOI 10.1007/978-1-4614-4720-7
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2012946727

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To Ruhama, Guy, and Nili who lived under the shadow of Gini, and Ido, Ella, and Roni who are a lovable distraction.

Shlomo

To Gideon

Edna

Acknowledgement

This book covers research that was carried out for about 40 years. However, it is clear to us that the task of investigating and implementing the Gini methodology is far from being completed. Many researchers, some of them anonymous referees, have contributed ideas that shaped the development of the ideas in this book. In some of the cases we can identify and acknowledge their contributions.

The interest of the first author in the Gini coefficient was when trying to convince academic members of the Ben-Shahar committee for tax reform in Israel, 1975, to belittle the recommendation of economic theory which claims that increasing the penalties is a sufficient instrument for decreasing tax evasion. The major argument was that if a crime is committed by many people then it is not considered as a crime. It was Martin Feldstein who suggested searching for an argument that is not related to the classical utility theory, which led to the connection between the Gini and the relative deprivation theory.

Ingram Olkin should be credited for the idea that one should use concentration curves instead of parameters to investigate the relationships between variables, an idea that led to concepts like marginal conditional stochastic dominance and inspecting whether relationships between variables are monotonic, a research agenda which is still in its initial stages.

Collaborations with other researchers brought into this book the specialization and the intimate knowledge of specific areas that helped spread the Gini methodology. The late Joachim Frick, Yoel Finkel, Jan Goebel, Bob Lerman, Joram Mayshar, Branko Milanovic, Joel Slemrod, Dan Slottje, Wayne Thrisk, Gert Wagner, and Quentin Wodon were influential in shaping the implementations in the areas of public finance and income distribution, while Haim Shalit was in charge of the area of finance, Oded Stark and Ed Taylor shaped the implementation in demography, and Manuel Trajtenberg in diffusion processes. Joseph Heller introduced us to the fascinating world of zoology. Yevgeny Artsev, Yolanda Golan, Vika Roshal, Taina Pudalov, and Amit Shelef assisted in implementing the Gini methodology, while Alexandra Katzenelbogen and Tamir Erez programmed the computer programs used in the book. The mathematical skills of Peter Lambert and Gideon Schechtman were crucial in overcoming some obstacles, while Joel Slemrod was the reader and

advisor of many papers. Correspondence with Jim Heckman and conversations with Josh Angrist were helpful in understanding the issues in Econometrics.

Over 25 years of cooperation between the two authors has led to the development of the statistical background of the Gini methodology, while Peter Lambert and Gideon Schechtman have helped in overcoming some mathematical difficulties. The reading and advice of Joel Slemrod provided the intellectual support needed to overcome some of the difficulties in the understanding and the exposition of the theory.

The idea of writing the book was initiated by John Kimmel, while Hillel Bar Gera, Michael Beenstock, Ingram Olkin, Peter Lambert, David Johnson, and Christian Toft kindly commented on the final draft. The tedious task of bringing the book into its final form is carried out by Hannah Bracken under the supervision of Marc Strauss but all the mistakes are ours.

Contents

Chapter 1
Introduction

Gini's mean difference (hereafter, GMD) was first introduced by Corrado Gini in 1912 as an alternative measure of variability. GMD and the parameters which are derived from it (such as the Gini coefficient, also referred to as the concentration ratio) have been in use in the area of income distribution for almost a century, and there is evidence that the GMD was introduced even earlier (Harter, 1978). In other areas it seems to make sporadic appearances and to be "rediscovered" again and again under different names. It turns out that GMD has at least 14 different alternative representations. Each representation can be given its own interpretation and naturally leads to a different analytical tool such as L_1 metric, order statistics theory, extreme value theory, concentration curves, and more. Some of the representations hold only for nonnegative variables while others need adjustments for handling discrete distributions. On top of that, the GMD was developed in different areas and in different languages. Corrado Gini himself mentioned this difficulty (Gini, 1921). Therefore in many cases even an experienced expert in the area may fail to identify a Gini when he or she sees one.

Covering all the approaches in detail can become tiresome and possibly uninteresting. Therefore, in order to overcome this "curse of the plenty" we set one target in mind. We shall focus on imitating the analyses that are based on the variance by replacing the variance by the GMD and its variants. We intend to show that almost everything that can be done with the variance as a measure of variability can be replicated by using the Gini. With this target in mind we will mainly focus our attention on one representation—the covariance-based approach—and limit the coverage of other approaches.

The use of GMD as a measure of variability is justified whenever the investigator is not ready to impose, without questioning, the convenient world of normality. When the underlying distribution is univariate and normal, the sample mean and variance are sufficient statistics to describe it and the GMD is redundant. Likewise, when dealing with multivariate distributions, the case of multivariate normality is fully described by the individual means, the individual variances, and Pearson's correlation coefficients. The GMD and the equivalents of the correlation coefficient have nothing to add to the understanding of the data, nor to the analysis. However

S. Yitzhaki and E. Schechtman, *The Gini Methodology: A Primer on a Statistical Methodology*, Springer Series in Statistics 272, DOI 10.1007/978-1-4614-4720-7_1,

when the distribution is not multivariate normal, then, as Lambert and Decoster (2005) put it—the GMD reveals more! As will be demonstrated in this book, it reveals whether the relationships between random variables (as described by the covariance and by the correlation) are symmetric or not, whether the population is stratified and to what extent, whether the assumption of linearity in regression analysis is supported by the data, and more. The use of GMD may add insight and understanding of the data at hand. For example, it can be used whenever one wants to see if the assumption that the underlying distribution is multivariate normal holds, or if the regression model is truly linear. However it comes with a price tag on it. It turns out that using the GMD as a substitute for the variance implies that the number of economic models is doubled because every variance-based model will now have beside it a Gini-based model that may give different results. We will show that many of the properties of the variance-based models are included as special cases of Gini-based models. As a result, we argue that if the estimates of the variance-based and Gini-based models are close to each other then we obtain reassurance that the model is robust in the sense that it is not sensitive to the implicit assumptions imposed on the data by treating it as if the underlying distribution is multivariate normal. On the other hand, if the estimates differ then it is an indication that the implicit assumptions of the variance-based model are responsible for the deviation. As far as we can see, in many cases using the Gini methodology in addition to the variance-based methods will lead to a reduction in the number of possibilities of generating "empirical proofs" that support the researcher's theory but are not supported by the data itself.

This book is a first attempt to present the family of parameters based on GMD and to illustrate its applications in different areas, mainly in the areas of economics and statistics. The main thrust is to "translate" the commonly used analyses based on the variance and the parameters based on it into the Gini world. Parameters such as the covariance and Pearson's correlation coefficient, as well as methodologies such as ANOVA and Ordinary Least Squares (OLS) regressions, are "translated" where the variance is replaced by (the square of) the GMD, the covariance and Pearson's correlation are replaced by Gini covariance and Gini correlation, ANOVA is changed to ANalysis Of GIni (ANOGI), and OLS regression is replaced by Gini regression. As will be shown, the above "translation" gives rise to additional parameters and the alternative approach reveals more when the underlying distribution deviates from the multivariate normal. The slogan of this book is "(almost) everything you can do (with the variance), we can do better (with the Gini)." By "doing better" we mean that the approach offers richer tools for statistical analyses and that the additional parameters that the Gini method offers enable the researcher to adjust the statistical analysis to the needs of the area of research. We argue that the convenience of the assumption of multivariate normality could blur some of the issues that are relevant in several areas of research such as risk analysis, income distribution, economics, and sociology. It should be stated that the task of "translating" the variance world into the Gini regime is not yet completed. In some sense we feel that we are touching the tip of the iceberg and plenty of

additional theoretical work as well as user-friendly software are called for to fully utilize the Gini methodology as an analytical tool.

One of the advantages of using the Gini methodology is that it provides a unified system that enables the user to learn about various aspects of the underlying distribution. Almost every property of the underlying distribution that the Gini method enables us to present or test can also be described and tested using other approaches but the advantage here is that we provide a systematic method and a unified terminology.

Let us illustrate this point. Consider the methodology of estimating and verifying a simple linear regression model. The Gini methodology enables the user to estimate the regression coefficient, draw inferences on it, check whether the model is linear, and verify that the residuals are normally distributed—all under one systematic method.

The variance is the most popular measure of variability. There are two properties which seem natural and are implicit when dealing with the variance: the symmetry and the decomposition (to be detailed below). The Gini approach deviates from this conventional (and convenient) approach. Understanding these two points will make the ideas that are stressed in the book easier to follow.

(a) Symmetric relationship: There are two kinds of symmetric relationships that are imposed in the conventional statistical analysis in general but are not followed in this book. The first one is the symmetry of the variability measure with respect to the underlying distribution and the second one is the symmetry in the relationship between variables. The first symmetry can be described as requiring that the variability of X will be equal to the variability of $(-X)$. The justification for not following this kind of symmetry is because some of the subject matters that we are dealing with such as the areas of risk and income distributions are governed by theories that call for asymmetric treatments of the distributions. This issue is handled in Chap. 6 which presents the extended Gini and in Chaps. 13, 17, and 19 which present applications in the areas of welfare economics and finance and in econometrics. The second deviation from the symmetry properties is concerned with the treatment of the relationship between two random variables. Most measures of association are symmetric with respect to the two variables, as is the case in $\mathrm{cov}(X, Y) = \mathrm{cov}(Y, X)$, even if the underlying bivariate distribution is not symmetric. Symmetry between random variables is a convenient property to have, but it comes with a price tag. The price paid is in the value imposed on the correlation. To see this consider two normally distributed random variables X and Y with a Pearson correlation coefficient of (-1). A researcher not knowing what the underlying distribution is may decide to use the exponential transformation to get e^X and e^Y changing the distributions to be lognormal. By doing that, the researcher inadvertently reduces the Pearson correlation coefficient to -0.36 (De Veaux, 1976). We argue that the change of the Pearson correlation from (-1) to (-0.36) should be attributed to the symmetry imposed by the covariance. We can also reverse the example. By taking the natural logarithm of two lognormally distributed

random variables with a Pearson correlation coefficient of (−0.36) the researcher changes the Pearson correlation coefficient to (−1). The above example is a bit extreme and only rarely occurs in practice. Consider a more plausible story. Given two normally distributed random variables with a Pearson correlation of 1, a researcher transforms one of them into a binary variable, a procedure intended to describe the participating/nonparticipating dichotomy. This is a common practice when applying Instrumental Variable procedure in regression analysis. In this case the researcher reduces the Pearson correlation from 1 to 0.8. The conclusion from these two examples is that a transformation can change the correlation, enabling the researcher to change the conclusion of the research. The Gini approach offers a remedy to this problem. There are two correlation coefficients defined between each pair of variables. These two correlations are equal if the distributions are exchangeable up to a linear transformation, which we will refer to as symmetric relationships. Applying a transformation to a variable changes only one (Gini) correlation coefficient leaving the other intact. Hence the difference in the correlations enables one to see the vulnerability of the correlation. This issue will be dealt with in Chaps. 3, 4, 8, 18, and 19.

There are at least two other major applications of having two (Gini) correlations between each pair of variables instead of one. First, as will be shown in Chap. 8, every optimization results in first order conditions that can be described as "orthogonality conditions." Those conditions can be interpreted as setting a covariance to zero. Having two correlations (and covariances) between two variables enables one to test whether the other covariance is also equal to zero so that one can have a specification test with respect to the underlying model. The second application is related to the properties of the decomposition of Gini of a linear combination of random variables as is discussed next.

(b) Decompositions: There are two types of decompositions. One is the decomposition of a variability measure of a linear combination of random variables into the contributions of the individual variables and the contributions of the relationships between them. The other decomposition is the one that decomposes the variability of a population that is composed of several subpopulations into the contributions of the subpopulations and some extra terms. In both cases the decomposition of the GMD includes the structure of the decomposition of the variance as a special case. We refer to the assumptions that lead to the structure of the decomposition of the variance as hidden assumptions imposed on the data that lead to the simplicity of the structure of the variance decompositions. We refer to this property as the property of "revealing more." The Gini of a linear combination of random variables does not, in general, decompose into two components as neatly as the variance does. (In the variance decomposition one component is based on the individual variances and the other is based on the correlations among the variables.) Instead, it extracts more information about the underlying distributions, as will be discussed in Chaps. 3 and 23.

The decomposition of the Gini leads, under certain conditions, to a decomposition formula with the same structure as the decomposition of the variance. This fact enables us to test for the hidden (implicit) assumptions that are leading to the simplicity that has made the variance-based analysis so convenient. More specifically, the Gini of a population does not decompose neatly (i.e., additively) into intra- and inter-group Ginis. For this reason it was rejected by several economists who tried to imitate Analysis of Variance. As will be shown in this book, this disadvantage may turn into an advantage. The decomposition of the Gini coefficient of a population extracts more information from the underlying distribution than just the inter- and intra-components. It gives a quantitative measure of the amount of the overlapping between the subgroups which is important whenever one is interested in stratification and/or in evaluating the quality of the classification of a general population into groups. The decomposition will be discussed in Chap. 4 while the empirical applications will be demonstrated in Chaps. 13 and 22.

The usefulness of the GMD and its contribution to our statistical analysis is especially important whenever the concepts that are used are not symmetric by definition. Among those concepts are regression in statistics and elasticity in economics. The properties of the Gini enable one to check the validity of the implicit use of symmetry whenever those concepts are used. In the regression concept the use of the Gini plays an important role in checking whether the assumptions that led to the estimates are supported by the data or not. For example, the Gini methodology can be used to check whether the relationship between Y and X is monotonic over the entire range of X or not by a simple graphical technique. This will be demonstrated in Chaps. 5, 19, and 21.

Having listed these advantages of using the Gini, it is worth mentioning the "cost" of using it. First, its use is cumbersome because sometimes the additional information that the Gini offers may be redundant. Second, in order to use the Gini one has to ignore some of the intuition and conventional wisdom that come with the variance. As will be shown, the Gini describes the variability by two attributes: the variate and its rank. For the economist, this should resemble the intuition that comes with what is known as "the index number problem" that is taught in intermediate economic theory. The index number problem arises whenever one tries to describe a phenomenon by two attributes: the price and the quantity of a commodity. In these cases one attribute is kept unchanged, while the other is allowed to change. Because in real life the two attributes can change simultaneously, the choice of which attribute is held constant and which one is allowed to change may result in some cases in contradicting conclusions. The cases of contradictions are the cases that diverge from the analysis based on the variance, and remembering them may help in understanding the intuition needed for evaluating the results.

An alternative approach to be taken when reading this book is to view the GMD and the parameters that are related to it as representing several theories that originate in the social sciences. Among these theories are (a) the expected utility hypothesis which represents the main paradigm in the area of risk and social welfare, (b) the relative deprivation theory which plays a major role in explaining social unrest, and (c) mobility, horizontal equity, and similar concepts that are used

in the social sciences. In this respect the book presents the essence of these theories and advocates the use of the decomposition properties of the Gini so that one can offer statistical tools for understanding, analyzing, and developing these theories. These theories and the relationships with the Gini are presented in the applications part of this book.

To be able to fully utilize the properties of the Gini we will not make any assumptions concerning the distribution of the random variable throughout this book. The only case in which we will assume a particular distribution is to illustrate a point.

Finally, we wish to add an apology. Darling, in his Annals of Mathematical Statistics paper (1957), writes:

> The reader is advised that the relative amount of space and emphasis allotted to the various phases of the subject do not reflect necessarily their intrinsic merit and importance, but rather the author's personal interest and familiarity. Also, for the sake of uniformity the notation of many of the writers quoted has been altered so that when referring to the original papers it will be necessary to check their nomenclature (Darling, 1957, p. 823).

We could not find better words for describing our approach in this book. Also, we apologize in advance for not giving the appropriate credit to the appropriate authors in some occasions. One serious difficulty is to define the meaning of an innovation and to decide to whom to give the credit for it in this area of research. The reason is that on top of independent developments, where researchers could not read the language or were not aware of the developments in their or other areas, there is a difficult issue in this crowded area. Is the person who wrote a formula in passing the one who should be credited for it, or is it the person who correctly interpreted it and developed its implications? In order to illustrate this issue let us investigate the history of expressing the GMD as a covariance. The fact that one can express GMD as a covariance and use the covariance properties to further develop the theoretical aspects is in our opinion a major breakthrough.

As far as we know, the first step in this direction was to write the Gini as a covariance without noticing that it actually *is* a covariance. This was done by Corrado Gini (1914). The next step, some 40 years later, was to realize that Gini can be expressed as a covariance, with no further implications. This fact was realized by Stuart (1954). Fei, Ranis, and Kou (1978) constructed the Gini-covariance, referring to it as pseudo-Gini. Pyatt, Chen, and Fei (1980) used the term covariance in constructing the pseudo-Gini. The final breakthrough was made by Lerman and Yitzhaki (1984) who pointed out that because the Gini can be expressed as a covariance it is possible and helpful to use the properties of the covariance in handling it. This observation opened the way to investigating the Gini covariances and correlations and their properties.

On the anecdotal side, the person who triggered Lerman and Yitzhaki to write the Gini as a covariance was an anonymous referee of Yitzhaki (1982a). He/she argued that the covariance is more important than the variance in the area of finance, and therefore a sentence should be added to say whether it is possible to develop a covariance that is suitable for the Gini or not.

Similar issues arose concerning the development of the extended Gini which was discovered independently and from different angles by Donaldson and Weymark (1980, 1983) and Yitzhaki (1983). Kakwani (1980) mentions the possibility of the extended Gini in passing. Moreover, all the above-mentioned papers and Chakravarty (1983) can be classified as the Gini response to Atkinson (1970) who suggested an inequality measure that depends on a parameter. Clearly, there can be other scenarios for describing the development of the extended Gini and the expression of the Gini as a covariance. In order not to enter into such a debate, we apologize in advance for not taking the appropriate actions to attribute each concept to the original person who developed and coined it. In addition, in order to keep the presentation flowing, and to avoid sidetracking the reader into what we consider as dead end from the point of view of our target, some papers that may be important in the future are only mentioned in passing.

The target audience of this book is mainly applied economists, statisticians, and econometricians who are interested in applications for which the variance is not suitable. These applications arise mostly (but not only) when the underlying distribution deviates from normality. Possible areas of application are welfare economics, finance, and general econometric theory. As will be seen in this book, Gini-based analyses are robust to the asymmetry of the distribution and to the existence of outliers. In addition, the use of the Gini allows one to identify and test the existence of implicit assumptions about the underlying distributions that make the variance-based analyses so simple to apply, yet may not be satisfied by the data, or, alternatively, violate basic principles of economic theory.

The complexity and the different representations and applications of the GMD in different fields forced us to use different notations to represent the GMD in different areas. The reason is that in some areas it is convenient to use GMD/4 as the GMD, and in other areas GMD/2 or simply GMD. This implies the need to carry constants that affect all equations in a specific application and complicate the representation without adding any content. To overcome this problem, we use different representations of the GMD in different chapters and we will state in the introduction of each chapter which definition is used.

The book consists of two main parts. The first part (Chaps. 2–11) contains the theory while the second part (Chaps. 12–22) deals with applications. The applications chapters contain a short review of the needed theory to make them readable on their own. The structure of the book is the following: In Chap. 2 we provide the various definitions of the Gini. The Gini covariance, correlation, and regressions are introduced in Chap. 3. In Chap. 4 we present the decompositions of the Gini while Chap. 5 deals with the relation to the Lorenz and the concentration curves. The extended Gini family of measures is introduced in Chap. 6. Next, two chapters are devoted to Gini regression: the simple regression case is detailed in Chap. 7 while the extension to the multiple regression case is detailed in Chap. 8. The next three chapters are devoted to the statistical inference. Estimation of the Gini-based parameters is the topic of Chap. 9, a selection of formal tests is presented in Chap. 10, while tests that are related to the intersection of concentration curves are the topic of Chap. 11.

The second part of the book contains applications of the Gini methodology in various areas. We start with an introduction to the applications part (Chap. 12). In Chap. 13 we demonstrate the role of the Gini coefficient in two major competing theories that dominate the theoretical considerations in the area of income distribution, namely: the social welfare function approach and the theory of relative deprivation.

In Chap. 14 we illustrate the use of the concentration curves and the Gini methodology in the areas of taxation and progressivity of public expenditure.

Chapter 15 deals with the usefulness of several decompositions of the Gini and the extended Gini in analyzing government policies by non-marginal analyses, while in Chap. 16 the marginal analysis is illustrated. The applications in finance are the topic of Chaps. 17 and 18. These applications are relevant whenever one is interested in decision making under risk. Chapters 19–21 are devoted to applications of the Gini regression: in Chap. 19 we apply the simple Gini and extended Gini regressions, in Chap. 20 the multiple regression is applied, and in Chap. 21 we apply the mixed OLS, Gini, and extended Gini regressions. Chapter 22 deals with one application of the GMD and the Gini coefficient in statistics—an application that replicates the commonly used ANOVA and is denoted by ANOGI (ANalysis Of GIni). The last chapter (Chap. 23) concludes and lists several topics for further research.

Readers who will read the book will find some repetitions between the theoretical and the applications parts of the book. The reason for those repetitions is that each chapter in the applications part is written as a self-contained application. This approach is intended to enable the specialist in a field to read the relevant application chapter without having to read the whole book. Readers who want to see a proof for an argument are referred to the theoretical part.

Part I
Theory

Chapter 2
More Than a Dozen Alternative Ways of Spelling Gini

Introduction

Gini's mean difference (GMD) as a measure of variability has been known for over a century.[1] It has more than 14 alternative representations.[2] Some of them hold only for continuous distributions while others hold only for nonnegative variables. It seems that the richness of alternative representations and the need to distinguish among definitions that hold for different types of distributions are the main causes for its sporadic reappearances in the statistics and economics literature as well as in other areas of research. An exception is the area of income inequality, where it is holding the position as the most popular measure of inequality. GMD was "rediscovered" several times (see, for example, Chambers & Quiggin, 2007; David, 1968; Jaeckel, 1972; Jurečková, 1969; Olkin & Yitzhaki, 1992; Kőszegi & Rabin, 2007; Simpson, 1949) and has been used by investigators who did not know that they were using a statistic which was a version of the GMD. This is unfortunate, because by recognizing the fact that a GMD is being used the researcher could save time and research effort and use the already known properties of GMD.

The aim of this chapter is to survey alternative representations of the GMD. In order to simplify the presentation and to concentrate on the main issues we restrict the main line of the presentation in several ways. First, the survey is restricted to

This chapter is based on Yitzhaki (1998) and Yitzhaki (2003).

[1] For a description of its early development see Dalton (1920), Gini (1921, 1936), David (1981, p. 192), and several entries in Harter (1978). Unfortunately we are unable to survey the Italian literature which includes, among others, several papers by Gini, Galvani, and Castellano. A survey on those contributions can be found in Wold (1935). An additional comprehensive survey of this literature can be found in Giorgi (1990, 1993). See Yntema (1933) on the debate between Dalton and Gini concerning the relevant approach to inequality measurement.

[2] Ceriani and Verme (2012) present several additional forms in Gini's original writing that as observed by Lambert (2011) do not correspond to the presentations used in this book.

S. Yitzhaki and E. Schechtman, *The Gini Methodology: A Primer on a Statistical Methodology*, Springer Series in Statistics 272, DOI 10.1007/978-1-4614-4720-7_2,

quantitative random variables. As a result the literature on diversity which is mainly concerned with categorical data is not covered.[3] Second, the survey is restricted to continuous, bounded from below but not necessarily nonnegative variables. The continuous formulation is more convenient, yielding insights that are not as accessible when the random variable is discrete. In addition, the continuous formulation is preferred because it can be handled using calculus.[4] As will be shown in Sect. 2.4 there is an additional reason for the use of a continuous distribution: there is an inconsistency between the various tools used in defining the GMD when the distribution is discrete. This inconsistency complicates the presentation without adding any insight. To avoid problems of existence, only continuous distributions with finite first moment will be considered. The distinction between discrete and continuous variables will be dealt with in Sect. 2.4, while properties that are restricted to nonnegative variables will be discussed separately whenever they arise. Third, the representations in this chapter are restricted to population parameters. We deal with the estimation issue in Chap. 9.

Finally, as far as we know these alternative representations cover most, if not all, known cases but we would not be surprised if others turn up. The different formulations explain why the GMD can be applied in so many different areas and can be given so many different interpretations. We conclude this chapter with a few thoughts about the reasons why Gini was "rediscovered" again and again and with four examples that illustrate this point.

The structure of this chapter is as follows: Section 2.1 derives the alternative representations of the GMD. Section 2.2 investigates the similarity between GMD and the variance. Section 2.3 deals with the Gini coefficient and presents some of its properties. In sect. 2.4 the adjustments to the discrete case are discussed and Sect. 2.5 gives some examples. Section 2.6 concludes.

2.1 Alternative Representations of GMD

There are four types of formulas for GMD, depending on the elements involved: (a) a formulation that is based on absolute values, which is also known to be based on the L_1 metric; (b) a formulation which relies on integrals of cumulative distribution functions; (c) a formulation that relies on covariances; and (d) a formulation that

[3] For the use of the GMD in categorical data see the bibliography in Dennis, Patil, Rossi, Stehman, and Taille (1979) and Rao (1982) in biology, Lieberson (1969) in sociology, Bachi (1956) in linguistic homogeneity, and Gibbs and Martin (1962) for industry diversification. Burrell (2006) uses it in informetrics, while Druckman and Jackson (2008) use it in resource usage, Puyenbroeck (2008) uses it in political science while Portnov and Felsenstein (2010) in regional diversity.

[4] One way of writing the Gini is based on vectors and matrices. This form is clearly restricted to discrete variables and hence it is not covered in this book. For a description of the method see Silber (1989).

relies on Lorenz curves (or integrals of first moment distributions). The first type is the most convenient one for dealing with conceptual issues, while the covariance presentation is the most convenient whenever one wants to replicate the statistical analyses that rely on the variance such as decompositions, correlation analysis, ANOVA, and Ordinary Least Squares (OLS) regressions.

Let X_1 and X_2 be independent, identically distributed (i.i.d.) continuous random variables with F(x) and f(x) representing their cumulative distribution and the density function, respectively. It is assumed that the expected value μ exists; hence $\lim_{t\to-\infty} tF(t) = \lim_{t\to\infty} t[1 - F(t)] = 0$.

2.1.1 Formulas Based on Absolute Values

The original definition of the GMD is the expected absolute difference between two realizations of i.i.d. random variables. That is, the GMD in the population is

$$\Delta = E\{|X_1 - X_2|\}, \tag{2.1}$$

which can be given the following interpretation: consider an investigator who is interested in measuring the variability of a certain property in the population. He or she draws a random sample of two observations and records the absolute difference between them.

Repeating the sampling procedure an infinite number of times and averaging the absolute differences yield the GMD.[5] Hence, the GMD can be interpreted as the expected absolute difference between two randomly drawn members from the population. This interpretation explains the fact that for nonnegative variables the GMD is bounded from above by twice the mean because the mean can be viewed as the result of infinite repetitions of drawing a single draw from a distribution and averaging the outcomes, while the GMD is the average of the absolute differences between two random draws. Note, however, that this property does not necessarily hold for random variables that are not restricted to be nonnegative.

Equation (2.1) resembles the variance, which can be presented as

$$\sigma^2 = 0.5E\{(X_1 - X_2)^2\}\ . \tag{2.2}$$

Equation (2.2) shows that the variance can be defined without a reference to a location parameter (the mean) and that the only difference between the definitions of the variance and the GMD is the metrics used for the derivations of the concepts. That is, the GMD is the expected absolute difference between two randomly drawn

[5] See also Pyatt (1976) for an interesting interpretation based on a view of the Gini as the equilibrium of a game.

observations, while the variance is the expected square of the same difference. It is interesting to note that replacing the power 2 by a general power r in (2.2) is referred to as the generalized mean difference (Gini, 1966; Ramasubban, 1958, 1959, 1960). However, as far as we know, they were not aware of the fact that when $r = 2$ it is identical to the variance.

An alternative presentation of the GMD that will be helpful when we describe the properties of the Gini regressions and their resemblance to quantile regressions can be developed in the following way:

Let Q and X be two i.i.d. random variables; then by the law of iterated means the GMD can be presented as the average (over all possible values of Q) of all absolute deviations of X from Q. In other words

$$\Delta = E_Q E_{X|Q}\{|X - Q|\}. \tag{2.3}$$

Next, we note that Q in (2.3) can represent the quantile of the distribution. The reason is that the quantile can be assumed to have the same distribution function as X does, and can be assumed to be independent of X. To see that let $F_X(Q) = P$; then $F_X(Q)$ is uniformly distributed on [0, 1]. It follows that $Q = F_X^{-1}(P)$ is distributed as X,

$$G_Q(t) = P(Q \le t) = P(F_X^{-1}(P) \le t) = P(P \le F_X(t)) = F_X(t),$$

and independent of it. Therefore the term $E_{X|Q}\{|X - Q|\}$ in (2.3) can be viewed as the conditional expectation of the absolute deviation from a given quantile Q of the distribution of X. Hence equation (2.3) presents the GMD as the average absolute deviation from all possible quantiles.

From (2.3) one can see that minimizing the GMD of the residuals in a regression context (to be discussed in Chap. 7) can be interpreted as minimizing an *average* of all possible *absolute deviations* from all possible quantiles of the residual. We note in passing that (2.3) reveals the difference between the GMD and the expected absolute deviation from the mean. The former is the expected absolute difference from every possible value of Q, while the latter is the expected absolute deviation from the mean. We will return to this point in Chap. 23.

A slightly different set of representations relies on the following identities: let X_1 and X_2 be two i.i.d. random variables having mean μ. Then

$$\begin{aligned} |X_1 - X_2| &= (X_1 + X_2) - 2\text{Min}\{X_1, X_2\} = \text{Max}\{X_1, X_2\} - \text{Min}\{X_1, X_2\} \\ &= 2\text{Max}\{X_1, X_2\} - (X_1 + X_2). \end{aligned} \tag{2.4}$$

Using the first equation from the left of (2.4), the GMD can be expressed as

$$\Delta = 2\mu - 2E[\text{Min}\{X_1, X_2\}]. \tag{2.5}$$