OXFORD BULLETIN OF ECONOMICS AND STATISTICS, 47, 2 (1985)
0305-9049 $3.00

PRACTITIONERS' CORNER*

Calculating the Gini Index of inequality for Individual Data

Haim Shalit†

I. INTRODUCTION

For the practitioner and the layman, the Gini index rhymes with income inequality. Being equal to 1 minus twice the area under the Lorenz curve, the Gini index of concentration appeals to most economists who rank income distributions in empirical studies. However, this ranking is unambiguous, if and only if, Lorenz curves do not intersect (see Atkinson, 1970). Therefore, the curve and the Gini index should be constructed together. This argument was advanced in a recent contribution by Brown and Mazzarino who provided an algorithm to draw the Lorenz curve from grouped data and then derive the Gini index of concentration from that figure.

Recently, Gini's difference was used in risk analysis (Yitzhaki, 1982) and in finance theory (Shalit and Yitzhaki, 1984). The equivalence between decision-making under uncertainty models and income inequality problems was perceived by Atkinson (1970). Hence, it is not surprising to see the Gini as a measure of dispersion in portfolio analysis. What has prevented the wider use of the Gini index in applied economics was apparently related to its complex calculation from grouped data, often the only information available. In this note, I provide a simple formula to compute the Gini index for a single variable such as income and also for the various elements that compose it when individual data points are available. Furthermore, the Lorenz curve drawings are not required and there is no need to assume a specific frequency function.

II. THE GINI INDEX OF CONCENTRATION

Consider a random variable $X \in (a,b)$, $a \geqslant 0$ and $b \leqslant \infty$ with a density function $f(x)$ and a distribution $F(x)$.[1] The Gini index of concentration

[1] Although it is not imperative to have positive variates, the simple Gini's index of concentration requires a positive mean. See Kendall and Stuart (1977) for the following presentation.

* The purpose of Practitioner's Corner is to publish brief methodological notes of interest to applied economists. The Editors welcome submissions of this sort.

† The note was written when visiting the University of Maryland.

is defined as

$$G = \Gamma/\mu \tag{1}$$

where

$$\mu = \int_a^b xf(x)\,dx$$

and

$$\Gamma = 1/2 \int_a^b \int_a^b |x-y|\, f(x) f(y)\,dx\,dy \tag{2}$$

is one half of Gini's mean difference (MD) and μ is the mean of X. Formula (2) is the standard representation of Gini's MD which has some intuitive appeal as a measure of dispersion. Indeed, Gini's MD is the expected distance between two realizations of variate X. Using the identity $|x-y| = x + y - 2\min(x, y)$, we rewrite (2) as[2]

$$\Gamma = \int_a^b [1-F(x)]\,dx - \int_a^b [1-F(x)]^2\,dx \tag{3}$$

which, for finite values of a, it becomes

$$\Gamma = \mu - a - \int_a^b [1-F(x)]^2\,dx \tag{4}$$

Integrating formula (4) by parts (with $v = [1-F(x)]^2$ and $u = x$), using

$$\int_a^b [1-F(x)]\, f(x)\,dx = 1/2$$

changing the variable of integration with $f(x)\,dx = dF$, and replacing its limits with $F(a) = 0; F(b) = 1$, one obtains:

$$\Gamma = 2 \int_0^1 (F - 1/2)\, x\,dF \tag{5}$$

Since F is uniformly distributed between 0 and 1, its mean is 1/2. Therefore, equation (6) expresses Gini's mean difference as twice the

[2] See Dorfman (1979) for the derivation.

covariance between the variate and its cumulative distribution. Thus, Gini's index of concentration is simply

$$G = 2\,\mathrm{Cov}[x, F(x)]/\mu \tag{6}$$

and is obtained by computing the moments of variate X. In practice, we analyse only a population sample and estimate the Gini index from these data points. For simplicity, we will assume that the data is not subject to sampling error. Therefore the Gini index will be estimated as follows.

Step 1: Rank the n data points in increasing order and calculate the mean as

$$\bar{x} = \sum_{i=1}^{n} x_i/n \qquad x_1 \leqslant x_2 \leqslant x_i \leqslant x_n$$

Step 2: Calculate the covariance between the newly ranked variable and a vector of positive integers 1 to n. Multiply by 2 and divide by n to obtain one half of Gini's mean difference.

Step 3: Divide by the mean $\bar{x}$ to obtain Gini's index of concentration.

Once the individual data is ranked in increasing order, the method is as simple as calculating the variance. Since most computers and some hand calculators have regression packages, it is easy to compute the Gini index by regressing the ranked variate on the set of natural numbers as follows

$$x_i = \alpha + \beta i + u \qquad i = 1, \ldots, n \tag{7}$$

where $x_1 \leqslant x_2 \leqslant x_3 \ldots \leqslant x_i \ldots \leqslant x_n$.

The estimated slope coefficient of this regression is

$$\hat{\beta} = \mathrm{Cov}(x, i)/\mathrm{var}(i) = \mathrm{Cov}(x, i)/((n^2 - 1)/12) \tag{8}$$

Hence from equation (7), the Gini index is calculated as

$$G = ((n^2 - 1)/6n) \cdot (\hat{\beta}/\bar{x}) \tag{9}$$

III. THE GINI INDEX FOR FACTOR COMPONENTS

When the variable to be analysed is the sum of several components the practitioner is interested in evaluating the contribution of the various elements to the Gini index. In the field of income distribution, the purpose of the approach is to decompose the inequality index among all the factors that contribute to income (see Pyatt *et al.*, 1980). The same method is used in finance theory where the portfolio is a weighted sum of individual securities, and the emphasis is placed on evaluating the relative performance of individual securities on the portfolio they compose (see Shalit and Yitzhaki, 1984).

Consider variate X as the weighted sum of several components $\{Z_k \mid k = 1, \ldots, K\}$

$$X = \sum_{k=1}^{K} \alpha_k Z_k \tag{10}$$

where $\{\alpha_k\}$ are the weights attributed to the components. Let G_k be the Gini index of concentration of component z_k as

$$G_k = 2 \operatorname{Cov}(z_k, F_k)/\mu_k \tag{11}$$

where F_k is the cumulative distribution of Z_k and μ_k its mean. The Gini index of x is by definition of a covariance as follows

$$\begin{aligned} G_x &= 2 \operatorname{Cov}(x, F_x)/\mu = 2 \operatorname{Cov}\left(\sum_{k=1}^{K} \alpha_k z_k, F_x \right) \Big/ \mu \\ &= 2 \sum_{k=1}^{K} \alpha_k \operatorname{Cov}(z_k, F_x)/\mu \end{aligned} \tag{12}$$

where $\operatorname{Cov}(z_k, F_x)$ is the concentration index of component z_k with respect to the distribution of x. Hence the Gini of a sum of variables can be decomposed as

$$G_x = \frac{1}{\mu} \sum_{k=1}^{K} \alpha_k \theta_k \cdot G_k \cdot \mu_k \tag{13}$$

where $\theta_k = \operatorname{Cov}(z_k, F_x)/\operatorname{Cov}(z_k, F_k)$ is a correlation coefficient between z_k and the ranking of x.

In practice, one ranks the various components Z_k according to the increasing order of X and regresses each individual component on the set of natural numbers (the Z_k are no necessarily ranked!). The slope coefficient of that regression is the numerator of θ_k. Then one ranks each Z_k individually, according to its own increasing order, and regresses it on the set of positive natural numbers to obtain the denominator of θ_k.[3]

IV. CONCLUSION

Since Gini's mean difference is the covariance between the random variable and its rank, the concentration index can be simply computed without resorting to graphical approximation when individual data points are available. Furthermore, the practitioner possesses now a simple measure of dispersion and concentration whose fields of application are not confined to income distribution. However, to obtain an

[3] Since the variance of natural numbers remains the same in the two regressions, there is no need for the adjustment made in equation (9).

unambiguous ranking of income distributions, the drawing of Lorenz curves is necessary to determine whether or not they intersect. For that purpose, precision is not relevant and grouped data can be used without losing efficiency.

Hebrew University of Jerusalem

Date of Receipt of Final Manuscript: February 1985.

REFERENCES

Atkinson, A. B. (1970). 'On the Measurement of Inequality', *Journal of Economic Theory*, Vol. 2, pp. 244–63.

Brown, T. A. C. and Mazzarino, G. (1984). 'Drawing the Lorenz Curve and Calculating the Gini Concentration Index from Grouped Data by Computer', *BULLETIN*, Vol. 46, pp. 273–8.

Dorfman, R. (1979). 'A Formula for the Gini Coefficient', *Review of Economics and Statistics*, Vol. 61, pp. 146–9.

Kendall, M. and Stuart, A. (1977). *The Advanced Theory of Statistics*, 4th Edition, London, Charles Griffin and Company.

Pyatt, G., Chen, C. and Fei, T. (1980). 'The Distribution of Income by Factor Components', *Quarterly Journal of Economics*, Vol. 94, pp. 451–73.

Shalit, H. and Yitzhaki, S. (1984). 'Mean Gini, Portfolio Theory, and the Pricing of Risky Assets', *Journal of Finance*, Vol. 39, pp. 1449–68.

Yitzhaki, S. (1982). 'Stochastic Dominance, Mean Variance and Gini's Mean Difference', *American Economic Review*, Vol. 72, pp. 178–85.

BULLETIN PRIZE

The Editors are pleased to announce that the 1984 Prize has been awarded to Dr Friedrich K. Struth for his article, *Modelling Expectations Formation with Parameter Adaptive Filters. An Empirical Application to the Livingston Forecasts*, published in the August issue. The Prize kindly offered by our Publishers is valued at £250.