# The Gini Coefficient for a Mixture of Ln-Normal Populations*

Alwyn Young
London School of Economics (December 2011)

Abstract

I present a formula which allows for the calculation of the Gini coefficient when the overall population distribution is unknown, but there is some information on the Gini coefficients or moments of population sub-groups. When applied to data on the global and US income distribution, it proves to be extremely accurate, producing estimates with errors that are small fractions of one percent.

---

*The programs and data extracts used in this paper are available on my website http://personal.lse.ac.uk/YoungA/.

One often has limited information on the moments of sub-groups of a population and would like to come to an assessment of the overall degree of inequality within the metric given by the famous Gini coefficient. For example, drawing from diverse sources, one might only know the mean income and Gini coefficient of each population sub-group. Aitchison and Brown (1957) showed that if a population's income is ln-normally distributed, then its Gini coefficient is given by $G = 2N[\sigma/\sqrt{2}]-1$, where $\sigma$ is the standard deviation of ln income. In this note I present an extension of Aitchison and Brown's theorem to the case of a population formed from a mixture of ln-normal distributions. While the resulting aggregate distribution is not ln-normal, its Gini coefficient is easily calculated using a few moments of the sub-populations. Drawing on detailed estimates of the distribution of income globally and within the United States, I show that the formula produces extraordinarily accurate estimates of Gini coefficients, i.e. for the purposes of calculating the Gini aggregate populations are very closely approximated as mixtures of ln-normal sub-populations. I begin by presenting the formula, then examine its accuracy in datasets with full population information, and finally illustrate its usefulness in calculating Ginis for datasets with incomplete distributional information.[1]

## I. The Ln-Normal Mixture Gini

I begin by stating the central result of the paper, concerning the Gini coefficient for a mixture of ln-normal distributions:

Theorem: Consider a population composed of N sub-groups. Within each sub-group i, income is ln-normally distributed with mean $Y_i=exp[\mu_i+.5\sigma_i^2]$, where

[1]Readers of this paper have brought to my attention the fact that Modalsli (2011) independently derives the same formula for the Gini coefficient of a mixture of ln-normals. Modalsli's emphasis is on providing an interesting application of the formula to the study of income inequality in pre-industrial societies, while mine is on confirming its accuracy as an approximation of actual income distributions.

$\mu_i$ and $\sigma_i^2$ are the mean and variance of ln income. With $\omega_i$ denoting the population share of sub-group *i*, *Y* aggregate mean income and *N[]* the cumulative standard normal distribution, the Gini coefficient for the aggregate population is given by:

$$(1)\quad G = \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\omega_i\omega_j Y_i}{Y}\left(2N\left[\frac{\ln(Y_i)-\ln(Y_j)+.5\sigma_i^2+.5\sigma_j^2}{\sqrt{\sigma_i^2+\sigma_j^2}}\right]-1\right)$$

The proof of (1) is lengthy, involving a number of properties of the ln-normal distribution, and is left for the appendix. It is apparent, however, that when N=1, i.e. there is only one sub-group, the formula reduces to Aitchison and Brown's result $G = 2N[\sigma/\sqrt{2}]-1$. While (1) is not part of a formal estimation framework, it operates in the spirit of a semi-parametric approximation of the population Gini, allowing $Y_i$ and $\sigma_i$ to vary freely across groups while imposing structure on the distribution within each group.[2]

Simple substitution into (1) allows one to calculate the contribution of within and between group inequality to aggregate inequality:

Corollary: For the population described above, inequality in the absence of differences between group means, $Y_i = Y$, is given by:

$$(2)\quad G = \sum_{i=1}^{N}\sum_{j=1}^{N}\omega_i\omega_j\left(2N\left[\sqrt{\sigma_i^2+\sigma_j^2}/2\right]-1\right)$$

while inequality in the absence of within group differences, $\sigma_i^2 = 0$, is given by:

$$(3)\quad G = \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\omega_i\omega_j Y_i}{Y}\text{sign}[Y_i - Y_j]$$

[2]In a formal estimation setting, one could allow *Y* and $\sigma$ to be smooth non-parametric functions of observable sub-group characteristics.

This decomposition is not additive,[3] in that (2) + (3) is greater than (1). I view this as a positive feature, rather than a weakness, as it matches one's intuition that the impact on inequality of differences between group means and within group dispersion might each be ameliorated by the presence of the other.[4]

## II. Confirming the Accuracy of the Ln-Normal Mixture Gini

As a preliminary, I must consider how one might come by the parameters necessary to calculate the mixture Gini. The most likely situation, I believe, is the one mentioned in the introduction, where one receives from other sources data on the mean income of each sub-group ($Y_i$) and its Gini coefficient ($G_i$). In this case, one can invert the Aitchison-Brown formula to find $\sigma_i=\sqrt{2N^{-1}[(G_i+1)/2]}$ and apply (1)-(3) above.[5] Alternatively, one might be in the situation of having to estimate the parameters of the ln-normal distribution or use moment information provided by others. Given information on individual incomes in a sub-group one can calculate the population mean and variance of $lnY$ to arrive at the maximum likelihood estimates of $\mu_i$ and $\sigma_i^2$ which, using the formula $Y_i = exp(\mu_i + .5\sigma_i^2)$, provides the information on $Y_i$ and $\sigma_i^2$ necessary to use equations (1)-(3). Alternatively, one can take a standard methods of moments approach and, noting

[3]As shown by Bourguignon (1979) and Shorrocks (1980), the only income distribution measures which are additively separable and homogenous of degree zero in income and population are the square of the coefficient of variation and Theil's population and income weighted entropy indices. Of these, the only one with weights that are not dependent upon the distribution of income is the population weighted index, or mean log deviation.

[4]Thus, imagine in your mind's eye a diagram with the population density on the vertical axis and incomes on the horizontal axis. Next, consider the case where there are two groups, with differences between group means but no within group inequality (i.e. the distribution is made up of two spikes). As one introduces within group inequality, the outer tails of the two distributions move apart, but their inner tails move together. Thus, the impact of within group inequality is less than it would be if there were no differences between means. A similar mental exercise explains why the introduction of between group inequality has less of an impact in an environment with pre-existing within group inequality. In thinking through these exercises, it is useful to bear in mind that the Gini is calculated from the absolute differences of incomes, not their squares.

[5]I assume that the population weights $\omega_i$ are always independently available.

| Table I: Accuracy of Estimates of Ln-Normal Distribution | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 20 observations | | | 100 observations | | | 1000 observations | | |
| | Gini | MM | MLE | Gini | MM | MLE | Gini | MM | MLE |
| σ | | | | | | | | | |
| mean ln error | -.0986 | -.0697 | -.0267 | -.0208 | -.0140 | -.0028 | -.0024 | -.0014 | .0003 |
| rmse | .2111 | .1849 | .1662 | .0897 | .0784 | .0698 | .0297 | .0258 | .0221 |
| E(Y) | | | | | | | | | |
| mean ln error | -.0336 | -.0336 | .0047 | -.0025 | -.0025 | .0075 | -.0001 | -.0001 | .0015 |
| rmse | .2748 | .2748 | .2696 | .1267 | .1267 | .1231 | .0420 | .0420 | .0390 |

Notes: rmse – root mean squared ln error. Each entry involves 1000 estimation runs using the sample sizes listed of the parameters of a ln normal random variable with $\mu=0$ and $\sigma=1$. Gini method calculates and inverts the Gini coefficient to estimate $\sigma$ and uses the population mean income to estimate $E(Y)$. MM method uses the square root of 2 times the difference between the ln of the mean income and the mean of ln income to estimate $\sigma$ and uses the population mean income to estimate $E(Y)$. MLE uses the mean and standard deviation of the ln of income to estimate $\mu$ and $\sigma$ and calculates $E(Y)=exp(\mu+.5\sigma^2)$.

that $\sigma^2 = 2*[ln[E(Y)]-E(lnY)]$, use the mean of income and the difference between the ln of the mean of income and the mean of the ln of income to calculate $Y_i$ and $\sigma_i^2$.

Given the maintained distributional assumption, in a sampling framework the maximum likelihood estimates (MLE) will provide the most efficient estimates of the parameters of the distribution, although both the method of moments (MM) and Gini inversion[6] (G) approaches will be consistent. This is illustrated in Table I, where I simulate the estimation of the parameters of a ln normal distribution with $\mu=0$ and $\sigma=1$ in sample sizes of 20, 100 and 1000 observations. Across all sample sizes the MLE has a smaller root mean squared ln error, although the advantage largely disappears in larger samples[7] and is minimal, in all samples, in

[6]This is, in itself, a methods of moments estimator, but I distinguish it from the other MM estimator by calling it the Gini inversion or, later, Gini mixture estimator.

[7]Thus, ironically, while the MLE's known superiority is asymptotic, in this case it is greatest in small samples.

the estimation of functions of the parameters such as $E(Y) = exp(\mu+.5\sigma^2)$. Of greater interest, however, is the sensitivity of each estimator to deviations from the distributional assumptions. In this regard, I note that the variance of $lnY$ is quite sensitive to skewness. Consequently, the MLE estimates of $\sigma^2$ will be more sensitive to a slight departure from the strict distributional assumptions, as will be seen below.

To assess the accuracy of each procedure, in what follows I will take complete income distribution data, calculate the "actual" Gini by integrating the income distribution, and then separately calculate mixture Ginis, deriving the necessary parameters by (a) calculating sub-group Ginis and population means (the Gini inversion mixture); (b) using the methods of moments on the sub-group population data (the MM mixture); and (c) using the maximum likelihood equations on the sub-group population data (the MLE mixture). Obviously, if one has micro data for each sub-group, one can integrate the entire distribution to calculate the Gini and there is no need to actually apply the mixture formula at all. However, the real world application I have in mind is one where one receives estimates from others and/or is in the position of having micro data for only some sub-groups, and in these cases it is useful to know how accurate the different procedures tend to be. In each case I will start by using all the population data, to illustrate bias in the presence of specification error, and then draw repeated samples of 20, 100 and 1000 observations from the population data to illustrate the combination of bias and sampling variability. In practice, I believe, most estimates of population moments will involve samples of at least 100, but I include calculations with smaller samples to allow the sampling superiority of the MLE estimator to overcome some of its sensitivity to distributional assumptions.

(a) The Global Gini (Sala-i-Martin 2006)

To begin, I draw on Xavier Sala-i-Martin's (2006) estimates of the distribution of income inequality by country around the world. Sala-i-Martin uses Penn-World Table data and the UNU income inequality database to arrive at estimates of the distribution of the population by 100 income categories in 126 countries from 1970 to 2000.[8] Table II below compares the results arrived at by applying ln normal approximation formulas with those arrived at by integrating the entire distribution provided by Sala-i-Martin. Although I calculate measures for each year between 1970 and 2000, to save space I only present decadal numbers and summary statistics.

As shown in the table, global Ginis calculated as a mixture of ln-normal populations, whose parameters are estimated from the Gini coefficients of the sub-populations, are extraordinarily accurate, with a root mean squared ln error of three hundredths of a percent. The error in the estimate of the contribution of within inequality is greater, but still only half of one percent, while the calculation of between inequality is exactly correct, as the estimate of $Y_i$ uses the population data. MM methods increase the root mean squared error on within inequality by about a half, while MLE methods double it. Combined with the MLE error in the estimation of levels[9] and between inequality, the MLE rmse on aggregate inequality is about half of one percent, vastly greater than the other methods, but still quite small. As shown in the final two columns of the table, crude ln-normal approximations, calculating one $\sigma$ for the entire global population and applying the original Aitchison-Brown formula, perform worst of all, with the MLE methods producing a root mean squared ln error of about 4 percent.

[8]To maintain consistent regional definitions, I recombine the 14 post-1989 former Soviet Republics back into the Soviet Union.

[9]Recall that the MLE estimate of $Y_i = exp(\mu_i + .5\sigma_i^2)$, which generally does not equal the population $Y_i$.

| Table II: Global Inequality (Sala-i-Martin 2006 data) | | | | | | |
|---|---|---|---|---|---|---|
| | Actual Gini | Gini Mixture | MM Mixture | MLE Mixture | MM LnNormal | MLE LnNormal |
| Aggregate Gini | | | | | | |
| 1970 | .656 | .656 | .656 | .660 | .646 | .626 |
| 1980 | .663 | .663 | .663 | .666 | .654 | .635 |
| 1990 | .655 | .655 | .655 | .659 | .645 | .627 |
| 2000 | .639 | .639 | .639 | .643 | .634 | .631 |
| mean ln error | | -.0001 | .0000 | .0052 | -.0135 | -.0382 |
| rmse | | .0003 | .0003 | .0052 | .0137 | .0394 |
| Within Inequality (no between inequality) | | | | | | |
| 1970 | .365 | .362 | .363 | .362 | .371 | .377 |
| 1980 | .378 | .375 | .374 | .373 | .380 | .383 |
| 1990 | .398 | .396 | .395 | .395 | .399 | .403 |
| 2000 | .423 | .421 | .420 | .421 | .426 | .433 |
| mean ln error | | -.0057 | -.0088 | -.0109 | .0055 | .0162 |
| rmse | | .0058 | .0088 | .0111 | .0070 | .0174 |
| Between Inequality (no within inequality) | | | | | | |
| 1970 | .565 | .565 | .565 | .569 | .571 | .559 |
| 1980 | .572 | .572 | .572 | .575 | .577 | .564 |
| 1990 | .559 | .559 | .559 | .563 | .554 | .532 |
| 2000 | .524 | .524 | .524 | .527 | .521 | .508 |
| mean ln error | | 0 | 0 | .0064 | .0006 | -.0278 |
| rmse | | 0 | 0 | .0064 | .0101 | .0329 |

Notes: rmse = root mean squared ln error. Within inequality measures calculated by rescaling individual incomes so that average country incomes are all equal. Between income inequality measures calculated by giving each individual within a country the mean country income. Mean ln error and rmse calculated using 31 observations between 1970 and 2000. Gini, MM and MLE refer to methods of calculating the sub-group parameters, as described earlier above. Mixture uses equations (1)-(3) earlier, while LnNormal (in the final two columns) simply applies the Aitchison-Brown formula to the aggregate population.

| Table III: Accuracy of Sampled Mixture Estimates of Global Gini (2000) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 20 observations | | | 100 observations | | | 1000 observations | | |
| | Gini | MM | MLE | Gini | MM | MLE | Gini | MM | MLE |
| Aggregate Gini | | | | | | | | | |
| mean ln error | -.0059 | .0003 | .0151 | -.0005 | .0010 | .0085 | .0001 | .0005 | .0068 |
| rmse | .0257 | .0242 | .0295 | .0115 | .0112 | .0146 | .0037 | .0036 | .0078 |
| Within Inequality | | | | | | | | | |
| mean ln error | -.0645 | -.0459 | -.0143 | -.0186 | -.0167 | -.0071 | -.0075 | -.0093 | -.0059 |
| rmse | .0751 | .0587 | .0400 | .0252 | .0230 | .0175 | .0094 | .0108 | .0078 |
| Between Inequality | | | | | | | | | |
| mean ln error | .0119 | .0119 | .0178 | .0030 | .0030 | .0087 | .0001 | .0001 | .0061 |
| rmse | .0440 | .0440 | .0492 | .0204 | .0204 | .0234 | .0065 | .0065 | .0093 |
| Notes: as in Tables I and II above. The rescaling of income figures to perform within and between inequality calculations is done on the basis of the sampled means. | | | | | | | | | |

The sensitivity of the different measures to sampling variability is explored in Table III, where I focus on the calculation of the mixture Ginis for the year 2000. The smaller sampling variability of the MLE procedure overcomes its relative bias to produce a smaller root mean squared ln error only in the estimation of the contribution of within inequality. Otherwise, extremely small and unlikely samples of 20 individuals per sub-group are necessary before its rmse approaches those of the other estimators, although its bias remains substantially larger. Finally, I note that the sampling bias and rmse of the MM mixture estimator is quite close to that of the Gini mixture estimator.

Further evidence on the greater sensitivity to distributional assumptions, for the purposes of calculating the Gini coefficient, of the MLE estimator is presented in Table IV below. For each of the 126 country x 31 year observations I calculate the country specific Gini coefficient by integrating the income distribution and then compare it to the MM and MLE ln-normal approximations using the complete population data, applying the Aitchison and Brown one region

| Table IV: Ln-Normal Approximation of Country Ginis (Sala-i-Martin 2006 data) | | | | |
|---|---|---|---|---|
| | Full Data | | Eliminating skewed 1% of the income distribution | |
| | MM Gini | MLE Gini | MM Gini | MLE Gini |
| mean ln error | -.0053 | -.0483 | -.0032 | -.0106 |
| rmse | .0204 | .1108 | .0154 | .0534 |
| regression $R^2$ | .9946 | .8405 | .9946 | .9344 |

Notes: 3906 country x year observations. $R^2$ is of regression of the ln of the ln-normal approximation Gini on the ln of the actual Gini. Skew adjustment involves eliminating top (bottom) 1% of the population in countries with positively (negatively) skewed ln income. Other terms as in tables above.

formula.[10] As can be seen, the ln-normal approximation with parameters calculated using the MLE is quite poor, with a mean ln error of -4.8 percent and a root mean squared ln error of 11 percent. Regressing the ln MLE Gini on a constant and the ln actual Gini, I get an $R^2$ of .841. In contrast, the MM approximation has a mean error of -.5 percent, a rmse of 2 percent, and an $R^2$ of .995. In the right hand panel I remove the top one percent of population observations in countries with positive skewness in their income distribution, and the bottom one percent in countries with negative skewness, recalculating all of the actual, MM and MLE Ginis. This vastly improves the relative accuracy of the MLE Gini, reflecting my comments earlier above on the sensitivity of the various measures to skewness.

(b) The US Gini (Krueger-Perri 2006)

As a second application, I take the Krueger-Perri (2006) cleaned data files for the United States consumer expenditure survey from 1980 to 2003. Krueger and Perri remove observations with anomalous data (e.g. no food expenditure or

[10]Obviously the Gini inversion cannot be applied in this case, as it simply returns itself.

only food expenditure) and impute expenditures for the flow of services from durable goods.[11] As their data set divides the United States into only four regions, whose consumption levels are quite similar, there is not much point in pursuing a geographical breakdown. Instead, I cross the sex of the survey reference person, whether or not they are in their prime working years (ages 30 to 60), and their educational attainment divided into six categories[12], to divide the population into 24 sub-groups.

As shown in Table V, the Gini mixture ln-normal calculation of the US consumption Gini has an average ln bias of 7 hundredths of a percent and a root mean squared error of a quarter of a percent. Its error, in the measurement of within inequality, is on the order of one half of a percent. As before, the MM and MLE mixtures perform less well, with the MLE mixture in particular showing 4 to 7 times the rmse of the Gini mixture. Nevertheless, both mixtures achieve about one half the error, in the aggregate Gini, of the comparable estimates arrived at by crudely applying the Aitchison-Brown formula to the aggregate population.

Although the accuracy achieved by the Gini mixture in approximating the aggregate US Gini is below that achieved in the analysis of global data earlier above, the two sets of results are actually remarkably consistent. In both cases the mean ln error and root mean squared ln error on the measurement of the contribution of within equality are about -.5 and +.5 percent, respectively. In the US case, however, the contribution of within group inequality is much greater than in the global case, and the contribution of between group inequality is much

[11] I take their ndpbe0 (benchmark) measure of consumption divided by the number of equivalent adults in the household as consumption per capita in the household. The Gini is then calculated over this measure, taking the total population as the sample sum of the household weight times the number of equivalent adults in the household. The sample sizes are around three to four thousand households per year.

[12] None or primary, some high school, high school graduate, some college, college graduate, and more than college.

| Table V: US Consumption Inequality (Krueger-Perri 2006 data) | | | | | | |
|---|---|---|---|---|---|---|
| | Actual Gini | Gini Mixture | MM Mixture | MLE Mixture | MM LnNormal | MLE LnNormal |
| Aggregate Gini | | | | | | |
| 1980 | .257 | .256 | .261 | .267 | .264 | .270 |
| 1990 | .278 | .278 | .282 | .282 | .283 | .284 |
| 2000 | .284 | .284 | .288 | .288 | .289 | .288 |
| mean ln error | | -.0007 | .0125 | .0152 | .0202 | .0307 |
| rmse | | .0025 | .0129 | .0178 | .0211 | .0343 |
| Within Inequality (no between inequality) | | | | | | |
| 1980 | .231 | .227 | .234 | .240 | .235 | .240 |
| 1990 | .244 | .243 | .247 | .249 | .248 | .249 |
| 2000 | .249 | .248 | .252 | .253 | .253 | .253 |
| mean ln error | | -.0050 | .0124 | .0185 | .0161 | .0191 |
| rmse | | .0056 | .0130 | .0212 | .0167 | .0216 |
| Between Inequality (no within inequality) | | | | | | |
| 1980 | .120 | .120 | .120 | .118 | .125 | .126 |
| 1990 | .141 | .141 | .141 | .140 | .140 | .141 |
| 2000 | .141 | .141 | .141 | .140 | .144 | .143 |
| mean ln error | | 0 | 0 | -.0067 | .0236 | .0367 |
| rmse | | 0 | 0 | .0088 | .0274 | .0410 |

Notes: As in Table II, except that mean ln error and rmse are calculated using 24 annual observations between 1980 and 2004.

smaller, as the differences between the mean consumption levels of the US sex x age x education sub-groups (on the order of 3 or 4 to 1 at most) are simply dwarfed by the observed 60+ fold differences in country mean income levels. If one thinks of the ln-normal mixture as providing a semi-parametric approximation to the population income distribution, then in the US case the non-parametric component is simply much smaller, leaving a bigger role for the error in the parametric approximation of sub-group distributions, which seems to be around half of a percent in both cases.

| Table VI: Accuracy of Sampled Mixture Estimates of US Gini (2000) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 20 observations | | | 100 observations | | | 1000 observations | | |
| | Gini | MM | MLE | Gini | MM | MLE | Gini | MM | MLE |
| Aggregate Gini | | | | | | | | | |
| mean ln error | -.0164 | .0033 | .0203 | -.0022 | .0116 | .0138 | .0004 | .0130 | .0119 |
| rmse | .0489 | .0446 | .0477 | .0208 | .0231 | .0239 | .0064 | .0143 | .0133 |
| Within Inequality | | | | | | | | | |
| mean ln error | -.0603 | -.0322 | -.0053 | -.0146 | .0036 | .0097 | -.0048 | .0116 | .0130 |
| rmse | .0762 | .0554 | .0451 | .0257 | .0207 | .0227 | .0082 | .0132 | .0145 |
| Between Inequality | | | | | | | | | |
| mean ln error | .0742 | .0742 | .0676 | .0179 | .0179 | .0107 | .0012 | .0012 | -.0061 |
| rmse | .1179 | .1179 | .1137 | .0505 | .0505 | .0478 | .0148 | .0148 | .0158 |

Table VI explores the impact of the combination of bias and sampling variability on the accuracy of the various measures in estimating the US Gini in the year 2000. Once again, the MLE mixture requires relatively small samples to match or improve upon (in the case of within inequality) the rmse of the other measures. While the MM mixture rmse is quite close to that of the Gini mixture in small samples, in larger samples of 1000 per sub-region the Gini inversion mixture is by far the most accurate, reinforcing the results of the tables above.

## III. Applying the Formula

In this section I present an example of a practical "real world" application of the formula. In constructing his dataset, Sala-i-Martin made a number of extrapolations and interpolations of the limited data in the UNU world income inequality database to calculate income inequality within each country year by year. One might wonder how these impacted his conclusions, namely that income inequality in the world is declining, particularly in the past couple of decades. To this end, I draw on two other datasets constructed using the UN database. Bhalla (2002) put together his own estimates of country specific Ginis in 1980 and 2000, taking in each case the most recent earlier year for which data were available or,

when none existed, the most recent later year. Jones and Klenow (2011) averaged Gini data that they found to be of acceptable quality from 1974-1986 and 1994-2006 to construct estimates for 1980 and 2000, respectively. While Bhalla and Jones-Klenow did not generate estimates of the full distribution of world income, their data provide enough information to apply the ln-normal mixture formula. I invert the Aitchison-Brown formula for a ln-normal population to convert each of the Bhalla and Jones-Klenow country Gini observations into estimates of $\sigma$, the standard deviation of ln income, and combine these with Sala-i-Martin's data on country mean levels, which he took from the Penn World tables.

Table VII presents the results of the comparison. In column (1) I report the global Gini calculated using the full distribution of income in Sala-i-Martin's 126 country sample. In column (2) I restrict attention to the 96 countries which overlap the Bhalla and Jones-Klenow data. 26 of the 30 countries lost in merging the samples are countries for which the UNU database had no observations whatsoever and Sala-i-Martin, as he explains, had to extrapolate from the trends of neighbours. As can be seen, in expanding the sample to include countries for which no data were available, Sala-i-Martin worked against his main conclusion (i.e. the decline in the second column is greater than in the first), because in the larger group the growth of within inequality was greater and the decline in between inequality smaller. The third column of the table recalculates the 96 country Sala-i-Martin global Gini using the Gini ln-normal mixture. The results are almost identical, as would be expected from the analysis earlier above. Finally, the last two columns report the 96 country ln-normal mixture Ginis calculated using the Bhalla and Jones-Klenow country inequality data. As can be seen, their estimates indicate a greater proportional decline in the global Gini. We can conclude that the extrapolations and interpolations Sala-i-Martin made to generate a complete country x year dataset were biased, if anything, against the results he put forward.

| Table VII: Global Ginis Using Alternative Estimates of Within Inequality | | | | | |
|---|---|---|---|---|---|
| | 126 countries | 96 countries | 96 countries, Gini ln-normal mixture | | |
| | Sala-i-Martin | Sala-i-Martin | Sala-i-Martin | Bhalla | Jones-Klenow |
| Aggregate Gini | | | | | |
| 1980 | .663 | .683 | .683 | .677 | .670 |
| 2000 | .639 | .645 | .645 | .638 | .629 |
| ln change | -.038 | -.058 | -.057 | -.060 | -.063 |
| Within Inequality (no between inequality) | | | | | |
| 1980 | .376 | .380 | .377 | .362 | .355 |
| 2000 | .423 | .418 | .416 | .397 | .379 |
| ln change | .115 | .097 | .097 | .093 | .064 |
| Between Inequality (no within inequality) | | | | | |
| 1980 | .572 | .591 | .591 | .591 | .591 |
| 2000 | .524 | .534 | .534 | .534 | .534 |
| ln change | -.087 | -.101 | -.101 | -.101 | -.101 |
| Notes: Bhalla and Jones-Klenow measures calculated using their country level Gini estimates, but Sala-i-Martin's data on country income levels. Consequently, the between estimates are identical to those of Sala-i-Martin. | | | | | |

## IV. Conclusion

With regards to the Gini coefficient, real world income distributions appear to be very closely approximated as mixtures of ln-normals. In calculating the sub-group parameters of the ln-normal, inverting the Aitchison-Brown formula for sub-group Ginis to arrive at estimates of the sub-group standard deviation of ln income produces the best results, with methods of moments estimation coming in second. Maximum likelihood estimates of the variance of ln income are quite sensitive to skewness, i.e. a deviation from the distributional assumption, producing less accurate estimates, outside of extremely small samples, of the aggregate Gini. In all cases, breaking the population down into sub-groups and calculating a mixture formula improves upon the application of the Aitchison-

Brown formula to the aggregate population. It is my hope that the formulas presented above will allow for the easy calculation and comparison of measures of inequality where income distribution data are incomplete.

## Appendix: Proof of the Theorem

As a preliminary to the proof, it is first necessary to review some of the properties of the ln-normal distribution.[13] We say that $x>0$ is distributed ln-normally if the ln of $x$ is distributed normally, so that the cumulative distribution and density functions of $x$, $F(x)$ and $f(x)$, are given by:

$$\text{(P1)}\quad F(x\,|\,\mu,\sigma^2) = N[(\ln x-\mu)/\sigma]$$

$$f(x\,|\,\mu,\sigma^2) = n[(\ln x-\mu)/\sigma]/x\sigma$$

where I use $N[]$ and $n[]$ to denote the cumulative distribution and density functions of the standard normal and where I keep track of the parameters $\mu$ and $\sigma^2$, the mean and variance of ln $x$, in describing the distribution of $x$ as it will be necessary to do so in what follows.

From the properties of the normal distribution, we know that if $X_1 \sim F(\mu_1,\sigma_1^2)$ then $X_1^b \sim F(b\mu_1,b^2\sigma_1^2)$, while if $X_1 \sim F(\mu_1,\sigma_1^2)$ and $X_2 \sim F(\mu_2,\sigma_2^2)$, then $X_1X_2 \sim F(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$. This allows for the very useful property

$$\text{(P2)}\quad \int_0^\infty F(ax^b|\mu_1,\sigma_1^2)\, f(x|\mu_2,\sigma_2^2)\,\mathrm{dx} \;=\; F(a|\mu_1-b\mu_2,\sigma_1^2+b^2\sigma_2^2)$$

(P2) is proven by noting that the probability $X_1X_2^{-b} = x$ is given by

$$\int_0^\infty f(xv^b|\mu_1,\sigma_1^2)f(v|\mu_2,\sigma_2^2)\,\mathrm{dv}$$

so that the probability $X_1X_2^{-b} \le a$ equals

[13] I review these properties, presented and derived by Aitchison and Brown (1957), as they are necessary to understand my extension of their Gini theorem to the case of a mixed distribution of ln-normal populations.

$$\int_0^a \int_0^\infty f(xv^b|\mu_1,\sigma_1^2)f(v|\mu_2,\sigma_2^2)\,\mathrm{dv}\,\mathrm{dx} = \int_0^\infty F(av^b|\mu_1,\sigma_1^2)\,f(v|\mu_2,\sigma_2^2)\,\mathrm{dv}$$

where I arrive at the last step by reversing the order of integration. Since $X_1X_2^{-b}\sim F(\mu_1-b\mu_2,\ \sigma_1^2+b^2\sigma_2^2)$, this establishes that the left and right-hand sides of (P2) are equal.

The jth moments of the ln-normal distribution are given by:

$$\text{(P3)}\quad \lambda^j(\mu,\sigma^2) = \int_0^\infty x^j f(x)\,\mathrm{dx} = \int_{-\infty}^\infty e^{jy} n(y)\,\mathrm{dy} = e^{j\mu+\frac{1}{2}j^2\sigma^2}$$

where the second equality follows from the substitution $y = ln(x)/\sigma$ and the last from the moment generating function of the normal distribution. The jth normalized incomplete moment of the ln-normal distribution, $F^j(x|\mu,\sigma^2)$, is itself a ln-normal cumulative density function:

$$\text{(P4)}\quad F^j(x\,|\,\mu,\sigma^2) = \frac{1}{\lambda^j(\mu,\sigma^2)}\int_0^x u^j f(u\,|\,\mu,\sigma^2)\,\mathrm{du}$$

$$= e^{-j\mu-\frac{1}{2}j^2\sigma^2}\int_0^x e^{j\ln u}\frac{1}{u\sigma\sqrt{2\pi}}\exp\left[-\frac{(\ln u-\mu)^2}{2\sigma^2}\right]\mathrm{du}$$

$$= \int_0^x \frac{1}{u\sigma\sqrt{2\pi}}\exp\left[-\frac{(\ln u-\mu-j\sigma^2)^2}{2\sigma^2}\right]\mathrm{du} \quad = \quad F(x|\mu+j\sigma^2,\sigma^2)$$

Taking the derivative of top and bottom right hand sides with respect to $x$, we have the corollary:

$$\text{(P5)}\quad \frac{x^j f(x\,|\,\mu,\sigma^2)}{\lambda^j(\mu,\sigma^2)} = f(x|\mu+j\sigma^2,\sigma^2)$$

With the above, we can proceed to the proof of the theorem. Let $p()$ denote the population density function of a population composed of $N$ subgroups, each of which is distributed ln-normally. Thus,

$$\text{(A1)}\quad p(x) = \sum_{i=1}^N \omega_i f(x|\mu_i,\sigma_i^2) \qquad Y = \sum_{i=1}^N \omega_i Y_i$$

where $\omega_i$ is the population share of group *i*, *Y* the aggregate population mean and $Y_i=\lambda^1(\mu_i,\sigma_i^2)$ the sub-group mean. The Gini coefficient is defined as one-half the relative mean difference, i.e. the mean-normalized expected value of the absolute difference between the incomes of any two individuals in the population:

$$\text{(A2)}\quad G = \frac{1}{2}\frac{1}{Y}\int_0^\infty\int_0^\infty |u-v|\,p(u)p(v)\,\mathrm{dv\ du}$$

$$= \frac{1}{2Y}\int_0^\infty\int_0^u (u-v)p(u)p(v)\,\mathrm{dv\ du} + \frac{1}{2Y}\int_0^\infty\int_u^\infty (v-u)p(u)p(v)\,\mathrm{dv\ du}$$

$$= \frac{2}{2Y}\int_0^\infty\int_0^u (u-v)p(u)p(v)\,\mathrm{dv\ du}$$

$$= \frac{1}{Y}\int_0^\infty\int_0^u (u-v)\sum_{i=1}^N \omega_i\, f(u|\mu_i,\sigma_i^2)\sum_{j=1}^N \omega_j\, f(v|\mu_j,\sigma_j^2)\,\mathrm{dv\ du}$$

$$= \sum_{i=1}^N\sum_{j=1}^N \frac{\omega_i\omega_j}{Y}T(i,j) \quad \text{where} \quad T(i,j) = \int_0^\infty\int_0^u (u-v)f(u|\mu_i,\sigma_i^2)f(v|\mu_j,\sigma_j^2)\,\mathrm{dv\ du}$$

One simplifies *T(i,j)* to a function of the cumulative normal by breaking the integral into two parts and using the properties described above:

$$\text{(A3)}\quad \int_0^\infty\int_0^u uf(u|\mu_i,\sigma_i^2)f(v|\mu_j,\sigma_j^2)\,\mathrm{dv\ du} - \int_0^\infty\int_0^u vf(u|\mu_i,\sigma_i^2)f(v|\mu_j,\sigma_j^2)\,\mathrm{dv\ du}$$

$$= \int_0^\infty uf(u|\mu_i,\sigma_i^2)F(u|\mu_j,\sigma_j^2)\,\mathrm{du} - Y_j\int_0^\infty f(u|\mu_i,\sigma_i^2)F(u|\mu_j+\sigma_j^2,\sigma_j^2)\,\mathrm{du}$$

$$= Y_i\int_0^\infty f(u|\mu_i+\sigma_i^2,\sigma_i^2)F(u|\mu_j,\sigma_j^2)\,\mathrm{du} - Y_j\int_0^\infty f(u|\mu_i,\sigma_i^2)F(u|\mu_j+\sigma_j^2,\sigma_j^2)\,\mathrm{du}$$

$$= Y_iF(1,\mu_j-\mu_i-\sigma_i^2,\sigma_i^2+\sigma_j^2) - Y_jF(1,\mu_j+\sigma_j^2-\mu_i,\sigma_i^2+\sigma_j^2)$$

$$= Y_i N\left[\frac{\sigma_i^2+\mu_i-\mu_j}{\sqrt{\sigma_i^2+\sigma_j^2}}\right] - Y_j N\left[\frac{\mu_i-\mu_j-\sigma_j^2}{\sqrt{\sigma_i^2+\sigma_j^2}}\right]$$

where I arrive at the second line by integrating and using (P4), the third line by using (P5), the fourth line by using (P2) with $a = b = 1$, and the fifth line by invoking (P1).

To finish, I note that:

$$\text{(A4)}\ \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\omega_i\omega_j}{Y}\left(Y_i N\left[\frac{\sigma_i^2+\mu_i-\mu_j}{\sqrt{\sigma_i^2+\sigma_j^2}}\right] - Y_j N\left[\frac{\mu_i-\mu_j-\sigma_j^2}{\sqrt{\sigma_i^2+\sigma_j^2}}\right]\right)$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\omega_i\omega_j}{Y}\left(Y_i N\left[\frac{\sigma_i^2+\mu_i-\mu_j}{\sqrt{\sigma_i^2+\sigma_j^2}}\right] - Y_i N\left[\frac{\mu_j-\mu_i-\sigma_i^2}{\sqrt{\sigma_i^2+\sigma_j^2}}\right]\right)$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\omega_i\omega_j Y_i}{Y}\left(2N\left[\frac{\sigma_i^2+\mu_i-\mu_j}{\sqrt{\sigma_i^2+\sigma_j^2}}\right] - 1\right)$$

where I arrive at the second line by taking the second term, reversing the order of the double summation, and renaming *j* as *i* and *i* as *j*, while the third line uses the well known fact that $N[-x] = 1-N[x]$. As, from (P3), $ln(Y_i) = \mu_i + .5\sigma_i^2$, simple substitution yields:

$$\text{(A5)}\ G = \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\omega_i\omega_j Y_i}{Y}\left(2N\left[\frac{\ln(Y_i)-\ln(Y_j)+.5\sigma_i^2+.5\sigma_j^2}{\sqrt{\sigma_i^2+\sigma_j^2}}\right] - 1\right)$$

which completes the proof of the theorem.

# References


Aitchison, J. and Brown, J.A.C. (1957). The Lognormal Distribution, with special reference to its uses in economics. Cambridge University Press, Cambridge, 1957.

Bhalla, Surjit S. (2002). Imagine There's No Country: Poverty, Inequality, and Growth in the Era of Globalization. Institute for International Economics, Washington D.C., 2002.

Bourguignon, Francois (1979). "Decomposable Income Inequality Measures." Econometrica, Vol. 47(4 ), 1979: 901-920.

Jones, Charles I. and Klenow, Peter J. (2011). "Beyond GDP? Welfare across Countries and Time." Manuscript, Stanford University, 2011.

Krueger, Dirk and Perri, Fabrizio (2005). "Does Income Inequality Lead to Consumption Inequality? Evidence and Theory." Review of Economic Studies, Vol. 73(1), 2006: 163-193.

Modalsli, Jorgen (2011). "Inequality and growth in the very long run: inferring inequality from data on social groups." Department of Economics, University of Oslo, Memorandum No. 11/2011.

Sala-i-Martin, Xavier (2006). "The World Distribution of Income: Falling Poverty and ... Convergence, Period." Quarterly Journal of Economics CXXI (May 2006): 351-397.

Shorrocks, Anthony F. (1980). "The Class of Additively Decomposable Inequality Measures." Econometrica, Vol. 48(3), 1980: 613-625.