Journal of Development Economics 12 (1983) 75–81. North-Holland Publishing Company

AN APPROXIMATION TO THE GINI COEFFICIENT FOR A POPULATION BASED ON SPARSE INFORMATION FOR SUB-GROUPS

Michael BRAULKE

Universität Konstanz, D-7750 Konstanz 1, West Germany

Received June 1981, final version received March 1982

Based on the assumption that the sub-groups' income distributions follow a Pareto distribution, a simple approximation to the Gini coefficient for the entire population is developed that requires merely information on the group Ginis, the group mean incomes and the shares in total population. A comparison with actual data indicates a satisfactory accuracy of this approximation. It is finally demonstrated that rather than to use it as an approximation to the aggregate Gini coefficient it may reversely serve as an approximative decomposition device that works on a minimum of group-specific information.

1. Introduction

It is well known that the Gini coefficient is, in general, not additively decomposable in the sense that it can be neatly split into a (weighted) sum of sub-group Ginis and a Gini capturing inequality between groups. As Bhattacharya and Mahalanobis (1967) have shown, an exact decomposition of the Gini coefficient gives rise to a third term which depends on the extent to which the sub-group income distributions overlap and which vanishes only if there is no overlapping.[1] It is therefore usually impossible to derive the aggregate Gini coefficient on the basis of the knowledge of the sub-group Ginis, the group mean incomes and the group shares in population alone. In addition, one has to have information on the extent of overlapping, which means essentially that the entire income distribution by groups has to be known. It is obvious that in applied work such extensive information will, in general, not be available and it may therefore be worthwhile to have an approximative aggregation that can do without it.

It is the purpose of this short paper to develop an approximation to the aggregate Gini coefficient for a two-sector economy that requires solely information on the sector-specific Ginis, the mean incomes and the population shares. This is certainly a modest informational requirement as compared to the information needed for an exact decomposition, yet in

[1] Compare also the appealing interpretation of these three components in Pyatt (1976).

0304–3878/83/0000–0000/$03.00 © 1983 North-Holland

practical application even this little will often turn out to be more than what is readily available. As regards the question of precision, our approximation appears to perform rather well. This will be demonstrated on the basis of actual data for various developing countries in the concluding section of this note. There we will also briefly indicate how the approximation may conversely be utilized as a simple quasi decomposition of the aggregate Gini coefficient.

2. The approximation

The simplifying assumption that the sectoral incomes follow a Pareto distribution is the basis of our approximation to the aggregate Gini coefficient. The sectoral Lorenz curves, y_i, will then belong to the one-parameter family

$$y_i = 1-(1-x)^{\alpha_i}, \qquad 0 \leqq \alpha_i \leqq 1, \tag{1}$$

where y_i denotes the share in sector i's income of the poorest fraction x_i of the population of that sector. The parameter α_i characterizes the sectoral income distribution and may itself be regarded as a measure of equality within a sector that ranges from perfect inequality ($\alpha_i = 0$) to perfect equality ($\alpha_i = 1$). This is already evident from (1) but also reflected in the fact that the sectoral Gini coefficients, G_i, associated with (1) amount to

$$G_i = (1-\alpha_i)/(1+\alpha_i). \tag{2}$$

Given (1), the derivation of the sectoral Gini coefficients is simple enough. The derivation of the associated aggregate Gini coefficient for the entire population, G, is hardly more demanding but it is tiresome and, if there are many sectors, it would become rather unwieldy. Accordingly, we will merely indicate the steps which are involved in the aggregation and deal only with the case of two sectors. The first step is easy. Using the definition of a Lorenz curve, the sectoral income distributions can be directly obtained from (1) by differentiation. In a second step then these two income distributions have to be merged in an order-preserving way to establish the aggregate distribution for the entire economy. Here, some care has to be applied to the question how and in what range the two income distributions overlap. This investigation reduces essentially to determining which sector houses the poorest members of the society and necessitates the differentiation between two different constellations. The aggregate Lorenz curve is then found by integration, as is eventually the aggregate Gini coefficient for the entire population. Writing $\beta = \mu_1/\mu_2$ for the ratio of the first sector's mean income to the second sector's mean income and γ for the first sector's share in total population, the result for the two-sector economy is

$$G=\frac{\gamma^2\beta}{\gamma\beta+1-\gamma}\,\frac{1-\alpha_1}{1+\alpha_1}+\frac{(1-\gamma)^2}{\gamma\beta+1-\gamma}\,\frac{1-\alpha_2}{1+\alpha_2}+\frac{\gamma(1-\gamma)(1-\beta)}{\gamma\beta+1-\gamma}+\frac{2\gamma(1-\gamma)}{\gamma\beta+1-\gamma}A, \quad (3)$$

where

$$\begin{aligned} A&=\beta\frac{(1-\alpha_1)^2}{1-\alpha_1\alpha_2}\left(\frac{\alpha_1\beta}{\alpha_2}\right)^{\alpha_1/(1-\alpha_1)} && \text{if} \quad \alpha_1\beta\leqq\alpha_2, \\ &=\beta-1+\frac{(1-\alpha_2)^2}{1-\alpha_1\alpha_2}\left(\frac{\alpha_1\beta}{\alpha_2}\right)^{\alpha_2/(\alpha_2-1)} && \text{if} \quad \alpha_1\beta\geqq\alpha_2. \end{aligned}$$

Thus, given sectoral Lorenz curves of the Pareto type as assumed in (1), the aggregate Gini coefficient for the two-sector economy is simply a function of the two distribution parameters α_1, α_2, of the ratio of sectoral mean incomes β, and of the distribution of population as represented by the first sector's share γ. We can therefore write $G=G(\alpha_1,\alpha_2,\beta,\gamma)$. Or, since the α_i and the G_i are one to one, since $\beta=\mu_1/\mu_2$ and since $\gamma=n_1/(n_1+n_2)$, where n_i denotes population in sector i, we could equally well write $G=H(G_1,G_2;\mu_1,\mu_2;n_1,n_2)$. This implies that (3) is in fact a decomposition of the aggregate Gini coefficient in the sense of Cowell and Shorrocks (1980).[2]

For various special cases, G is simplified considerably. If the population is concentrated in either of the sectors ($\gamma=1$ or $\gamma=0$), the aggregate Gini reduces, of course, to the corresponding sectoral Gini, i.e., we have from (3) and (2)

$$\begin{aligned} &G(\alpha_1,\alpha_2,\beta,1)=(1-\alpha_1)/(1+\alpha_1)=G_1, \\ &G(\alpha_1,\alpha_2,\beta,0)=(1-\alpha_2)/(1+\alpha_2)=G_2. \end{aligned} \quad (4)$$

Also, setting $\alpha_1=\alpha_2=\alpha$ and taking the limit of (3) as α approaches 1 from below, we find

$$G(1,1,\beta,\gamma)=\left|\frac{\gamma(1-\gamma)(1-\beta)}{\gamma\beta+1-\gamma}\right|. \quad (5)$$

Thus, for the case of perfect equality within both sectors ($\alpha_1=\alpha_2=1$), the aggregate Gini coefficient does indeed collapse to the simple formula (5) which Knight (1976) derived already by direct calculation.

[2] This may sound like, but is certainly not, in conflict with Cowell and Shorrock's finding that the Gini coefficient does *in general* not belong to the class of decomposable or 'aggregative' inequality measures.

We mention these special cases explicitly because they may help to understand more clearly the substance of our approximation (3) to the aggregate Gini coefficient. Note first that $\gamma\beta/(\gamma\beta+1-\gamma)$ and $(1-\gamma)/(\gamma\beta+1-\gamma)$ represent the first and the second sector's share in aggregate income, respectively. The first two terms on the right-hand side of (3) are consequently nothing but weighted sectoral Ginis where the weights are products of the respective sector's shares in income and in population. They may therefore be interpreted to represent the component in aggregate inequality that is due to within-sector inequality (if there is no overlapping). The third term or, if $\beta>1$, the negative of the third term in (3) has already been identified as the amount of inequality that would arise if there were no inequality within sectors. It can therefore be seen as representing the component that reflects inequality between sectors.[3] Now, since these components correspond so far exactly to the first two components in the Bhattacharya–Mahalanobis–Pyatt framework, the final term of the right-hand side of (3) must correspond then, if properly adjusted,[4] to the degree of inequality that is due to overlapping of sectoral income distributions. Our approximation (3) to the aggregate Gini coefficient, which is exact only if the sectoral income distributions are exactly Pareto, is thus, in a sense, an approximation of this third term due to overlaps.[5]

3. Applications

When it comes to using the approximation (3), knowledge of the individual parameters α_1, α_2, β and γ is indispensable; this is exactly where a practical application will find its natural limitation. From the point of view of data availability, the most difficult part is certainly to find the distribution parameters α_i. If the sectoral Ginis are known, one may, of course, solve for the α_i directly on the basis of (2). If, instead, information on the sectoral Lorenz curves is given, one would perhaps prefer to estimate the α_i by fitting (1) to the known points on the Lorenz curves. Compatible data on sectoral mean incomes may also be difficult to come by, but if they exist, $\beta=\mu_1/\mu_2$ can immediately be determined. Finally, if γ is not available already, knowledge of aggregate mean income μ will do as well, since γ may be retreived also from the definitorial identity $\mu=\gamma\mu_1+(1-\gamma)\mu_2$.

For the purpose of checking on the accuracy of the approximation formula (3), the compilation of national income statistics by Jain (1975) proved to be

[3] As Shorrocks (1980, pp. 624f.) has stressed, the definition of what should reasonably be termed, e.g., the between-sectors component in aggregate inequality is not at all clear. Nevertheless, we continue here to use this ambiguous terminology because it is well established within the Bhattachary–Mahalanobis–Pyatt framework.

[4] If $\beta\leq 1$, an adjustment is not required. But if $\beta>1$, the term reflecting between-sectors inequality is the negative of the third term on the right-hand side of (3) so that the term due to overlaps becomes $2\gamma(1-\gamma)(A+1-\beta)/(\gamma\beta+1-\gamma)$.

[5] I owe this interesting interpretation to an anonymous referee.

Table 1

Actual and approximated aggregate Gini coefficients for 10 developing countries.

Country	Year/ source[a]	Parameters[b]				Aggregate Gini coefficients		Error
		α_1	α_2	β	γ	Actual	$G(\alpha_1, \alpha_2, \beta, \gamma)$	(8)–(7)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Bangladesh	1967(2)	0.499	0.430	0.671	0.966	0.3420	0.3391	−0.003
Brazil	1970(4)	0.381	0.285	0.356	0.423	0.5770	0.5749	−0.002
Colombia	1970(6)	0.355	0.289	0.431	0.388	0.5615	0.5642	0.003
Honduras	1968(1)	0.346	0.333	0.186	0.521	0.6188	0.6278	0.009
India	1968(5)	0.355	0.366	0.741	0.790	0.4775	0.4804	0.003
Korea	1971(7)	0.527	0.495	0.535	0.376	0.3601	0.3695	0.009
Malaysia	1970(3)	0.355	0.326	0.472	0.719	0.5179	0.5254	0.008
Pakistan	1967(2)	0.506	0.438	0.709	0.751	0.3551	0.3551	0.000
Philippines	1971(4)	0.364	0.372	0.481	0.700	0.4941	0.5048	0.011
Sri Lanka	1970(3)	0.480	0.418	0.588	0.868	0.3771	0.3767	−0.000

[a] The source codes (given in brackets) are Jain's [Jain (1975)].

[b] All parameters are derived directly from the information contained in Jain (1975). Subscript 1 represents here the rural and 2 accordingly the urban sector. As the income statistics refer mostly to household income, β stands for the ratio of mean rural to mean urban household income and γ consequently for the share of rural households in the total number of households.

very useful since it contains for a number of countries not just the actual aggregate Gini coefficients but also compatible disaggregated information that permits the calculation of the parameters α, β and γ. The results of the comparison between the actual aggregate Gini coefficients and their approximation $G(\alpha_1, \alpha_2, \beta, \gamma)$ based on sector-specific data are given in table 1. Note that the discrepancy between the two remains, in most cases, markedly below one percentage point. Judging from these minimal errors, the accuracy of the approximation $G(\alpha_1, \alpha_2, \beta, \gamma)$ appears to be quite acceptable.

Until now the function $G(\alpha_1, \alpha_2, \beta, \gamma)$ was seen primarily as an approximation to the aggregate Gini coefficient, if there is only minimal information on the sectoral income distributions. But since $G(\alpha_1, \alpha_2, \beta, \gamma)$ can be calculated for any set of parameters, including hypothetical ones, it may reversely serve as well as the basis for a wide variety of decomposition exercises. To give just one example, suppose that the change in the aggregate Gini coefficient between two points in time is to be analyzed. If sufficient sector-specific information exists, a meaningful approach could consist of considering a tautological extension of the form

$$\begin{aligned} G(\alpha^1, \beta^1, \gamma^1) - G(\alpha^0, \beta^0, \gamma^0) = {} & [G(\alpha^1, \beta^1, \gamma^1) - G(\alpha^0, \beta^1, \gamma^1)] \\ & + [G(\alpha^0, \beta^1, \gamma^1) - G(\alpha^0, \beta^0, \gamma^1)] \\ & + [G(\alpha^0, \beta^0, \gamma^1) - G(\alpha^0, \beta^0, \gamma^0)], \end{aligned} \qquad (6)$$

Table 2

Decomposition of change in aggregate inequality for 4 developing countries.

Country	Time span/	Change in aggregate Gini coefficient accounted for by change in[a]				
			Within-	Between-		
			sectors	sectors		Error
Country	source[b]	Actual	component	component	Migration	(3–4–5–6)
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Brazil	1960(4)–1970(4)	0.0724	0.0436	0.0215	−0.0001	0.0074
India	1965(3)–1968(5)	0.0566	0.0799	−0.0274	0.0088	−0.0048
Pakistan	1964(1)–1967(2)	−0.0313	−0.0373	−0.0030	0.0041	0.0049
Philippines	1961(2)–1971(4)	−0.0187	0.0158	−0.0201	−0.0061	−0.0083

[a] Calculated on the basis of data contained in Jain (1975). As to the definition of the components, see (6) in the text.
[b] The source codes (in brackets) are again Jain's (1975).

where the superscript indicates the time of reference and α is written in short for (α_1, α_2). When checking which parameters are being held constant within each of the three brackets, it will be apparent that the right-hand side of (6) decomposes the overall change in the aggregate Gini coefficient into three changes which may be interpreted to reflect, in this order, the change in the within-sectors component, the change in the between-sectors component, and a pure migration effect.[6] Since in practical application the single components need not always work in the same direction, such a decomposition may in fact unveil changes in the underlying elements that remain hidden at the aggregate level. For two of the few countries listed in table 2 for which Jain's compilation contain the necessary information to carry out the decomposition exercise suggested in (6), the change in the within-sectors component tends indeed to offset the change in the between-sectors component of aggregate inequality.

4. Summary

Based on the assumption that the sectoral income distributions follow a Pareto distribution, a simple approximation to the aggregate Gini coefficient for a two-sector economy was developed which works on a minimum of information. As a matter of fact, merely the sectoral Gini coefficients, the mean incomes, and the distribution of population between sectors are needed to apply it. Its accuracy appears to be satisfactory and it should therefore be particularly useful, at least as a first investigative step, in the analysis of the income distribution of developing countries where the data base is small and the dual concept represents a meaningful approach, i.e., of countries with a distinctive difference between the export sector, the modern sector or the urban sector etc. and the rest of the economy.

[6] We should add that there are five alternative decompositions of the form (6) suggesting the same interpretation but differing by the period of reference of the arguments being held fixed in each of the differences $[G(\cdot)-G(\cdot)]$. The actual choice of what is eventually labeled the change in the within-sectors component etc. is consequently somewhat arbitrary. However, in practical application it appears to make little difference which definition is chosen as long as the recorded changes in the arguments (α, β, γ) are not too dramatic.

References

Bhattacharya, N. and M. Mahalanobis, 1967, Regional disparities in household consumption in India, American Statistical Association Journal 62, 143–161.

Cowell, F.A. and A.F. Shorrocks, 1980, Inequality decomposition by population subgroups, London School of Economics discussion paper (London).

Jain, S., 1975, Size distribution of income, A compilation of data (The World Bank, Washington, DC).

Knight, J.B., 1976, Explaining income distribution in less developed countries: A framework and an agenda, Oxford Bulletin of Economics and Statistics 38, 161–177.

Pyatt, G., 1976, On the interpretation and disaggregation of Gini coefficients, Economic Journal 86, 243–255.

Shorrocks, A.F., 1980, The class of additively decomposable inequality measures, Econometrica 48, 613–625.