The Advantages of Using Group Means in Estimating the Lorenz Curve and Gini Index From Grouped Data

Merritt Lyon, Li C. Cheung, and Joseph L. Gastwirth

A recent article proposed a histogram-based method for estimating the Lorenz curve and Gini index from grouped data that did not use the group means reported by government agencies. When comparing their method to one based on group means, the authors assume a uniform density in each grouping interval, which leads to an overestimate of the overall average income. After reviewing the additional information in the group means, it will be shown that as the number of groups increases, the bounds on the Gini index obtained from the group means become narrower. This is not necessarily true for the histogram method. Two simple interpolation methods using the group means are described and the accuracy of the estimated Gini index they yield and the histogram-based one are compared to the published Gini index for the 1967–2013 period. The average absolute errors of the estimated Gini index obtained from the two methods using group means are noticeably less than that of the histogram-based method. Supplementary materials for this article are available online.

KEY WORDS: Gini index; Grouped data; Group means; Interpolation; Lorenz curve; Split histogram.

1. INTRODUCTION

The Lorenz curve and Gini index are used in a wide variety of areas, for example, economics (Sen 1973; Nygard and Sandstrom 1981; Atkinson 1983; Kleiber and Kotz 2003; Cowell 2011), genetics (Gianola, Perez-Enciso, and Toro 2003), summarizing insurance scores (Frees, Meyers, and Cummings 2011), equity and efficiency of the allocation process for donor kidneys (Massie et al. 2009), maximizing the benefit of public health (Gail 2009), and statistical analyses (Yitzhaki and Schechtman 2013). They were originally proposed to measure the inequality of income and wealth, and government agencies throughout the world use the Gini index as their main summary measure. Gini indexes of less than 0.3, 0.3–0.399, 0.4–0.499, and 0.5 or greater correspond to low, medium, high, and very high income inequality, respectively (Conference Board of Canada 2011). To preserve the confidentiality of income and other sensitive data, government agencies often publish it in grouped form. The household income data from the Current Population Survey (U.S. Census Bureau 2011a) is a well-known example. A method for deriving upper and lower bounds on the Lorenz curve and Gini index from grouped data with provided group means was proposed by Gastwirth (1972) and refined by Mehran (1975), Gastwirth and Krieger (1975), Giorgi and Pallini (1987), and Silber (1990).

Recently, Tillé and Langel (2012) proposed the use of a histogram-based density to estimate the Lorenz curve and Gini index from grouped data. Their method does not require the group means, and they provided an example showing that their estimate of the Gini index of the 2010 U.S. household income distribution using data from the Current Population Survey organized into four selected groups is 0.4874. This estimate is greater than the lower bound 0.4414 yielded by the method in Gastwirth (1972). They reported that the formula for the upper bound in Gastwirth (1972) gives 0.5565 and noted that the difference in the estimates is far from negligible. However, in their calculation of the Gini bounds, Tillé and Langel (2012) did not use the reported group means, which enter into the formulas given in Section 2. Incorporating the improved bounds in Gastwirth (1972) for densities having a decreasing hazard rate in the tail yields an upper bound of 0.4964, which is noticeably less than 0.5565. Since the U.S. Census Bureau reports the income data in 42 groups along with the group means (U.S. Census Bureau 2011b), it is useful to see how accurately the various estimates of the Gini index approach the reported estimate, 0.469 (DeNavas-Walt, Proctor, and Smith 2011), as the number of groups increases. It will be seen that as the number of groups increases, the distance between the upper and lower bounds from the Gastwirth (1972) method decreases and the bounds always contain the value 0.469. On the other hand, the Gini index obtained from the histogram-based approach does not necessarily become more accurate as the number of groups increases; indeed, it is less than the lower bound obtained using the group means when the full dataset with 42 groups is used.

The concepts and notation are introduced in Section 2. The extra information in the group means is defined, and Tillé and Langel’s histogram-based method is described. In Section 3, a linear interpolation method and the split histogram...
method for estimating the Lorenz curve and Gini index using
the group means are described. They avoid making parametric
assumptions that may be unreliable (Schader and Schmid 1994).
Formulas for the estimated Lorenz curve and Gini index using
the linear interpolation method are derived in Section 4. In
Section 5, these methods are compared to the histogram ap-
proach of Tillé and Langel (2012). The full publicly available
dataset for U.S. household income in 2010 is analyzed and re-
grouped into fewer intervals to illustrate how the bounds become
more accurate as the number of groups increases. Mathemati-
cally, it is demonstrated that when a grouping interval is divided
into two subgroups, the distance between the bounds derived
using the group means can only decrease. Another issue that
arises in comparing the methods is that Tillé and Langel (2012)
arbitrarily set the largest income to $500,000. Assuming different
values for the upper bound, which are consistent with the
largest actual observation (over $1,000,000), severely affects
the Gini index obtained from the histogram approach. In Section 6,
it is seen how the Gini bounds can be used to choose more
informative groupings.

2. BACKGROUND

2.1 Concepts

Given a population of values with distribution function $F$ and
mean $\mu$, the Lorenz curve is defined by (Gastwirth 1971)

$$L(p) = \frac{\int_0^p F^{-1}(t) \, dt}{\mu}, \quad 0 \leq p \leq 1,$$

and represents the share of the total held by the lowest 100$p$% of
the distribution. The Gini index $G$ equals twice the area between
the line of equality $h(p) = p$ and the Lorenz curve $L(p)$:

$$G = 2 \int_0^1 (p - L(p)) \, dp = 1 - 2 \int_0^1 L(p) \, dp.$$

$G$ also is the ratio of the mean difference $\Delta$ to twice the mean:

$$G = \frac{\Delta}{2\mu}.$$

Here, $\Delta$ equals $E[|X_1 - X_2|]$, where $X_1$ and $X_2$ are iid copies
of $F$. An alternative expression for $\Delta$ is (Stuart and Ord 1987,
p. 67)

$$\Delta = 4 \int_{-\infty}^{\infty} x \left( F(x) - \frac{1}{2} \right) \, dF(x)$$

$$= 4 \int_{-\infty}^{\infty} x F(x) \, dF(x) - 2\mu. \quad (3)$$

2.2 Sample Survey Context

Consider a sample of $n$ observations, $x_1, \ldots, x_n$, from a large
finite population, whose values are grouped into $J$ intervals de-
defined by $\{a_{j-1}, a_j\}$, $j = 1, \ldots, J$. For application in income
data, the first group $j = 1$ is bounded below with $a_0 = -\infty$, and
the final group $j = J$ is unbounded with $a_J = \infty$. Let $f_j$ be the estimated proportion contained in group $j$. Then
$\hat{F}_j = \sum_{k=1}^{j} \hat{f}_k$ is the estimated proportion less than $a_j$, where

$$\hat{F}_0 = 0 \text{ and } \hat{F}_J = 1. \text{ Let } \tilde{x}_j \text{ be the mean of group } j. \text{ Then the sample mean is } \bar{x} = \sum_{j=1}^{J} \hat{f}_j \tilde{x}_j. \text{ Finally, let } x_{j}^{\ast} = (a_{j-1} + a_j)/2 \text{ denote
the mid-point of interval } j. \text{ Table 1 contains an example
of grouped income data from the Historical Income Tables of the Current Population Survey (U.S. Census Bureau 2014). From this information, one can estimate the underlying
density of the data and then apply the formulas in Section 2.1 to obtain estimates of the Lorenz curve and Gini index.}

2.3 Estimating the Gini Index when Group Means are Reported

Yntema (1933) showed that the mean difference $\Delta$ can be
expressed as

$$\Delta = \sum_{j=1}^{J} \sum_{j=1}^{J} f_j f_k |\mu_j - \mu_k| + \sum_{j=1}^{J} f_j^2 \Delta_j^*, \quad (4)$$

where $f_j = F(a_j) - F(a_{j-1})$ is the proportion in group $j$, $\mu_j = \int_{a_{j-1}}^{a_j} xdF(x)/f_j$ is the mean of group $j$, and $\Delta_j^*$ is the mean
difference in group $j$. Ignoring the second term in (4), which
assumes all the values in each group are equal, the Gini index is
underestimated by the “grouping correction” (Goldsmith et al.
1954) $D = (2\mu)^{-1} \sum_{j=1}^{J} f_j^2 \Delta_j^*$. Gastwirth (1972) showed that $D$ is bounded above by

$$\hat{D} = 1 \sum_{j=1}^{J} f_j^2 (\mu_j - a_{j-1})(a_j - \mu_j) \quad (5)$$

This yields lower and upper bounds for the Gini index:

$$\text{GL} = \frac{1}{2\mu} \sum_{j=1}^{J} \sum_{j=1}^{J} f_j f_k |\mu_j - \mu_k| \quad \text{ and } \quad \text{GU} = \text{GL} + \hat{D}. \quad (6)$$

If one is willing to assume that the density function in the last
interval is decreasing or has a decreasing hazard rate, Gastwirth
(1972) provided tighter bounds.

When group means are reported, the natural estimate of $\Delta$ is

$$\hat{\Delta} = \sum_{j=1}^{J} \sum_{j=1}^{J} \hat{f}_j \hat{f}_k |\tilde{x}_j - \tilde{x}_k| + \sum_{j=1}^{J} \hat{f}_j^2 \hat{\Delta}_j^*, \quad (7)$$

and the natural estimate of $\hat{D}$ is

$$\hat{\hat{D}} = 1 \sum_{j=1}^{J} f_j \frac{(|\tilde{x}_j - a_{j-1}|)(a_j - \tilde{x}_j)}{(a_j - a_{j-1})}. \quad (8)$$

| j | $a_{j-1}$ | $a_j$ | $\hat{f}_j$ | $\tilde{x}_j$ | $\hat{\Delta}_j^*$ |
|---|---|---|---|---|---|
| 1 | 0 | 20,000 | 0.20 | 0.20 | 10,994 | 10,000 |
| 2 | 20,000 | 38,000 | 0.20 | 0.40 | 28,532 | 29,000 |
| 3 | 38,000 | 61,500 | 0.20 | 0.60 | 49,167 | 49,750 |
| 4 | 61,500 | 100,029 | 0.20 | 0.80 | 78,877 | 80,765 |
| 5 | 100,029 | 180,485 | 0.15 | 0.95 | 130,121 | 140,257 |
| 6 | 180,485 | $\infty$ | 0.05 | 1.00 | 287,201 | – |

NOTE: Source: U.S. Census Bureau, Current Population Survey 2011 Annual Social and
economic Supplement.
One can then estimate the lower and upper Gini bounds by
\[ GL = \frac{1}{2\tilde{\mu}} \sum_{j=1}^{J} \sum_{j=1}^{J} \tilde{f}_j \tilde{f}_k |\tilde{x}_j - \tilde{x}_k| \quad \text{and} \quad GU = GL + D. \] (9)

Cowell and Mehta (1982) and Needleman (1978) showed that the Gini index is accurately estimated by the linear combination,
\[ \frac{1}{2} GL + \frac{2}{3} GU. \]

2.4 The Additional Information in the Group Means

Denote the survival function \( t(x) = 1 - F(x) \). Gastwirth and Krieger (1975) showed that the area under \( t(x) \) over the interval \([a_{j-1}, a_j]\) is given by
\[ \int_{a_{j-1}}^{a_j} t(x)dx \]
\[ = \mu_j f_j + a_j [1 - F(a_j)] - a_{j-1} [1 - F(a_{j-1})]. \] (10)

Suppose one is trying to determine \( F \) based on the values \( F(a_j) \) and \( \mu_j \). When only the information contained in the \( F(a_j) \) is used, any cdf \( H \) satisfying \( H(a_j) = F(a_j), j = 1, \ldots, J \), fits the data. Using the additional information contained in the group means \( \mu_j \), the possible cdfs \( H \) consistent with the data must also satisfy the area constraint (10) in each of the \( J \) intervals, that is,
\[ \int_{a_{j-1}}^{a_j} (1 - H(x))dx \]
\[ = \mu_j f_j + a_j [1 - F(a_j)] - a_{j-1} [1 - F(a_{j-1})]. \]

Krieger (1983) provided several illustrative examples of the gain in information provided by the group means.

2.5 The Histogram-Based Interpolation Method

Tillé and Langel (2012), hereafter T&L, proposed a histogram-based method for estimating the Lorenz Curve and Gini index that does not require the group means. Instead, they chose an arbitrary upper bound of \$500,000 for the income distribution and assumed that the values in each interval follow a uniform distribution. Under this assumption, the mid-point \( x_j^c \) is the mean of group \( j \), and the estimated overall mean income is
\[ \hat{\mu}_u = \sum_{j=1}^{J} \hat{f}_j x_j^c, \] (11)
and the estimated density function is
\[ \hat{h}_u(x) = \begin{cases} \hat{f}_j, & a_{j-1} \leq x < a_j, \quad j = 1, \ldots, J, \\ 0, & x < a_0 \quad \text{or} \quad x \geq a_J. \end{cases} \]
The estimated Lorenz curve obtained from formula (1) is piecewise quadratic. The corresponding estimate of the Gini index is
\[ \hat{G}_u = \frac{1}{2\hat{\mu}_u} \sum_{j=1}^{J} \sum_{j=1}^{J} \hat{f}_j \hat{f}_k |x_j^c - x_k^c| \\
+ \frac{1}{\hat{\mu}_u} \sum_{j=1}^{J} \hat{f}_j (a_j - a_{j-1}) / 6. \] (12)

T&L estimated lower and upper bounds for the Gini index by replacing the mean income and group means in (6) with the estimated overall mean income \( \hat{\mu}_u \) and group mid-points \( x_j^c \), respectively.

The T&L method can be refined by fitting an exponential tail to the last income group using the frequencies in the final two intervals \([a_{j-2}, a_{j-1}) \) and \([a_{j-1}, \infty) \). The exponential density is anchored at the second-to-last cut point
\[ h_{j-1}(x) = \frac{\eta}{\lambda} e^{-(x-a_{j-2})/\lambda}, \quad x \in [a_{j-1}, \infty), \]
where the parameters \( \eta \) and \( \lambda \) are estimated from \( \hat{f}_{j-1} \) and \( \hat{f}_j \).
The estimating equations are
\[ \int_{a_{j-1}}^{\infty} \frac{\eta}{\lambda} e^{-(x-a_{j-2})/\lambda} dx = \hat{f}_j \]
\[ \int_{a_{j-2}}^{\infty} \frac{\eta}{\lambda} e^{-(x-a_{j-2})/\lambda} dx = \hat{f}_{j-1} + \hat{f}_j, \]
whose solution is
\[ \hat{\eta} = \hat{f}_{j-1} + \hat{f}_j \quad \text{and} \quad \hat{\lambda} = \frac{a_{j-1} - a_{j-2}}{\ln(\hat{f}_{j-1} + \hat{f}_j) - \ln(\hat{f}_j)}. \]

3. A LINEARLY INTERPOLATED DENSITY USING THE GROUP MEANS

An interpolation method should account for the finite intervals and unbounded tail. A linear density was assumed for the finite intervals. Because most of the distributions, for example, Pareto (Arnold 2015) and lognormal, used to model the upper tail of the income distribution have a decreasing hazard rate and the exponential distribution provides a bound on the tail probabilities for such distributions (Barlow and Proschan 1965), the exponential distribution will be fit to the unbounded right tail.

3.1 Finite Intervals

For group \( j < J \), a linear density is assumed:
\[ h_j(x) = \alpha_j + \beta_j x, \quad x \in [a_{j-1}, a_j). \] (13)
The estimated density \( \hat{h}_j(x) \) must satisfy three constraints:
\[ \int_{a_{j-1}}^{a_j} \hat{h}_j(x)dx = \hat{f}_j, \quad \int_{a_{j-1}}^{a_j} x \hat{h}_j(x)dx = \hat{f}_j \tilde{x}_j, \quad \text{and} \quad \hat{h}_j(x) \geq 0. \] (14)

3.1.1 Estimation

Solving for \( \alpha_j \) and \( \beta_j \) in (13) and (14) yields
\[ \hat{\beta}_j = \hat{f}_j \frac{12 (\tilde{x}_j - x_j^c)^2}{(a_j - a_{j-1})^3}, \]
\[ \hat{\alpha}_j = \frac{\hat{f}_j}{a_j - a_{j-1}} - \hat{\beta}_j x_j^c = \frac{\hat{f}_j}{a_j - a_{j-1}} \left[ 1 - \frac{12 (\tilde{x}_j - x_j^c)^2}{(a_j - a_{j-1})^2} \right]. \] (15)
The sign of \( \hat{\beta}_j \) indicates whether the group mean is greater than the group mid-point, and the intercept \( \hat{\alpha}_j \) is the usual histogram density \( \hat{f}_j/(a_j - a_{j-1}) \) adjusted by a term that accounts for the
slope. Only if \( \bar{x}_j = x_j^{*} \) does (13) become the uniform density used by T&L.

### 3.1.2 Nonnegativity

Cowell (2011) warned readers that the above estimated density is not guaranteed to be nonnegative. Lemma 1 states that the density will be nonnegative as long as the mean is contained in the middle third of the interval:

\[
\hat{h}_j(x) \geq 0 \iff 2\alpha_{j-1} + 3\alpha_j \leq \bar{x}_j \leq 3\alpha_{j-1} + 2\alpha_j.
\]

**Proof.** See the Appendix.

### 3.1.3 Split Histogram

An alternative method that ensures a nonnegative density is the “split histogram density” method given by Cowell and Mehta (1982) and Cowell (2011). Given \( \bar{x}_j \) and \( \hat{f}_j \), the split histogram density over \([\alpha_{j-1}, \alpha_j] \) is

\[
\hat{h}_j(x) = \begin{cases} 
\frac{\hat{f}_j}{\alpha_j - \alpha_{j-1}} \bar{x}_j - \alpha_{j-1}, & \alpha_{j-1} \leq x < \bar{x}_j \\
\frac{\hat{f}_j}{\alpha_j - \alpha_{j-1}} \bar{x}_j - \alpha_{j-1}, & \bar{x}_j \leq x < \alpha_j
\end{cases}
\]

When the conditions of Lemma 1 are not satisfied, Hoffman (1984) modified the linear interpolator by setting it equal to zero in an appropriate sub-region of an interval.

### 3.2 Unbounded Interval

The final interval \([\alpha_{J-1}, \infty) \) has unbounded support. Cowell and Mehta (1982) chose an upper bound \( B \) for their split histogram density method and explored the sensitivity of their results by replacing \( B \) by \( 3B/4 \) and \( 2B \). An exponential density is assumed

\[
h_j(x) = \frac{\eta}{\lambda} e^{-\frac{(x-a_{j-1})}{\lambda}}, \quad x \in [\alpha_{j-1}, \infty),
\]

where the parameters \( \eta \) and \( \lambda \) are estimated from \( \hat{f}_j \) and \( \bar{x}_j \). The estimating equations are

\[
\int_{\alpha_{j-1}}^{\infty} \frac{\hat{f}_j}{\lambda} e^{-\frac{(x-a_{j-1})}{\lambda}} dx = \hat{f}_j \quad \text{and} \quad \int_{\alpha_{j-1}}^{\infty} \frac{\eta}{\lambda} e^{-\frac{(x-a_{j-1})}{\lambda}} dx = \hat{f}_j \bar{x}_j,
\]

whose solution is

\[
\hat{\eta} = \hat{f}_j \quad \text{and} \quad \hat{\lambda} = \bar{x}_j - \alpha_{j-1}.
\]

### 3.3 Comparison of Densities

Figure 1 plots the linear interpolation, T&L histogram, and split histogram densities for the grouping selected by T&L. For all three methods, an exponential density is fit to the final unbounded interval. The linear interpolation and split histogram densities overlap in this interval. In a typical interval where the density is decreasing, the Lorenz curve of the split histogram method is slightly higher than that of the linear interpolator at the beginning of an interval, then drops below and catches up at the end of the interval.

### 4. ESTIMATES OF THE LORENZ CURVE AND GINI INDEX FOR THE LINEAR INTERPOLATION METHOD

All of the estimated densities in Section 3 yield estimates of the quantile function and cdf. In turn, an estimate of the Lorenz curve is obtained via (1), and an estimate of the Gini index is obtained via (2) and (3).

#### 4.1 Estimate of the Quantile Function

Any value \( p \in [0, 1) \) will belong to one of the \( J \) intervals \([\hat{F}_{j-1}, \hat{F}_j)\) where \( \hat{F}_0 = 0 \) and \( \hat{F}_J = 1 \). The case of a finite interval with a linear density is considered first, followed by the case of the unbounded interval with an exponential tail.

**4.1.1 Finite Interval**

If \( p \in [\hat{F}_{j-1}, \hat{F}_j) \), then the \( p \)th quantile \( x^* = \hat{F}^{-1}(p) \) of the fitted density satisfies

\[
\hat{F}_{j-1} + \int_{\alpha_{j-1}}^{x^*} (\hat{\alpha}_j + \hat{\beta}_j x) dx = p,
\]

where \( \hat{\alpha}_j \) and \( \hat{\beta}_j \) are given in (15). The single positive solution \( x^* \) in \([\alpha_{j-1}, \alpha_j]\) is given by

\[
x^* = \frac{-\hat{\alpha}_j + \sqrt{2\hat{\beta}_j p + \hat{C}_j}}{\hat{\beta}_j},
\]

where \( \hat{C}_j = [\hat{\alpha}_j^2 - 2\hat{\beta}_j \hat{F}_{j-1} + 2\hat{\beta}_j \hat{\alpha}_j \hat{\alpha}_{j-1} + \hat{\beta}_j^2 (\alpha_{j-1})^2] \).
4.1.2 Unbounded Interval

If \( p \in \{ \hat{F}_{j-1}, 1 \} \), then the corresponding quantile \( x^* \) of the fitted exponential satisfies
\[
\hat{F}_{j-1} + \int_{a_{j-1}}^{x^*} \frac{\hat{\lambda} e^{-(x-a_{j-1})/\hat{\lambda}}}{\hat{\lambda}} \, dx = p.
\]
Solving for \( x^* \) gives
\[
x^* = a_{j-1} - \hat{\lambda}\ln \left( 1 - \frac{p - \hat{F}_{j-1}}{\hat{\lambda}} \right).
\]

4.1.3 Estimate of the Lorenz Curve

The estimate of the Lorenz curve is derived from (1) and the results in Section 4.1. See the Web Appendix for the formulas.

4.2 Estimate of the Gini Index

It is convenient to estimate the Gini index using Equations (2) and (3), which requires finding the estimated cdf. Given the linear density (13), the cdf is piecewise quadratic. For \( a_{j-1} \leq x < a_j \), with \( j < J \), the cdf equals
\[
\hat{F}(x) = \hat{F}_{j-1} + \hat{\lambda}(x - a_{j-1}) + \frac{\hat{\beta}_j}{2} \left[ x^2 - (a_{j-1})^2 \right],
\]
and for the final unbounded interval,
\[
\hat{F}(x) = \hat{F}_{j-1} + \hat{\eta} \left( 1 - e^{-\frac{x-a_{j-1}}{\hat{\eta}}} \right).
\]

Using this cdf, one calculates the integrals \( \bar{I}_j = \int_{a_{j-1}}^{x} \hat{F}(x) \, d\hat{F}(x) \). The estimated mean difference is obtained from (3). For details see the Web Appendix. Using (2), the resulting estimate of the Gini index is given by
\[
\hat{G} = \frac{2}{\hat{\lambda}} \left[ \sum_{j=1}^{J-1} \bar{I}_j + \hat{f}_j \bar{x}_j - \frac{\hat{f}_j^2}{4} (\bar{x}_j + a_{j-1}) \right] - 1. \quad (16)
\]

The difference between the estimated Gini bounds (9) on \( \hat{G} \) always decreases when a group is split into two. Formally, as the number \( r \) of groups increases, one has Lemma 2. Define \( \hat{G}_{r} \) and \( \hat{G}_{U} \) to be the lower and upper bounds (9) on the Gini index, respectively, when the number of groups equals \( r \) and \( \hat{G}_{r+1} \) and \( \hat{G}_{U} \) to be the bounds when a group is split.

Lemma 2. \( \hat{G}_{r} \leq \hat{G}_{r+1} \); \( \hat{G}_{L} \geq \hat{G}_{U} \). The inequalities are strict if the distribution has positive density in both of the groups created by the split.

Proof. See Web Appendix. \( \square \)

5. DATA ANALYSIS

Tables 2 and 3 contain estimates of the Gini index (12) derived from the following methods:

1. T&L’s histogram-based method using two choices of upper bound on income: $500,000 and $1,000,000.
2. The modified T&L method using an exponential tail.
3. The linear interpolation method that uses an exponential tail.
4. The split histogram method using two choices of upper bound on income: $500,000 and $1,000,000.
5. The modified split histogram using an exponential tail.
6. $\frac{1}{3}$ the Gini lower bound + $\frac{2}{3}$ the Gini upper bound.

5.1 Comparison of the Estimates to the Reported Gini Index of 0.469

In Table 2, four representative income groupings of the 2010 U.S. Census Bureau data were examined for each method:

1. The four groups selected by T&L with cut points at $50,000, $100,000, and $200,000.
2. The historical six groups given in Table 1.
3. Twelve groups with cut points at $20,000 increments from $20,000 to $200,000 and at $250,000.
4. The full 42 groups used by the Census Bureau with cut points at $5000 increments from $5000 to $200,000 and at $250,000 (U.S. Census Bureau 2011b).

5.2 Comparison of the Estimates to the Reported Gini Index for the Historical Data

The accuracy of the different estimates was calculated on the inflation-adjusted historical data for the 1967–2013 time period. An extract of the findings, reported in the Web Appendix, is given in Table 3, where the average absolute difference (AAD) for all years is also given.

5.3 Discussion

For 2010, the Census Bureau estimates the Gini index as 0.469. Tillé and Langel reported an estimate of 0.4874 using four income groups and an assumed upper income bound of $500,000. This estimate is fairly close to that of the Census Bureau. However, the accuracy of the T&L method depends heavily on the assumed maximum income and the choice of groupings. Using an upper bound of $1,000,000, which is reasonable as the largest income was larger, the T&L estimate of the Gini index is 0.5448. Even using the full 42 groups available for the 2010 data, the T&L Gini estimate can vary consider-
ably with the choice of upper bound. With the largest income assumed to be $500,000, the T&L estimate, 0.4639, of the Gini index is below the Gini lower bound, 0.4683, obtained using the group means. If the largest income is assumed to be $1,000,000, then the T&L estimate, 0.5015, of the Gini index is noticeably larger than the Gini upper bound, 0.4700, obtained using the group means. Although the incorporation of an exponential distribution in place of a uniform distribution in the final group clearly improves the Gini estimate from the T&L approach, in general the exponential tail does not ensure that the resulting estimates of the Gini index are greater than the lower bound based on group means.

For the linear interpolation method, the conditions for Lemma 1 hold for all intervals. The linear interpolation method yields accurate estimates even with only four groups (0.4705 using T&L’s grouping). Table 2 shows that both the linear interpolation and split histogram methods converge to the Census Bureau’s reported estimate of 0.469 very quickly as the number of groups increase. Indeed, they are both quite close to 0.469 when only the six groups given in Table 1 are used. The other methods using the group means yield similar results.

For the historical data, the AAD between each estimate and the published Gini index for the years 1967–2013 is about 0.001 for the methods that use the group means, while that of the original histogram-based method is 0.04. Incorporating an exponential tail into the histogram-based method reduces that to 0.008.

### 6. USING GINI BOUNDS TO DETERMINE A MORE INFORMATIVE GROUPING

#### 6.1 Current Census Grouping

Several authors (Gastwirth 1972; Mehran 1975; Aghevli and Mehran 1981; Davies and Shorrocks 1989) have used bounds on the Gini index to determine a grouping that achieves a desired degree of accuracy when estimating the Gini index from data grouped accordingly. For illustration, consider the current Census grouping, which uses 42 groups with cut points at $5000 increments from $5000 to $200,000 and at $250,000. This is a large number of groups, but the distribution of household income has a long right tail. The Census Bureau reports that the two highest income groups each contain more than 2% of households, while the groups just below them each contain less than 1% (U.S. Census Bureau 2011b). Moreover, the within-group mean differences $\Delta_j^*$ of the highest income groups are large because the group intervals are much wider ($50,000 and open-ended) than the interval length ($5000) used to group incomes in most of the rest of the distribution. Consequently, the contribution of the uncertainty in the bounds of (5) on the overall mean difference and Gini index from the two highest groups will be much greater than that from any of the six groups in the $170,000 to $200,000 range.

#### 6.2 Alternative Grouping

This insight into which groups contribute the most to the difference $\bar{D}$ between the Gini bounds can be used to reduce $\bar{D}$ while keeping the number of groups at 42. It is logical to split the highest income groups with large reported proportions $f_j$ and combine the groups just below them with smaller proportions $f_j$. The following changes to the current Census grouping were considered:

1. Combine the $180,000–185,000$ and $185,000–190,000$ groups
2. Combine the $190,000–195,000$ and $195,000–200,000$ groups
3. Split the $200,000–250,000$ group into $200,000–225,000$ and $225,000–250,000$
4. Split the $250,000+ group into $250,000–400,000$ and $400,000+

Public-use 2012 household income micro-data from the 2013 Current Population Survey are used to compare the groupings. The largest household income in the micro-data is more than $2$ million, four times larger than the upper bound of $500,000 considered by T&L.

For the current and alternative groupings, $\bar{G}_L$, $\bar{D}$, and $\bar{G}_U$ are calculated. The following estimates for the group proportions and means are used:

$$f_j = \frac{\sum_{k \in s_j} w_k}{\sum_{k \in s} w_k} \quad \text{and} \quad \bar{x}_j = \frac{\sum_{k \in s_j} w_k x_k}{\sum_{k \in s} w_k},$$

where $s_j$ is the sample of households belonging to group $j$ and $w_k$ and $y_k$ are, respectively, the weight $HSUP.WGT$ and household income HTOTVAL reported in the micro-data for household $k$. For the current Census grouping, some of these estimates $f_j$ and $\bar{x}_j$ differ slightly from the reported estimates because of possible top-coding and other adjustments to the micro-data.
Table 4 reports the Gini bounds for both groupings. The bounds for both groupings contain the Census Bureau’s 2012 estimate of the Gini index, 0.477 (DeNavas-Walt, Proctor, and Smith 2013). The value $\hat{D}$ for the alternative grouping is one half the value for the current Census grouping, which shows the potential of using the Gini bounds to aid in the choice of intervals for summarizing the income data.

**APPENDIX: PROOF OF LEMA 1**

Proof. $\hat{h}_j(x) \geq 0 \iff \frac{1}{3}a_{j-1} + \frac{1}{3}a_j \leq \hat{x}_j \leq \frac{1}{3}a_{j-1} + \frac{2}{3}a_j$.

Without loss of generality, assume $a_{j-1} = 0$ and $a_j = 1$. An estimated linear density $\hat{h}_j(x) = \hat{a}_j + \hat{\beta}_j x$ fit over the interval $[0, 1]$ with the group mean $\hat{x}_j$ given must satisfy the density constraint

$$\int_0^1 \hat{h}_j(x) dx = \frac{\hat{a}_j + \frac{\hat{\beta}_j}{3}}{2} = 1$$

and the mean constraint

$$\int_0^1 x \hat{h}_j(x) dx = \frac{\hat{a}_j}{2} + \frac{\hat{\beta}_j}{3} = \hat{x}_j.$$ 

Using these constraints, it is seen that $\hat{a}_j = 4 - 6\hat{x}_j$ and $\hat{\beta}_j = 12\hat{x}_j - 6$. For $\hat{h}_j(x)$ to be a valid linear density, both $\hat{h}_j(0) = \hat{a}_j$ and $\hat{h}_j(1) = \hat{a}_j + \hat{\beta}_j$ must be nonnegative, and this occurs if and only if $\frac{1}{3} \leq \hat{x}_j \leq \frac{2}{3}$. $\square$

**SUPPLEMENTARY MATERIALS**

The supplementary appendix contains details on estimating the Lorenz curve and Gini index using the linear interpolation method, the proof of Lemma 2 that showed that the difference between the Gini bounds decreases the number of groups increase, and gives a table reporting the comparison of the estimates of the Gini index obtained by the methods examined in the paper to the reported Gini index for the historical data.

[Received August 2014. Revised September 2015.]
Stuart, A., and Ord, J. K. (1987), *Kendall’s Advanced Theory of Statistics* (5th ed.), London: Griffin and Company. [26]

Tillé, Y., and Langel, M. (2012), “Histogram-Based Interpolation of the Lorenz Curve and Gini Index for Grouped Data,” *The American Statistician*, 66, 225–231. [25,27]

U. S. Census Bureau (2011a), *Current Population Survey, Annual Social and Economic (ASEC) Supplement*, Washington, DC: Author. [25]

——— (2011b), “Table HINC-06. Income Distribution to $250, 000 or More for Households: 2010,” *Current Population Survey, Annual Social and Economic (ASEC) Supplement* [online]. Available at [http://www.census.gov/hhes/www/cpstables/032011/hhinew06_000.htm](http://www.census.gov/hhes/www/cpstables/032011/hhinew06_000.htm). [25,29,30]

Yitzhaki, S., and Schechtman, E. (2013), *The Gini Methodology: A Primer on a Statistical Methodology*, New York: Springer. [25]

Yntema, D. B. (1933), “Measures of the Inequality in the Personal Distribution of Wealth or Income,” *Journal of the American Statistical Association*, 28, 423–433. [26]