THE INDEX OF INCOME CONCENTRATION IN THE 1970 CENSUS OF POPULATION AND HOUSING

Joseph J. Knott, Bureau of the Census*

Introduction

Publications showing results of the 1970 Census of Population will contain the Index of Income Concentration (also known as the Gini Index of Inequality) for families, unrelated individuals and for persons. They will be available for areas or cities with population over 50,000, counties, States, and for the United States. The primary purpose of this paper is to outline the procedure used to compute the Index so that the procedure may be duplicated by interested users. Also presented are results of the research undertaken to determine the effect of the various assumptions used in the estimation technique.

Section I outlines the procedure used to compute the Index of Income Concentration (or Index). Section II analyzes some of the effects of various assumptions and constraints used in developing the Index. It is divided into six parts: (A) The overall effect on the Index from using estimated means, (B) use of the midpoint of an income interval as the estimated mean of the income interval, (C) use of the Pareto formula to estimate the mean of the open-end interval, (D) assumption involved in splitting larger $2,000, $3,000, and $5,000 income intervals into $1,000 income intervals, (E) choice of the size of the open-end income interval, and (F) the range of acceptable Indexes. Section III summarizes key findings.

Procedure for Computing the Index of Income Concentration

The Index is defined in terms of the Lorenz curve, and may be represented as the ratio of the area between the diagonal and the Lorenz curve to the area under the diagonal. 1/ The computation of the Index uses an approximate integration technique and requires the percent distribution of units and the percent distribution of aggregate income both by income classes.

The 1970 Census publications show selectively income size distribution of the number of families, unrelated individuals, and persons. A percent distribution is obtained from a numerical distribution by dividing the units in each income class by the total number of units covered in the distribution. It is the computation of the percent distribution of aggregate income which usually presents problems in computing the Index. The Census publications do not show aggregate income by each income class and consequently the aggregate income for each income class must be estimated by multiplying the number of units by the assumed mean for each income class.

In general, in the computation of the Index, the midpoint of an income class is assumed to be the mean of the income interval. This is true for income intervals ranging between $1,000 to $15,000. For "less than $1,000," $500 is assumed to be the mean. For the $15,000 to $19,999 and $20,000 to $24,999 intervals, $17,000 and $22,000, respectively, are assumed to be the means. The Pareto formula is usually used to estimate the mean of the open-end interval.

In order to lessen the error associated with the linearity assumption applied in the approximate integration technique, larger income intervals are divided into smaller income intervals by relating the logarithm of units by the logarithm of income within the class interval. For example, the family income distributions contained in the Census detailed publications show the income interval $12,000 to $14,999. This composite interval is subdivided into three $1,000 intervals. (See table 1.)

Table 1.--INCOME SIZE DISTRIBUTION RELATIONSHIPS FOR SPLITTING THE $12,000-$15,000 INCOME INTERVAL INTO THREE $1,000 INTERVALS

Ratio of frequency of above $12,000 interval to frequency of above $15,000 interval	Percent of $12,000 to $15,000 interval	Percent of $12,000 to $13,000 interval	Percent of $13,000 to $14,000 interval	Percent of $14,000 to $15,000 interval
Under 1.5.....	100	40	33	27
1.5 to 2.5.....	100	44	32	24
2.5 to 3.5.....	100	49	31	20
3.5 and over.....	100	53	28	19

The above table is used as follows:

1. Compute the number of units with income over $15,000 (or F_{15+}). For example, F_{15+} = 349 units.

2. Compute the number of units with income over $12,000 (or F_{12+}). For example, F_{12+} = 425 units.

3. Compute the ratio $\frac{F_{12+}}{F_{15+}}$ or $\frac{425}{349} = 1.218$

4. Find the proper line in the above table for 1.218 (or line 1 above) and apply the percentages to the number of units in the $12,000 to $14,999 interval to get the frequency within the three $1,000 income intervals.

* Comments by Dr. Murray S. Weitzman, Assistant Division Chief for the Economic Statistics Programs, and staff members of the Consumer Income Statistics Branch, Population Division are gratefully acknowledged.

There are two open-end intervals ($15,000 and over; and $25,000 and over) used in the calculation of the Index. In most cases, the mean computed by using the Pareto Formula (the Pareto estimate) of the open-end is used. The Pareto estimate of the $25,000 and over open-end income interval is computed.

First derive the slope in the formula:

$$\text{Slope} = \frac{\log_{10}\left[\frac{F_{25+} + F_{15-25}}{F_{25+}}\right]}{\log_{10}\left[\frac{25,000}{15,000}\right]} = \frac{\log_{10}\frac{F_{15+}}{F_{25+}}}{.22185}$$

Where F_{25+} = Number of units with income over $25,000

F_{15+} = Number of units with income over $15,000

F_{15-25} = Number of units with income in the range $15,000 - $24,999

From the above, the Pareto estimate (of the $25,000 and over interval) is derived:

$$\frac{\text{Slope}}{\text{Slope (minus) } 1.0} \times \$25,000 = \text{Pareto estimate}$$

If the frequency in the $15,000 to $24,000 interval is zero, the Pareto estimate cannot be calculated and $36,000 is used as the estimated mean of the open-end interval. Also, if the Pareto estimate is outside the range of $25,000 to $75,000, it is not used and $36,000 is used as the mean of the open-end. 2/ This range constraint is seldom used, and is usually associated with a distribution having a very small base.

The Pareto estimate of the $15,000 and over income interval is computed similarly except that the acceptable range is $15,000 to $40,000. If the Pareto estimate falls outside of this range then the estimate of $23,000 is used.3/

When the percent distribution of units (P_i) and the accumulated percent distribution of aggregate income (A_i) are obtained on the expanded interval distribution (by the above method), the Index is then computed as follows:

$$\text{Index} = 1 - \sum_{i=1}^{n} \left(P_i\right) \times \left(A_{i-1} + A_i\right)$$

P_i = Percent of units in the ith income interval

A_i = Cumulative percent of aggregate income in the ith income interval (when i = 0, A_i = 0)

n = Number of income classes

Assumptions Used in Computing the Index

A. Overall Effect on the Index in Using Assumed Interval Means versus Tabulated Means

The problem is to determine the effect on the Index of using assumed interval means (midpoints) rather than tabulated interval means. The findings show that with relative few income intervals, the use of midpoints as interval means tends to result in estimates about as good as estimates of the Index using tabulated means.

To investigate this problem the Index was computed on a distribution with 190 income intervals using tabulated interval mean values. This is the "Perfect" Index in the sense the "bias" introduced by using the approximate integrated technique is greatly reduced. The smaller (19) interval distributions used to calculate the Index are simply collapses of the 190 interval distribution data. It should be noted that by definition, the number of intervals has an effect on the value of the Index in that a reduction in the number of intervals tends to bias the Index downward. (See table 2).

Table 2.--INDEX OF INCOME CONCENTRATION FOR FAMILIES AND UNRELATED INDIVIDUALS BY AGE BY THREE COMPUTATION METHODS IN 1969

AGE	"PERFECT" Index (190 intervals)	Tabulated Means (19 intervals)	Census Estimation Procedure 1/
Families.....Total	.349	.346	.346
14 - 24	.300	.298	.296
25 - 34	.274	.272	.270
35 - 44	.301	.298	.296
45 - 54	.323	.318	.323
55 - 64	.367	.363	.367
65 and over	.434	.432	.439
Unrelated Individuals.....Total	.480	.475	.469
14 - 24	.454	.447	.426
25 - 34	.370	.368	.343
35 - 44	.404	.401	.406
45 - 54	.428	.425	.429
55 - 64	.438	.434	.432
65 and over	.471	.458	.469

1/ The Census estimation procedure as detailed in the first part of this paper uses 14 tabulated income intervals expanded to 19 with assumed means used to compute the percent aggregate income distribution.

Source: Bureau of the Census, Current Population Survey

As compared with the "Perfect" Index, the Census estimation procedure based on assumed means approximates it fairly well. The slight overestimate of the interval means compensates for the underestimate of the Index caused by the reduction in the number of income intervals.

B. Midpoints as Means of Income Classes

The problem here is to test whether or not midpoints represent good estimates of the actual interval means. For income intervals between $1,000 and $15,000, the midpoint of the interval was used as the mean of the interval. For the under $1,000 interval, $500 was used and for the $15,000 to $19,999 and $20,000 to $24,999 intervals, $17,000 and $22,000, respectively, were used as the means. The use of the midpoint as mean of an income interval is supported by an Internal Revenue Service (IRS) tabulation of adjusted gross income (AGI) by AGI class. The mean AGI of the intervals from $1,000 to $10,000, all fell within $18 of the midpoint. (See table 3) The mean of the "under $1,000" class is not relevent because persons with AGI under $600 are not required to file a tax return.

As data in table 3 show, the CPS tabulated mean within each interval between $2,000 and $15,000 consistently falls below their midpoint in each income interval. This is contrary to what would be expected of a right skewed income frequency distribution. As the units increase in frequency from one interval to another it would seem logical the same increasing frequency would be found within the interval. However, this is not the case. A tabulation by $100 and $250 intervals clearly shows that there is a high frequency in the $100 or $250 interval which contains the even $1,000 amount. Attachment 1 is a bar graph showing the number of families tabulated by small income intervals. The high frequency in the intervals containing the even $1,000 amount is quite evident. This tendency is shown in total family income which is the sum of eight separate income questions per family member and more than one person. This apparent reporting bias is being studied further.

Table 3.--MEAN AGI AND TOTAL MONEY INCOME IN 1969 BY SIZE CLASS

Size Class	Mean Adjusted Gross Income 1/	Mean Total Family Income 2/
Total	$7,959	$10,577
Under $1,000	946 3/	51
$1,000 to $1,999	1,491	1,543
$2,000 to $2,999	2,493	2,475
$3,000 to $3,999	3,488	3,486
$4,000 to $4,999	4,502	4,475
$5,000 to $5,999	5,495	4,457
$6,000 to $6,999	6,497	6,436
$7,000 to $7,999	7,495	7,453
$8,000 to $8,999	8,490	8,443
$9,000 to $9,999	9,495	9,447
$10,000 to $11,999	12,134	10,876
$12,000 to $14,999		13,280
$15,000 to $19,999	17,013	18,284
$20,000 to $24,999	22,093	
$25,000 and over	46,132	35,786

1/ Preliminary Statistics of Income, 1969, "Individual Income Tax Return," Internal Revenue Service, Table 4, page 22.

2/ U.S. Bureau of the Census, Current Population Reports, Series P-60, No. 75, "Income in 1969 of Families and Persons in the United States," Table 1, page 19.

3/ Not comparable since persons with Adjusted Gross Income below $600 are not required to file a tax return.

C. Use of Pareto Formula to Compute the Mean of the Open-End Income Interval

This analysis shows that the use of the Pareto Formula tends to overestimate the mean of the open-end if compared with the tabulated mean of the open-end income interval.

Table 4 shows the Pareto estimate of the mean of the open-end interval and the actual tabulated value from the March 1970 CPS. The Pareto estimate of open-end income interval of $25,000 and over is clearly better for families, than it is for unrelated individuals. The difference between the Pareto estimate and the tabulated means indicates that the Pareto estimate should be used carefully. Unfortunately the tabulation of means by income interval is expensive in terms of computer core space and if tabulated means are not available, the use of the Pareto estimate is the most feasible alternative for estimating the mean of the open-end. It should also be noted that the tabulated means from the CPS are slight underestimates of the Census means since CPS income data by type cannot be coded above $99,900, while the Census items can be coded to $990,000.

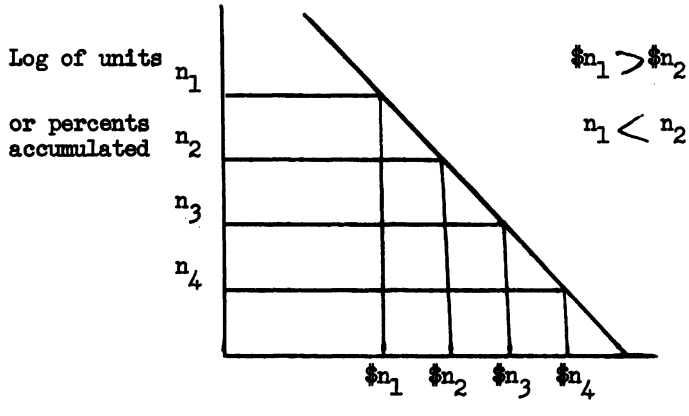
Table 4.--Pareto Estimates and Tabulated Mean Values of the $25,000 and Over and $15,000 Open-End Income Intervals for Families and Unrelated Individuals by Selected Characteristics for 1969

Selected Characteristics	$25,000 and over			$15,000 and over		
	Pareto	Tabulated	Percent Difference	Pareto	Tabulated	Percent Difference
All families	$35,975	$35,786	+0.5	$25,650	$21,625	+18.6
All unrelated individuals	39,500	38,480	+2.7	21,750	22,791	- 4.6
Negro and other races						
Families	33,000	31,117	+6.1	23,100	19,681	+17.4
Unrelated individuals	34,950	30,342	+15.2	19,800	16,717	+18.4

Source: Bureau of the Census--Estimates derived from data in the Current Population Survey.

D. Splitting Income Intervals

The assumption of a log-log relationship on which the broad intervals are split is a good assumption to use for the above $10,000 interval on almost all distributions. This is clearly shown by graphing distributions on log-log paper and observing the linear relationship. From about $6,000 or $7,000 to $10,000 the graph curve shows a shift from log-log to more of a log-normal relationship. The log-normal relationship is also clearly shown on log-normal graph paper. The tables for splitting six different income intervals are given in Attachment 2. These tables are constructed from the following formula.



$$\$n_1 > \$n_2 > \$n_3 > \$n_4$$

$$n_1 < n_2 < n_3 < n_4$$

$$\frac{\log n_1 - \log n_2}{\log \$n_2 - \log \$n_1} = \frac{\log n_1 - \log n_4}{\log \$n_4 - \log \$n_1}$$

$$\log n_2 = \log n_1 - \frac{(\log n_1 - \log n_4)\ (\log \$n_2 - \log \$n_1)}{(\log \$n_4 - \log \$n_1)}$$

Percent or number of units with income over n_2 = Antilog ($\log n_2$)

The tables were constructed by computing the values of the n_2 (all intermediate points desired) for various values of the ratio $\frac{n_1}{n_4}$ of curve (under 1.5, 1.5 to 2.5, 2.5 to 3.5, and 3.5 and over). The percent proportions of n_1 to n_2, n_2 to n_3, n_3 to n_4 to the n_1 to n_4 class were then computed for the midpoint of the 1.5 to 2.5 and 2.5 to 3.5 ranges; and for the under 1.5 and 3.5 and over income interval, 1.5 and 3.5 were used.

E. Choice of the Size of the Open-End Income Interval

For the computation of the Index for family income distributions, the $25,000 and over open-end income interval is used, and for unrelated individuals and persons, $15,000 is used in the 1970 Census. The choice of the open-end is important because it determines the relative importance of the Pareto estimate. Different open-end intervals were used for families and unrelated individuals because they make the Index more comparable in terms of the percent of units in the open-end interval. This gives more equal weight to the Pareto estimate.

Table 5.--ACCUMULATED PERCENT OF UNITS FOR FAMILIES AND UNRELATED INDIVIDUALS FOR SELECTED INCOME CLASSES

Total money income	Percent of units over the specified income level	
	Families	Unrelated Individuals
Over $12,000	32.9	4.9
Over $15,000	19.2	2.4
Over $25,000	3.6	0.6

Source: Bureau of the Census, Current Population Reports, Series P-60, No. 75, Table 16.

As the table shows, 3.6 percent of all families had incomes above $25,000, but only 0.6 percent of unrelated individuals was in the same interval. This difference would result in the Pareto mean having six times the weight for family distributions relative to unrelated individual distributions. This disparity is reduced by using the $15,000 and over interval as the open-end interval for unrelated individuals, and the $25,000 and over interval for families (i.e., 2.4 percent for unrelated individuals relative to 3.6 percent for families).

F. Range of Published Indexes

For publication purposes, only Indexes within the range of .200 to .650 will be published. An Index outside this range will be suppressed and three dots will be shown (...). Indexes outside this range, for the most part, represent Indexes computed on very small bases. In any case, users can compute Indexes, if desired, for these distributions by using the technique outlined in this paper.

Summary

In summary, the estimation technique used to compute the Index of Income Concentration from the Census publications appears to give good results in most cases. It is interesting to note that (when compared to an Index computed on the basis of 190 intervals), the estimation procedure results in estimates about as good as estimates of the Index produced by using tabulated number and aggregate income for 19 size income intervals.

The tendency for respondents to report estimated income to the nearest $1,000 is an interesting phenomenon which is being analyzed further.

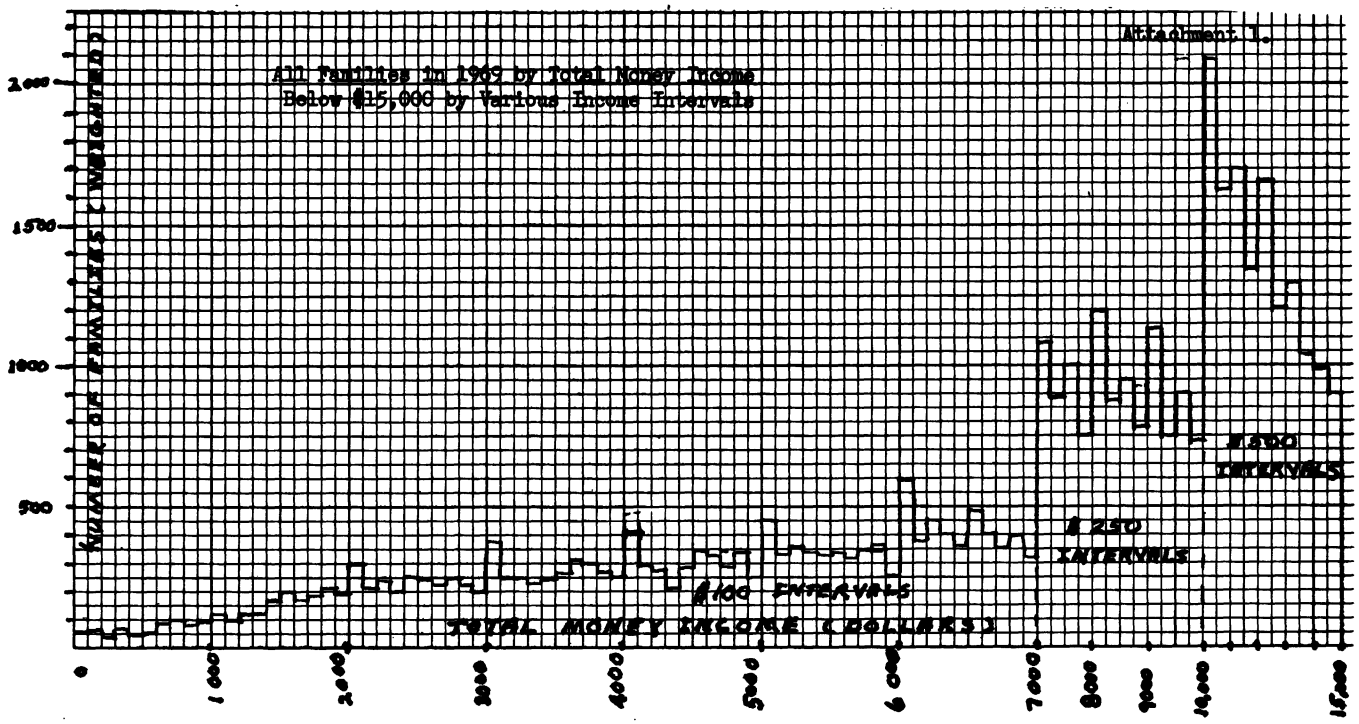
Findings showed that the various assumptions used to compute the Index do not invalidate the relative accuracy of the Index. The assumption of the midpoint as the mean of the income interval is supported by AGI data, but CPS income data suggest that midpoints are too high. The use of the Pareto formula also tends to overestimate the mean of the open-end interval, but not uniformly. Furthermore, data show that the number of intervals used to compute the Index makes a difference. Any comparison of Indexes requires that they be computed using the same number of income intervals.

FOOTNOTES

1/ An expanded discussion of the geometric interpretation of the Index of Income Concentration may be found in: Rich Man, Poor Man, by Herman P. Miller, Thomas Y. Cromwell Co., New York, 1971 appendix B, pp. 274 - 279.

2/ Implicit in this constraint is a ratio of $F_{25+}/F_{15+} = 2.15$. The value of $36,000 is obtained from CPS income data.

3/ Implicit in this constraint is a ratio of $F_{15+}/F_{12+} = 1.60$. The value of $23,000 is obtained from CPS income data.



Attachment 2.

Proportions to be Used to Split Broad Income Intervals into Smaller Income Intervals

A. $6,000 - $7,999

F6+ / F8+	F6-8	F6-7	F7-8
Under 1.5	100	56	44
1.5 - 2.5	100	62	38
2.5 - 3.5	100	66	34
3.5 and over	100	68	32

B. $8,000 - $9,999

F8+ / F10+	F8-10	F8-9	F9-10
Under 1.5	100	56	44
1.5 - 2.5	100	61	39
2.5 - 3.5	100	66	34
3.5 and over	100	68	32

C. $10,000 - $14,999

F10+ / F15+	F10-15	F10-11	F11-12	F12-13	F13-14	F14-15
Under 1.5	100	27	23	19	17	14
1.5 - 2.5	100	30	24	19	15	12
2.5 - 3.5	100	34	25	18	13	10
3.5 and over	100	36	25	17	13	9

D. $10,000 - $11,999

F10+ / F12+	F10-12	F10-11	F11-12
Under 1.5	100	55	45
1.5 - 2.5	100	61	39
2.5 - 3.5	100	66	34
3.5 and over	100	67	33

E. $12,000 - $14,999

F12+ / F15+	F12-15	F12-13	F13-14	F14-15
Under 1.5	100	40	33	27
1.5 - 2.5	100	44	32	24
2.5 - 3.5	100	49	31	20
3.5 and over	100	53	28	19

F. $15,000 - $24,999

F15+ / F25+	F15-25	F15-20	F20-25
Under 1.5	100	60	40
1.5 - 2.5	100	65	35
2.5 - 3.5	100	69	31
3.5 and over	100	71	29

F_{m+} - Number of units with income over $m,000

F_{m-n} - Number of units with income from $m,000 to $n,000