

## 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.<sup>1</sup> It has more than 14 alternative representations.<sup>2</sup> 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).

<sup>1</sup>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.

<sup>2</sup>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.

quantitative random variables. As a result the literature on diversity which is mainly concerned with categorical data is not covered.<sup>3</sup> 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.<sup>4</sup> 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

---

<sup>3</sup> 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, Lieberman (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.

<sup>4</sup> 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  $\mu$  exists; hence  $\lim_{t \rightarrow -\infty} tF(t) = \lim_{t \rightarrow \infty} t[1 - F(t)] = 0$ .

### 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

$$\Delta = E \{ |X_1 - X_2| \}, \quad (2.1)$$

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.<sup>5</sup> 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

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

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

---

<sup>5</sup> 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

$$\Delta = E_Q E_{X|Q} \{|X - Q|\}. \quad (2.3)$$

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  $X$ ,

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

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  $\mu$ . Then

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

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

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

That is, the GMD is twice the difference between the expected values of one random draw and the minimum of two random draws from the distribution. Alternatively, we can use the middle part of (2.4) to write

$$\Delta = E[\text{Max}\{X_1, X_2\}] - E[\text{Min}\{X_1, X_2\}]. \quad (2.6)$$

Here, the interpretation of the GMD is as the expected difference between the maximum and the minimum of two random draws. Finally, one can use the right-hand side of (2.4) to write the GMD as twice the expected value of the maximum of two random draws minus twice the expected value of one random draw. These presentations can be easily extended to involve more than two draws, leading to the extended Gini (Yitzhaki, 1983). (This issue will be discussed in Chap. 6.) They can be useful whenever the interpretation of the GMD is related to extreme value theory.

### 2.1.2 Formulas Based on Integrals of the Cumulative Distributions

This section focuses on representations of the GMD that are based on integrals of the cumulative distribution. The basic equation needed in order to develop such representations is an alternative expression for the expected value of a distribution.

**Claim** Let  $X$  be a continuous random variable distributed in the range  $[a, \infty)$ . Then the expected value of  $X$  is given by<sup>6</sup>

$$\mu = a + \int_a^\infty [1 - F(x)]dx. \quad (2.7)$$

*Proof* The standard definition of the expected value is  $\mu = \int_a^\infty xf(x)dx$ . Using integration by parts with  $u = x$  and  $v = -[1 - F(x)]$  yields (2.7).

Using (2.7) and the fact that the cumulative distribution of the minimum of two i.i.d. random variables can be expressed as  $\{1 - [1 - F(x)]^2\}$  we can rewrite (2.5) as

$$\Delta = 2 \int [1 - F(t)]dt - 2 \int [1 - F(t)]^2 dt, \quad (2.8)$$

and by combining the two integrals

---

<sup>6</sup>The GMD is based on the difference of two such formulae, so this restriction on the range (to be bounded from below) does not affect the GMD. See Dorfman (1979).

$$\Delta = 2 \int F(t)[1 - F(t)]dt. \quad (2.9)$$

See Dorfman (1979). Equation (2.9) can be given an interesting interpretation. Let  $F_n(x)$  be the empirical cumulative distribution of  $X$  based on a sample of  $n$  observations. Then for a given  $x$ ,  $F_n(x)$  is the sample mean of  $n$  i.i.d. Bernoulli variables with  $p = F(x)$ . The variance of  $F_n(x)$  is equal to

$$\sigma_{F_n(x)}^2 = F(x)[1 - F(x)]/n \quad (2.10)$$

(Serfling, 1980, p. 57) and the GMD can be interpreted as  $2n \int \sigma_{F_n(x)}^2 dx$ .

A similar (and older) variant of this formula is

$$\Delta = 2nE \left\{ \int [F_n(x) - F(x)]^2 dx \right\}, \quad (2.11)$$

which is the original Cramer–Von Mises–Smirnov criterion for testing goodness of fit of a distribution.<sup>7</sup> In some sense (2.11) can be viewed as a “dual” approach to the central moments of a distribution. Central moments are linear in the probabilities and power functions of deviations of the variate from its expected value. In the GMD, the power function is applied to the deviation of the cumulative distribution from its expected value while the linearity is applied to the variate itself. Hence the “duality.”<sup>8</sup> This interpretation also suggests a possible explanation to some robustness properties of the “dual” approach. The range of  $F(\cdot)$  is  $[0, 1]$  while the range of the variate can be unlimited. Using a power function as is done in the regular moments may lead to unboundedness of the statistics, while all the moments of the dual approach are bounded, provided that the mean is bounded.

---

<sup>7</sup> This formula, which is a special case of the statistic suggested by Cramer, plays an important role in his composition of elementary errors although it seems that he did not identify the implied GMD (see Cramer, 1928, pp. 144–147). Von Mises (1931) made an independent equivalent suggestion and developed additional properties of the statistic. Smirnov (1937) modified the statistic to be

$$w^2 = n \int [F_n(x) - F(x)]^2 dF(x).$$

Changing the integration from  $dx$  to  $dF(x)$  eliminates the connection to the GMD and creates a distribution-free statistic. The above description of the non-English literature is based on the excellent review in Darling (1957). Further insight about the connection between the Cramér–Von Mises test can be found in Baker (1997) which also corrects for the discrepancy in calculating the GMD in discrete distributions.

<sup>8</sup> This “duality” resembles the alternative approach to the expected utility theory as suggested by Yaari (1988) and others. While expected utility theory is linear in the probabilities and nonlinear in the income, Yaari’s approach is linear in the income and nonlinear in the probabilities. In this sense, one can argue that the relationship between “dual” approach and the GMD resembles the relationship between expected utility theory and the variance. Both indices can be used to construct a specific utility function for the appropriate approach (the quadratic utility function is based on the mean and the variance while the mean minus the GMD is a specific utility function of the dual approach).

Finally, we can write (2.9) as

$$\Delta = 2 \int_a^\infty \left[ \int_a^x f(t) dt \int_x^\infty f(t) dt \right] dx, \quad (2.12)$$

which is the way Wold (1935) presented it.

An additional presentation by Wold (1935, equation 12, p. 47)<sup>9</sup> which is valid for nonnegative variables is

$$\Delta = 2 \int_0^\infty \left[ \int_0^t F(u) du \right] dF(t). \quad (2.13)$$

Equation (2.13) is listed for completeness.

### 2.1.3 Covariance-Based Formulas

It is well known that the variance is a special case of the covariance, because it can be written as  $\text{var}(X) = \text{cov}(X, X)$ . In this section we show that the GMD can be expressed as a covariance as well. Once the GMD is written as a covariance, the properties of the covariance are called for to define the Gini correlation and the decomposition of a GMD of a linear combination of random variables, which naturally leads to Gini regressions, Gini Instrumental Variable, time-series Gini analysis, and numerous other applications. Generally speaking, one can take an econometrics textbook that is based on the variance and rewrite (most of) it in terms of the GMD. Another advantage of the covariance presentation is that the covariance formula opens the way to the decomposition of the Gini coefficient (to be defined later) of an overall population into the contributions of several subgroups. In addition, it opens the way to the extended Gini family of measures of variability (to be discussed in Chap. 6), which means replicating (almost) everything that was developed with the Gini and finding out which properties carry on to an infinite number of measures of variability.

Let us start with presentation (2.9). Applying integration by parts to (2.9), with  $v = F(t) [1 - F(t)]$  and  $u = t$ , one gets, after deleting zeros and rearranging terms,

$$\Delta = 2 \int F(t)[1 - F(t)] dt = 4 \int t[F(t) - 0.5]f(t) dt. \quad (2.14)$$

---

<sup>9</sup>Wold (1935) used a slightly different presentation, based on Stieltjes integrals.

Recall that the expected value of  $F$ , which is uniformly distributed on  $[0, 1]$ , is 0.5. Therefore one can rewrite (2.14) as

$$\Delta = 4E\{X(F(X) - E[F(X)])\} = 4 \operatorname{cov}[X, F(X)]. \quad (2.15)$$

Equation (2.15) lets us calculate the GMD using a simple regression program as will be shown next.<sup>10</sup> Recall that  $F(X)$  is uniformly distributed on  $[0, 1]$ . Therefore,  $\operatorname{cov}[F(X), F(X)] = 1/12$  (a constant) and we can write the GMD as

$$\Delta = (1/3)\operatorname{cov}[X, F(X)]/\operatorname{cov}[F(X), F(X)]. \quad (2.16)$$

In order to gain some intuition assume that the observations are arrayed in ascending order (say, by height as in the case of soldiers in a parade) with equal distance between each two observations (soldiers). The following proposition summarizes two interpretations of the GMD.

### Proposition 2.1

- (a) *The GMD is equal to one-third of the slope of the OLS regression curve of the observed variable (height, the dependent variable) as a function of the observation's position in the array ( $F(X)$ , the explanatory variable).*
- (b) *The GMD is a weighted average of the differences in, say, heights between adjacent soldiers (alternatively, it is a weighted average of the slopes defined by each two adjacent heights in the array). The weights are symmetric around the median, with the median having the highest weight.*

*Proof of (a)* Trivial. Recall that the OLS regression coefficient in a linear regression model is given by

$$\beta = \frac{\operatorname{cov}(Y, X)}{\operatorname{cov}(X, X)}$$

and see (2.16) above.

*Proof of (b)* Let  $X(p)$  be the height of a soldier as a function of his position,  $p$ . For example,  $X(0.5)$  is the height of the median soldier. That is,  $P(X < X(p)) = p = F(X(p))$ . Note that  $X(p)$  is the inverse of the cumulative

---

<sup>10</sup> See Lerman and Yitzhaki (1984) for the derivation and interpretation of the formula, see Jenkins (1988) and Milanovic (1997) on actual calculations using available software, and see Lerman and Yitzhaki (1989) on using this equation to calculate the GMD in stratified samples. As far as we know, Stuart (1954) was the first to notice that the GMD can be written as a covariance. However, his findings were confined to normal distributions. Pyatt, Chen and Fei (1980) also write the GMD as a covariance. Sen (1973) uses the covariance formula for the Gini, but without noticing that he is dealing with a covariance. Hart (1975) argues that the moment-generating function was at the heart of the debate between Corrado Gini and the western statisticians. Hence, it is a bit ironic to find that one can write the GMD as some kind of a central moment.



distribution of  $X$  at  $p$ . Writing explicitly the numerator in (2.16) we get  $\text{cov}(X, p) = \int X(p)(p - 0.5)dp$  and by using integration by parts with  $u = X(p)$  and  $v = (p - 0.5)^2/2$  we get

$$\text{cov}(X, p) = X(p)(p - 0.5)^2/2|_0^1 - 0.5 \int X'(p)(p - 0.5)^2 dp.$$

Substituting  $X(1) - X(0) = \int X'(p)dp$ , where  $X'$  denotes a derivative, we get

$$\text{cov}(X, p) = 0.5 \int X'(p)p(1 - p)dp. \quad (2.17)$$

Equation (2.17) shows that the GMD is equal to the weighted average of the slopes  $X'(p)$  and the weighting scheme  $p(1 - p)$  is symmetric in  $p$  around the median ( $p = 0.5$ ). The maximum weight is assigned to the median ( $p = 0.5$ ), and the weights decline the farther the rank of the observation gets from the median.

A consequence of (2.17) is that the flatter the density function of  $X$  is, the larger the GMD becomes (which is intuitively clear for a measure of spread). To sum up, according to these presentations the GMD is the weighted average change in a random variable as a result of a small change in the ranks. Because  $X(p)$  is the inverse of the cumulative distribution it is easy to see that  $X'(p) = \frac{1}{f(x)} dx$ . That is, the slope is the reciprocal of the density function.

Equation (2.15), the covariance representation of the GMD, can be used to show that  $R$ -regressions (Hettmansperger, 1984) are actually based on minimizing the GMD of the residuals in the regression. To see that, note that the target function in  $R$ -regression is to minimize  $\sum_i e_i R(e_i)$ , where  $e_i$  is the error term of the  $i$ -th observation in the regression while  $R(e_i)$  is its rank. Note that the mean of the residuals is equal to zero, and that the rank of the variable represents the cumulative distribution in the sample. Taking into account those facts, it is easy to see that  $R$ -regression is actually based on minimizing the GMD of the residuals. Therefore some properties of these regressions can be traced to the properties of the GMD. We will further elaborate on this point in Chap. 7.

We will be using the covariance formula of the GMD extensively in this book. It makes it very natural and convenient to “translate” the variance-based parameters such as the regression and the correlation coefficients into the Gini language. It is interesting to note that for the discrete case these facts were already mentioned in Gini (1914) and were repeated in Wold (1935). For the continuous case one can find the covariance presentation in 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 contribution of Lerman and Yitzhaki (1984) is in recognizing the implications of the term covariance when dealing with the decomposition of the variability measure and in producing the additional parameters which are based on it—a step that opened the way to applying the GMD method in the multivariate case.

### 2.1.4 Lorenz Curve-Based Formulas

The fourth set of representations of the GMD is based on the Absolute Lorenz Curve (ALC), which is also referred to as the generalized Lorenz curve.<sup>11</sup> The ALC and the concentration curve play important roles in the understanding of the compositions and the contributions of different sections of the distribution to the GMD and other related parameters such as Gini covariance and Gini correlation. Therefore they will be discussed in detail in Chap. 5. In this section we briefly mention the Lorenz curve-based formulas. There are several definitions of the ALC. We follow Gastwirth's (1971, 1972) definition, which is based on the inverse of the cumulative distribution. Let  $F(X(p)) = p$ , then  $X(p) = F^{-1}(p)$ . The ALC is plotted as follows:  $p$  is plotted on the horizontal axis while the vertical axis represents the cumulative value of the variate,  $-\infty \int^p X(t)dt$ . The familiar (relative) Lorenz curve (LC) is derived from the ALC by dividing the cumulative value of the variate by its expected value. The vertical axis is then  $(1/\mu) - \infty \int^p X(t)dt$ . The ALC has the following properties:

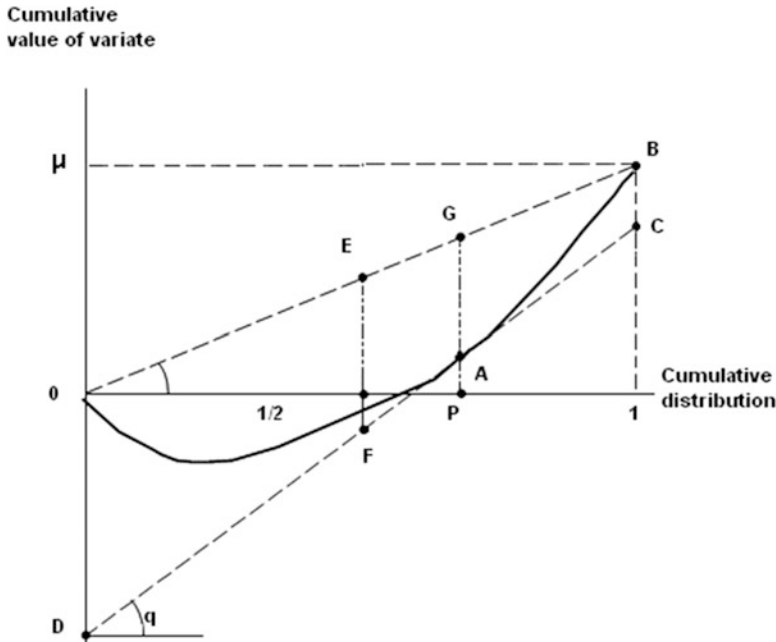
1. The ALC passes through  $(0, 0)$  and  $(1, \mu)$ . The LC passes through  $(0, 0)$  and  $(1, 1)$ .
2. The derivative of the curve at  $p$  is  $X(p)$ , which is the inverse of the cumulative distribution function; hence the curve is increasing (decreasing) depending on whether  $X(p)$  is positive (negative). Because  $X(p)$  is always a nondecreasing function of  $p$  the ALC is convex.

Figure 2.1 presents a typical ALC, the curve OAB. The slope of the line connecting the two extremes of the curve is  $\mu$ . We refer to this line as the Line of Equality (LOE), because when all observations are equal, the curve coincides with the line. The line OEGB in Fig. 2.1 represents the LOE. It can be shown (details will be given in Chap. 5) that the area between LOE and the ALC, OAB is  $\text{cov}(X, F(X))$  (that is, one-fourth of the GMD). We will return to this topic when dealing with the properties of ALC in Chap. 5.

As far as we know, we have covered all known interesting presentations of the GMD. The rest of this chapter is intended to supply some intuition and to compare the GMD with the variance.

---

<sup>11</sup> The term "generalized Lorenz curve" (GLC) was coined by Shorrocks (1983). Lambert and Aronson (1993) give an excellent description of the properties of GLC. However, it seems that the term "absolute" is more intuitive because it distinguishes the absolute curve from the relative one. Hart (1975) presents inequality indices in terms of the distribution of first moments, which is related to the GLC.



**Fig. 2.1** The absolute Lorenz curve. *Source:* Yitzhaki, 1998, p. 21. Reprinted with permission by Physica Verlag, Heidelberg

## 2.2 The GMD and the Variance

In this section we investigate the similarities and the differences between the GMD and the variance. As will be seen, on one hand they share many properties, but on the other hand there are some fundamental differences.

### 2.2.1 The Similarities Between GMD and the Variance

The first similarity between the GMD and the variance is the fact that both can be written as covariances. The variance of  $X$  is  $\text{cov}(X, X)$ , while the GMD of  $X$  is  $\text{cov}(X, F(X))$ . This similarity serves as the basis for the ability to “translate” the variance world into the Gini world.

The second similarity is the fact that the decomposition of the variance of a linear combination of random variables is a special case of the decomposition of the GMD of the same combination. The decomposition of the GMD includes some extra parameters that provide additional information about the underlying distribution, as will be developed in Chap. 4. If these additional parameters are equal to

zero then the decompositions of the GMD and the variance have identical structures. This property makes the GMD suitable for testing implicit assumptions that lead to the convenience of using the variance. This property is also the base for the claim that was put forward by Lambert and Decoster (2005) that “the Gini reveals more.”

The third similarity is the fact that both the variance and the GMD are based on averaging the distances between all pairs of observations (see (2.1), (2.2), and Daniels, 1944, 1948) or, alternatively, averaging the distances between random draws of two i.i.d. random variables. However, the difference between them is in the distance function used. The effects of the distance functions on the properties of the indices will be illustrated when we deal with the properties of OLS and the Gini regression coefficients in Chap. 7. The source of this difference will be discussed in the next section.

### 2.2.2 *The Differences Between the GMD and the Variance: City Block vs. Euclidean*

Let  $\Delta x_k$  denote the difference between adjacent observations. That is,  $\Delta x_k = X_{k+1} - X_k$ , where the observations are arranged in an increasing order. Then for any two ordered observations  $X_i > X_j$

$$X_i - X_j = \sum_{k=j}^{i-1} \Delta x_k. \quad (2.18)$$

The GMD and the variance can be presented as weighted averages of these distances between adjacent observations.<sup>12</sup> In both cases the weighting scheme attaches the highest weight to the mid-rank observation (i.e., the median), and the weights decline symmetrically the farther the rank of the observation is from the mid-rank. The fundamental difference between the two measures of variability is attributed to the distance function they rely on. The GMD’s distance function is referred to as the “city block” distance (or  $L_1$  metric), while the variance’s distance is Euclidean. It is interesting to note that other measures of variability (e.g., the mean deviation) also rely on the  $L_1$  metric, but they do not share the weighting scheme caused by the averaging of differences between all pairs of observations. To shed some light on the difference between the distance functions, note that the most basic measure of variability is the range, which is equivalent to the simple

---

<sup>12</sup> In the case of the GMD, the weights are not functions of  $\Delta x_k$  so that it is reasonable to refer to them as weights. In the case of the variance, the “weights” are also functions of  $\Delta x_k$  which makes the reference to them as weights to be incorrect. We refer to them as weights in order to compare with the GMD. See Yitzhaki (1996).

sum of the distances between adjacent observations, so that we end up with the difference between the most extreme parts of the distribution. If the distributions are restricted to have only two observations then the variance and the GMD (and all other measures of variability) will order all distributions in accordance with the ordering of the range. However, the range suffers from two major deficiencies: (1) it is not sensitive to the distribution of the non-extreme observations and (2) there are many important distributions with an infinite range.

In order to illustrate the difference between the distance functions embodied in the GMD and the variance one can ask, for a given range, what characterizes the distribution with the smallest/largest variance (GMD). Alternatively, for a given variance (GMD) one can ask what characterizes the distribution with the smallest/largest range. Presumably by answering those questions we will be able to form an opinion as to which distance function is more appropriate for a given situation and which one reflects our intuition better. To illustrate, let us restrict the distributions to have only three possible values and assume a given (normalized) range (equals to 1 in our example) so that the discussion is restricted to distributions of the type:  $[0, \delta, 1]$ . Which  $\delta$  will maximize or minimize each variability index? Ignoring constants, the GMD is

$$\Delta(\delta) = \sum \sum |X_i - X_j| = 1 + \delta + |1 - \delta|, \quad (2.19)$$

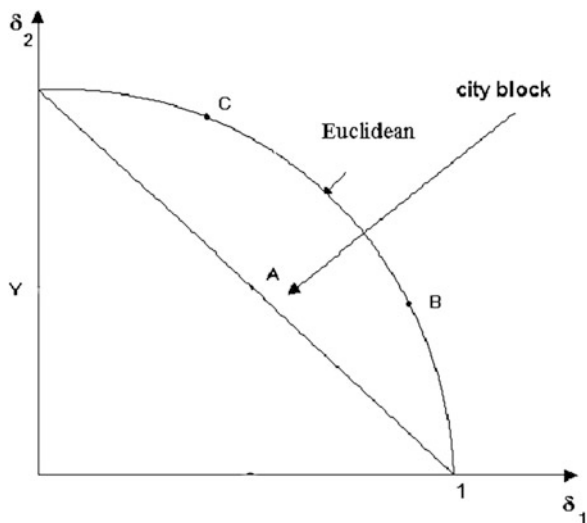
and it equals 2 regardless of the value of  $\delta$ . (More generally, it is equal to twice the range). Thus the position of the middle observation does not change the GMD. Repeating the same exercise with the variance yields (again, ignoring constants)

$$\sigma^2(\delta) = 1 + \delta^2 + (1 - \delta)^2 \quad (2.20)$$

and the variance is maximized for  $\delta = 0$  or 1 and minimized for  $\delta = 0.5$ . That is, for a given range, the more equal the distances defined by adjacent observations are, the smaller is the variance. The conclusion is that the variance is more sensitive to the variability in  $\Delta x_i$  than the GMD. In other words, if one of the differences will be extremely large the variance will be affected by it more than the GMD. This fact is responsible for the sensitivity of the variance to extreme observations.

An alternative way to illustrate the difference between the variance and the GMD can be presented geometrically. Let  $\delta_1$  be the difference between the second and first observations and let  $\delta_2$  be the difference between the third and second observations. Figure 2.2 presents an example of equal GMD and equal variance curves. That is, we allow the range to vary and instead, we are asking what should  $\delta_1$  and  $\delta_2$  be so that we get equal Ginis or equal variances. The horizontal axis represents  $\delta_1$  while the vertical axis represents  $\delta_2$ . The range, of course, is equal to  $\delta_1 + \delta_2$ . Gini (2.19) becomes  $1 + \delta_1 + \delta_2$  while the variance (2.20) is  $1 + \delta_1^2 + \delta_2^2$ . By changing the value of the GMD or the variance, one gets parallel curves of different distances from the origin, which can be referred to as equi-variance (GMD) curves. The point A is on the equi-GMD curve, so that all the points on the

**Fig. 2.2** Equal GMD and variance curves. *Source:* Yitzhaki, 2003, p. 292. Reprinted with permission by Metron International Journal of Statistics



graph have a value of GMD that is equal to the GMD at A. The points B and C are on the equi-variance curve. (In Fig. 2.2 the chosen value is 2).

As can be seen, the two measures represent different types of curves of equal distances. Imagine you are in a city. If you are allowed to only move in the east/west or north/south directions then you are in a GMD (city block) world. If, on the other hand, you are allowed to move in any direction you want, and you are Pythagorean, then you are in a variance world. It is hard to determine in general which distance function should be preferred. If one is traveling on the sea then the variance metric makes sense. However, the money metric which is extensively used by economists resembles the city block metric because the distance function embodied in the budget constraint is identical to the distance function of the GMD. One does not get a discount for spending equally on two commodities, as is the case of the variance (see Deaton, 1979; Jorgenson & Slesnick, 1984; McKenzie & Pearce, 1982 on uses of the money metric in economics). Hence when it comes to choosing a metric, the natural choice for the economists should be the GMD-type metric because spending money and the budget constraint follow the money metric rules.

The implication of the difference in metrics can also be seen from the following question which should be answered intuitively, without calculations. Consider the following distributions:  $[0, 0, 1]$  vs.  $[0, 0.575, 1.15]$ . Which distribution portrays a higher variability? If your intuitive answer points to the former (latter) distribution then you want to be in a variance (GMD) world (the variances are 0.222 and 0.220, respectively, while the Ginis are 0.667 and 0.767, respectively). The extension to more than three observations is straightforward.

This difference is responsible for the sensitivity of OLS regression to extreme observations and for the robustness properties of the GMD regressions as will be discussed in Chap. 7. Another implication of the difference in metrics is that the

GMD exists whenever the expected value exists while the existence of the variance requires the existence of a second moment.

It is interesting to note that if the underlying distribution is normal, then the increase in the distance between adjacent observations when moving from the middle to the extremes is identical to the decrease in the weight due to being farther away from the median, so that each observation gets an equal weight (Yitzhaki, 1996). Our conjecture is that this property leads to the statistical efficiency of variance-based statistics in cases of normality: the weights are distributed equally among observations. The main conclusion from the above discussion is that it is not obvious which metric is the preferred one, and the subject matter one is dealing with should also be taken into account when considering the appropriate metric.

Finally, an important property of the GMD is that it is bounded from above by  $\frac{2}{\sqrt{3}} \sigma_X$  (as is shown below). The advantage of having the bound from above by a function of the standard deviation is that whenever the standard deviation converges to zero, so does the GMD. Note that

$$1 \geq \rho(X, F(X)) = \frac{\text{cov}(X, F(X))}{\sigma_X \sigma_F} = \frac{\Delta_X}{4 \sigma_X \sigma_F}.$$

Recall that  $F$  is uniformly distributed. Hence its standard deviation is equal to  $\frac{1}{\sqrt{12}}$ . Therefore we get the bound

$$\Delta_X \leq \frac{2}{\sqrt{3}} \sigma_X. \quad (2.21)$$

Note that if  $X$  is uniformly distributed on  $[0, 1]$  then  $\rho(X, F(X)) = \rho(X, X) = 1$  and (2.21) holds as equality. This implies that (a) it is impossible to improve the bound and (b) for the uniform distribution the GMD is a constant multiplied by the standard deviation. (A similar case occurs under the normality, where the GMD  $= 2\sigma/\sqrt{\pi}$ .)

Our main purpose in this book is to imitate the applications of variance-based methods. Equation (2.21) enables us to simplify the analysis in this book by restricting the distributions to those with a finite lower bound. The reason for the above statement is that any convergence property that can be attributed to the variance can also be proved for the GMD. For additional bounds on the GMD see Cerone and Dragomir (2005, 2006) and Dragomir (2010).

We mention, for completeness, that there is also a bound which is related to the mean absolute deviation (MAD). The bound can be derived from

$$\frac{1}{2} \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}| \leq \frac{1}{n^2} \sum_{i < j} |x_i - x_j| \leq \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|. \quad (2.22)$$

For details see Cerone and Dragomir (2006). Further discussion on the relationship between MAD and GMD will be given in Chap. 23.

## 2.3 The Gini Coefficient

The most well-known member of the Gini family is the Gini coefficient. It is mainly used to measure income inequality. The Gini coefficient can be defined in two alternative ways:

- (a) The Gini coefficient is the GMD divided by twice the mean. For this definition to hold, the mean must be positive.
- (b) The Gini coefficient, also known as the concentration ratio, is the area enclosed between the 45° line and the actual Lorenz curve divided by the area between the 45° line and the Lorenz curve that yields the maximum possible value that the index can have. This definition which is based on the areas enclosed by the actual and potential Lorenz curves holds for nonnegative variables only. (Zenga (1987) describes the historical development of the connection between the concentration ratio and the GMD.)

There are two differences between the two alternative definitions. The first difference is that the first definition applies only when the expected value of the variable is positive while the second imposes the restriction that the variable is bounded to be nonnegative (otherwise the maximum inequality may be unbounded). The second difference is that the first definition is valid only for continuous distributions while the second definition has a built-in correction for discrete distributions with finite number of observations. To see that assume that the distribution is composed of  $n$  observations. Then the upper bound of the Gini coefficient is  $(n - 1)/n$ . (It is attained when all observations except one are equal to zero). As a result, the area enclosed between the 45° line and the Lorenz curve is divided by  $(n - 1)/n$ . This correction plays a similar role as the correction for degrees of freedom.

The Gini coefficient was developed independently of the GMD, directly from the Lorenz curve and for a while it was called “the concentration ratio.” Gini (1914) has shown the connection between the GMD and the concentration ratio. Ignoring the differences in definitions, the relationship between the GMD and the Gini coefficient is similar to the one between the variance and the coefficient of variation,  $CV = \frac{\sigma}{\mu}$ , a property that was already known in 1914. That is, the Gini coefficient is a normalized version of the GMD and it is unit-free (measured in percent). In order to calculate it one only needs to derive the GMD, and then easily convert the representation into a Gini coefficient by dividing by twice the mean.

The best known version of the Gini coefficient is as twice the area between the 45° line and the Lorenz curve (definition (b) above). For this definition the range of the coefficient is  $[0, 1]$ , with 0 representing perfect equality while 1 is reached when one observation is positive and all other observations are zero. Similar to the coefficient of variation, the Gini coefficient can be defined for distributions with negative lower bound, provided that the expected value is positive (definition (a) above). However in this case the upper bound for the Gini coefficient can be greater than one. Also similar to the coefficient of variation, the Gini coefficient is not defined for distributions with expected value of zero. Being a unit-free index, the Gini coefficient is unaffected by multiplication of the variable by a constant.



Although the normalization seems innocent—normalization of the units in which the Gini is measured—it may have implications on the notion of inequality and variability. It is worth mentioning that reference to “variability” or “risk” (most common among statisticians and finance specialists) implies the use of the GMD, whereas reference to “inequality” (usually in the context of income distribution) implies the use of the Gini coefficient. To see the implication of the normalization, assume a distribution that is bounded in the range  $[a, b]$ . Try to answer the following question: What characterizes a distribution that is the most unequal according to a relative measure (either the Gini coefficient or the coefficient of variation), and what characterizes a distribution that is most unequal according to an absolute measure like the GMD or the variance? It is easy to see that when an absolute measure is used to rank inequality or variability, then the most unequal distribution is the one with half of the population at  $a$  and the other half at  $b$ . On the other hand, the answer according to a relative measure will be that the most unequal distribution is the one with almost all the population at  $a$  and only a tiny fraction at  $b$ . Therefore when dealing with issues of justice, a minor unnoticeable change may reflect a major change of opinion. By a seemingly innocent division by (twice) the mean one can switch between what Kolm (1976) refers to as “leftist” and “rightist” measures of inequality, a point that we will discuss at length in Chap. 13 when we deal with applications of the Gini methodology.

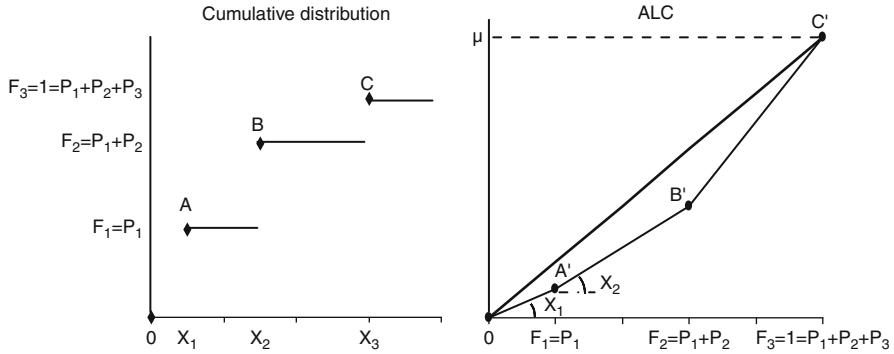
## 2.4 Adjustments Needed for Discrete Distributions

The discussion so far was limited to continuous distributions. When dealing with discrete distributions, or with empirical distributions that are discrete in nature, or even when the distributions are continuous while one is interested to do a decomposition of the variability according to population subgroups, one encounters a problem of inconsistent definitions of the basic concepts which imply a serious problem of incompatible calculations. For a survey of some of the problems arising in discrete distributions see Niewiadomska-Bugaj and Kowalczyk (2005).

In what follows we point out two inconsistencies. The first is that the definitions of the absolute and relative Lorenz curves and the cumulative distribution are incompatible (to be detailed below) and the second is similar to the issue of degrees of freedom.

### 2.4.1 *Inconsistencies in the Definitions of Lorenz Curves and Cumulative Distributions*

Assume the following pairs of observed values and their probabilities of occurrence:  $(x_1, p_1), \dots, (x_k, p_k), \dots, (x_n, p_n)$ . For simplicity of exposition we assume that the observations are ordered in a nondecreasing order. That is, if  $i < j$



**Fig. 2.3** The cumulative distribution and the ALC for discrete distributions

then  $x_i \leq x_j$ . This means that we observe at most  $n$  points on the cumulative distribution and on the absolute (or relative) Lorenz curve. We also assume that at least one  $p_k$  is not equal to the others. The cumulative distribution at  $x_k$  is

$$F_k = F(x_k) = \sum_{i=1}^k p_i. \quad (2.23)$$

The commonly used definition of a cumulative distribution function (cdf) is as a step function, continuous from the right, as shown on the left side of Fig. 2.3 for the case  $n = 3$ .

That is, the horizontal axis representing the variate is assumed to be continuous almost everywhere, while the vertical axis representing the cumulative distribution is discontinuous and jumps between the points. The right-hand side of Fig. 2.3 presents the ALC. In the absolute (or relative) Lorenz curve the horizontal axis represents the cdf while the vertical axis represents the cumulative value of the weighted mean. Formally, the vertical axis of the absolute Lorenz curve,  $q_k$ , is

$$q_k = q(F_k) = \sum_{i=1}^k p_i x_i. \quad (2.24)$$

When plotting a Lorenz curve, the usual procedure is to connect the points  $(F_k, q_k)$  by linear segments, making the curve continuous in  $F$ , and discontinuous in the variate (which is the slope of the Lorenz curve). These definitions of the cumulative distribution and the Lorenz curve are not compatible with each other because if we use definition (2.23) in plotting the Lorenz curve, we should plot the Lorenz curve as a step function and this would change the value of the Gini coefficient. This complicates the curves, the convexity property is lost, and even if this treatment solves the problem for the Gini, it is not clear how to solve this issue when dealing with other parameters of the Gini method such as the Gini

covariance, to be presented in Chap. 3. Whenever the number of points is small one should expect large discrepancies between the different methods of calculations.

Several solutions are available in the literature to handle those problems.

The inconsistency between the definitions of the Lorenz and the cumulative distribution and the effect on the Gini as defined in (2.1) have already been mentioned in the original paper by Gini who suggested to plot the Lorenz curve as a step function (see the translation in Metron, 2005, p. 25). But this solution has several disadvantages: (1) the convexity property of the Lorenz curve disappears and (2) to be consistent, the 45° line has to be changed into a step function as well.

Another solution was suggested by Lerman and Yitzhaki (1989). They suggested using a mid-point approximation of the cumulative distribution in the covariance formula. That is,

$$\Delta = 4 \sum_{i=1}^n p_i (x_i - \mu) \left( \frac{F(x_i) + F(x_{i-1})}{2} \right), \quad \text{with } F(x_0) = 0. \quad (2.25)$$

This solution overcomes the problem of inequality between the Lorenz and the covariance formulas. However, it may raise difficulties in interpretation because the mid-point cumulative distribution is not formally a cumulative distribution. Also, this solution is useful with respect to the Gini but it does not solve the problem in an extended Gini context. See Chap. 6.

### 2.4.2 Adjustment for a Small Number of Observations

When dealing with discrete distributions two additional problems arise. The first is that the upper bound of the Gini coefficient ceases to be one; instead, it is equal to  $(n - 1)/n$ . Therefore, the number of observations affects the estimated inequality. In some sense this is similar to the correction for degrees of freedom. The other problem arises when trying to implement (2.1). The issue is how to handle ties, whether to add them with zero value to the numerator which reduces the average absolute deviation or to omit them both from the numerator and the denominator.

## 2.5 Gini Rediscovered: Examples

The GMD is an intuitive measure of variability. Therefore it can be easily conceived. It turns out that (as was shown above) there is a large number of seemingly unrelated presentations of the GMD (and other parameters that are derived from it), making it hard to identify that one is dealing with a GMD and also to identify which version of the GMD one is dealing with. The GMD takes different forms for discrete vs. continuous distributions and for nonnegative vs.

bounded-from-below random variables. The fact that one has to differentiate between discrete and continuous distributions and between nonnegative and bounded-from-below variables, together with the wealth of alternative presentations, makes it hard to identify a GMD. In addition, as will be seen throughout the book, the GMD falls in between parametric and nonparametric statistics. Some of the properties resemble nonparametric while the others resemble parametric statistics. This can make it complicated for the “purists” to grasp. And finally, the alternative representations are scattered throughout many papers and languages, spread over a long period of time and in many areas of interest, and not all are readily accessible.<sup>13</sup>

The advantage of being able to identify a GMD is that it enables the investigator to use the existing literature in order to derive additional properties of the parameter at hand and rewrite it in an alternative, more user-friendly way. It also enables the investigator to find new interpretations of the GMD and of Gini-related parameters as well as draw inference about them. In order to illustrate this point we present four cases in details. We start with a recent reinvention of GMD that appeared in 2007 in a paper published in the *American Economic Review* by Kőszegi and Rabin, entitled: “Reference-Dependent Risk Attitudes.” In Appendix A of the paper, entitled “Further Definitions and Results” on p. 1063, the authors write: “In this appendix we present an array of concepts and results that may be of practical use in applying our model but that are not key to any of the main points of the paper.” Definition 5 is the following:

“DEFINITION 5: *The average self-distance of a lottery F is*

$$S(F) = \iint |x - y| dF(x) dF(y).$$

The average self-distance of a lottery is the average distance between two independent draws from the lottery. A lower average self-distance is a necessary but not sufficient condition for one lottery to be unambiguously less risky than another.” p. 1063. Anyone who is familiar with GMD will recognize the index (see Yitzhaki, 1982a). The Kőszegi and Rabin article may be the starting point of a new branch in the literature that will not be recognized as related to the GMD, and therefore will not rely on the already proven properties, and maybe several years down the road a future author will notice that the properties of the GMD were investigated again, using new terminology.

The second example is what is referred to as R-regression (Hettmansperger, 1984). As will be clear in Chap. 7, R-regression is actually a regression technique based on minimizing the GMD of the residuals of the regression. Knowing this fact can simplify many of the proofs of the properties of R-regression.

---

<sup>13</sup> This phenomenon seems to be a characteristic of the literature on the GMD from its early development. Gini (1921) argues: “probably these papers have escaped Mr. Dalton’s attention owing to the difficulty of access to the publications in which they appeared.” (Gini, 1921, p. 124).

A third example is the debate between Corrado Gini and the Anglo-Saxon statisticians. The most popular presentation of the variance is as a second central moment of the distribution. The most popular presentation of the GMD is as the expected absolute difference between two i.i.d. variables. See Giorgi (1990) for a bibliographical portrait. Using the expected absolute difference between two i.i.d. variables in order to measure variability characterized the Italian school, led by Corrado Gini, while reliance on moments of the distribution characterized the Anglo-Saxon school. However, as shown by Hart (1975) and the covariance presentation, and as will be shown in Chap. 6, the GMD can also be defined as a central moment. Had both sides known about the alternative presentations of the GMD, this debate which was a source of confrontation between the Italian school and what Gini viewed as the Western schools could be avoided (see Gini, 1965, 1966, p. 199; Hart, 1975).

A fourth example is the presentation of the GMD as four times the covariance between the variate and its cumulative distribution (Lerman & Yitzhaki, 1984). This formula can be seen in Wold (1935, p. 43) except that it was not referred to as covariance. Understanding that the GMD is actually a covariance enables the imitation of the decomposition properties of the variance. This property turns Gini into an analytical tool and enables replicating ANOVA, regression, and correlations—which in some sense doubles and triples the possibilities of modeling in economics and statistics. The result is that almost every model that can be constructed using the variance can be replicated using the Gini.

## 2.6 Summary

This chapter surveys all (known to us and relevant to the purpose of the book) alternative representations of the GMD and the Gini coefficient. While it is hard to make an accurate count of how many independent alternative definitions exist, there are clearly more than a dozen of them. Each representation is naturally related to a specific area of application. For example, the covariance formulation is natural when one is interested in regression analysis or in the decomposition of a Gini of a population into the contributions of the subpopulations.

The fact that the GMD is an intuitive measure and the need to distinguish between discrete and continuous and between negative and nonnegative variables may explain why the Gini has been “reinvented” so often. It also explains why it is harder to work with the Gini than with the variance.

The Gini is an alternative measure of variability. Therefore it is only natural that it shares some properties with the variance on one hand, and exhibits some differences on the other hand. These similarities and differences are discussed in this chapter. The main difference between the two measures lies in the distance function used. While the Gini uses the absolute value as the distance, the variance uses the square. This difference has practical implications that will be discussed later.

The Gini Methodology

A Primer on a Statistical Methodology

Yitzhaki, S.; Schechtman, E.

2013, XVI, 548 p., Hardcover

ISBN: 978-1-4614-4719-1