A NOTE ON THE MEASUREMENT OF INCOME INEQUALITY WITH INTERVAL DATA*

BY DANIEL A. SEIVER

Miami University, Oxford, Ohio

Concern with income and its (mal) distribution is rising rapidly in the United States and around the world. At the same time, the debate over the measurement of income inequality has intensified.[1] This note shows that the measurement of income inequality, almost exclusively based on grouped data, is sensitive to the number of intervals chosen and the assignment of interval means. These effects could overwhelm cross-section comparisons or time-series results. The effect of grouping error on the Gini index and Theil's index has been analyzed by Gastwirth (1972, 1975), but he assumed that interval means were known, and they often are not. This note presents examples of the magnitudes of grouping errors with unknown interval means for the Gini index and Atkinson's index. These errors in some cases are substantial, in part because of income-reporting biases first noted by Knott.

I. DATA

The data on which this study is based are taken from the 1970 one-percent U.S. Census Public Use Sample (State 5 percent) magnetic tape files.[2] This massive sample of individual records contains information on 1969 wages and salary earnings in hundred-dollar intervals, from \$100–\$199 to \$49,900–\$49,999, with an open interval of \$50,000 or more. The data on wage and salary earnings for males, married with spouse present, was collected in another study,[3] and a 10 percent random sample of this group (25,781) was selected for analysis. Table 1 contains a general description of the sample.[4]

II. RESULTS

Any single-valued index of inequality that is based on interval data will suffer from grouping error. Two popular indexes, the Gini concentration ratio[5] and

*This research was supported in part by the National Science Foundation. I am indebted to the Editors for helpful comments.

[1]See Paglin (1975,1977), Danziger *et al.*, Johnson, Kurien, Minarik, Nelson, on the current debate and Taussig and Lebergott for discussions of the major issues.

[2]Complete documentation of the data and sampling methods for Public Use Samples is contained in Department of Commerce (1972).

[3] Seiver (1978).

[4] Almost all published 1970 Census income data conforms to the interval set used in this paper, although there are some instances of tables with "\$15,000 and over" as the open-ended upper interval.

[5]The calculation of the Gini index is based on the standard formula presented, among other places, in Miller (p. 221).

TABLE 1
DESCRIPTION OF DATA

Number of Observations: 25,781
Mean: $9,131
Median: $8,054
Variance: $390,600
Log Variance: 0.438

Frequency Distribution:

Income Class $	Percent in Class
<1,000	1.27
1,000–1,999	2.49
2,000–2,999	2.45
3,000–3,999	4.15
4,000–4,499	5.67
5,000–5,999	7.64
6,000–6,999	9.56
7,000–7,999	11.35
8,000–9,999	20.98
10,000–14,999	24.40
15,000–24,999	8.09
25,000 and over	1.95
	100 percent

Note: 1970 U.S. Public Use Sample (State 5%) magnetic tape data.

Atkinson's I,[6] will suffice to make this point.

It is standard practice to assign the midpoint of an interval as the mean income for the group in the interval. The midpoints then determine the total income of the population (and thus the shares of each group) and the overall mean, which are necessary to the calculation of inequality indexes. Knott has shown with CPS data that reported earnings (or incomes) tend to "heap" at levels ending in "0", or to a much lesser extent, "5".[7] Thus, the true mean of an interval will almost always be lower than the midpoint, for both small and large intervals, on both sides of the mean, given intervals starting with "0". The asymmetric shape of most income distributions will also influence the differential between midpoint and true mean, as will the relative size and location of the interval.[8] The

[6] Atkinson's index has been gaining rapidly in popularity and is relied on heavily in a recent paper by Williamson. Atkinson's index is calculated according to the formula

$$I = 1 - \left[\sum_i \left(\frac{Y_i}{\mu}\right)^{1-\varepsilon} f(Y_i)\right]^{1/1-\varepsilon}$$

where Y_i is the mean income of interval i, μ the mean income of the entire distribution, $f(Y_i)$ the proportion in interval i, and ε is a measure of the degree of inequality aversion. Williamson uses $\varepsilon = 1.5$ and 2.5, the two values used in this paper. The Gini index is probably closer to an $\varepsilon = 1$ Atkinson index.

[7] The existence of this phenomenon was noted in passing by Budd. It appears also in the Census data used here.

[8] T. P. Schultz refers to the latter of these problems and based on an assumed lognormal distribution uses geometric means of intervals rather than midpoints. This procedure could reduce the error caused by "heaping" somewhat. Spiers summarizes interpolation methods which can reduce grouping error, but his Pareto interpolation method can still result in errors of up to 0.005 in Gini indexes (Spiers, p. 50, Table 4).

TABLE 2
GINI AND ATKINSON INDEXES FOR CENSUS STANDARD AND EQUAL INTERVAL DISTRIBUTIONS, USING MIDPOINTS AND MEANS OF INTERVALS

	Midpoints		Means	
Interval Type / Index	Census Standard	Equal \$2,500	Census Standard	Equal \$2,500
Gini	0.301	0.294	0.293	0.291
Atkinson I ($\varepsilon = 1.5$)	0.261	0.244	0.255	0.236
Atkinson I ($\varepsilon = 2.5$)	0.503	0.431	0.501	0.413

effect of this error on the Gini coefficient and Atkinson's I will be considered concurrently with the effect of alternative sets of intervals.

Table 2 shows the value of the Gini coefficient and the value of Atkinson's I (for $\varepsilon = 1.5$ and 2.5) using the standard Census intervals and equal \$2,500 intervals. It is immediately clear from inspection of this table that: 1) the use of Census standard intervals results in more inequality in all cases; 2) the use of midpoints of intervals results in more inequality in all cases; 3) the Gini index is much less sensitive to these changes than Atkinson's I ($\varepsilon = 1.5$) which, in turn, is less sensitive that Atkinson's I ($\varepsilon = 2.5$).[9] This comparison suggests that "interaction" effects are present: in fact, the sensitivity of Atkinson's I ($\varepsilon = 2.5$) to switches from midpoints to means depends on the interval set used, while the sensitivity of Atkinson's I ($\varepsilon = 1.5$) to this switch is much less, and the Gini sensitivity to the same switch smaller still. Given the similarity between Atkinson's and Theil's indexes of inequality, these findings are consistent with Gastwirth's (1975) findings that Theil's index is sensitive to the assumptions that all incomes in an interval are at the midpoint.

Although there are no formal tests for statistical significance for Atkinson's I, we can compare the variations in Table 2 with cross-section results reported in Atkinson, in order to gauge substantive significance. Atkinson ranks the income distributions of 12 countries according to his index with $\varepsilon = 1.5$ (p. 259). Eight of the eleven distances between countries are equal to or smaller than the 0.017 to 0.019 variation caused by interval shifts, as reported in Table 2.

The variations in the Gini index are smaller, but the Gini is a much less sensitive index. Although a formal test of significance was inconclusive,[10] Atkinson, for example, also ranked the same twelve countries' income distributions by the Gini index (p. 259), and six of the differences between countries are less than the 0.008 variation of Table 2 (Census Standard, means vs. midpoints).

[9]Gastwirth (1972, p. 310–11) discusses the effects of interval selection on the Gini, but with known interval means. Gastwirth and Glauberman (1976) also mention in passing that interval selection can affect the Gini, but do not develop the point.

[10]Reynolds and Smolensky (pp. 72–4) use a test devised by Kakwani and Podder in which a smooth Lorenz curve estimated by OLS is fitted to the data, which gives a slope coefficient and a standard error for significance tests. While this test suggests the Gini coefficients of Table 2 are not significantly different, it must be noted that: 1) the Lorenz curves intersect, 2) the unexplained variation in my data is 10 times as large as reported by Reynolds and Smolensky (p. 73, Table 5.3) and 3) the constant term is larger than its standard error, even though the linearized function should go through the origin, suggesting the functional form is inappropriate for my data.

The overall thrust of the tabular results is clear: popular single-index measures of income inequality are sensitive to the interval set used, and also to the assignment of means to the intervals. Since midpoints are almost universally used as measures of closed-interval means, the extent of income inequality tends to be overstated. This overstatement is offset by the use of a greater number of intervals. Thus, comparative measures of inequality for distributions with different interval sets, or comparisons with the same set of intervals with different means (or different degrees of income "heaping") could produce misleading results. The validity of these conclusions can be tested through estimation of other measures of inequality, further experimentation with alternative interval sets,[11] and use of alternative data sets.

REFERENCES

Atkinson, A. B., "On the Measurement of Inequality," *J. Econ. Theory*, Sept. 1970, 2, 244–63.

Blaug, M., "The Empirical Status of Human Capital Theory: A Slightly Jaundiced Survey," *J. Econ. Literature*, Sept. 1976, 14, 827–55.

Budd, E. C., "Postwar Changes in the Size Distribution of Income in the U.S.," *Amer. Econ. Review*, May 1970, 60, 247–60.

Danziger, S. *et al.*, "The Measurement and Trend of Inequality: Comment," *Amer. Econ. Review*, June 1977, 67, 505–12.

Gastwirth, J. L., "The Estimation of a Family of Measures of Economic Inequality," *Journal of Econometrics* 1975, 3, 61–70.

———, "The Estimation of the Lorenz Curve and Gini Index," *Rev. Econ. Stat.*, August 1972, 54, 306–16.

——— and M. Glauberman, "The Interpretation of the Lorenz Curve and Gini Index from Grouped Data," *Econometrica*, May 1976, 44, 479–83.

Johnson, W. R., "The Measurement and Trend of Inequality: Comment," *Amer. Econ. Review*, June 1977, 67, 502–04.

Kakwani, N., "On the Estimation of Income Inequality Measures from Grouped Observations," *Rev. Econ. Studies*, October 1976, 43, 483–92.

Knott, J. J., "Analysis of the Effect of Income Rounding in the Current Population Survey," *1971 American Statistical Association Proceedings* (Social Statistics Section), Washington, 1972.

Kurien, C. J., "The Measurement and Trend of Inequality: Comment," *Amer. Econ. Review*, June 1977, 67, 517–19.

Lebergott, S., "Dimensions of Income Inequality," unpublished paper, 1977.

Metcalf, C. E., *An Econometric Model of the Income Distribution*, Chicago, 1972.

Miller, H. P., *Income Distribution in the United States*, Washington, 1966.

Minarik, J. J., "The Measurement and Trend of Inequality: Comment," *Amer. Econ. Review*, June 1977, 67, 513–16.

Nelson, E. R., "The Measurement and Trend of Inequality: Comment," *Amer. Econ. Review*, June 1977, 67, 497–501.

Paglin, M., "The Measurement and Trend of Inequality: A Basic Revision," *Amer. Econ. Review*, Sept. 1975, 65, 598–609.

———, "The Measurement and Trend of Inequality: Reply," *Amer. Econ. Review*, June 1977, 67, 520–31.

Reynolds, M., and E. Smolensky. *Public Expenditures, Taxes and the Distribution of Income*, New York, 1977.

Schultz, T. P., "Secular Trends and Cyclical Behavior of Income Distribution in the United States, 1947–1965," in Lee Soltow, ed., *Six Papers on the Size Distribution of Wealth and Income*, New York, 1969.

Seiver, D. A., "A Note on Income Heaping in U.S. Census Data," unpublished paper, 1977.

———, "Which Couples at Given Parities Have Additional Births?" *Research in Population Economics*, October, 1978.

[11]The suggestion that interval sets with midpoints at "0" be used is discussed in Seiver (1977). This is a rather naive suggestion, given institutional and comparability constraints, and the "digit preference" of data-disseminating agencies.

Spiers, E., "Estimation of Summary Measures of Income Size Distribution from Grouped Data," *1977 American Statistical Association Proceedings* (Social Statistics Section), Washington, 1978.
Taussig, M. K., "Trends in Inequality of Well-Offness in the United States Since World War II," in *Conference on the Trend in Income Inequality in the U.S.*, Madison, 1976.
U.S. Bureau of the Census, *Public Use Samples of Basic Records from the 1970 Census*, Description and Technical Documentation, Washington, 1972.
Williamson, J. G., "'Strategic' Wage Goods, Prices, and Inequality," *Amer. Econ. Review*. March 1977, 67, 29–41.