Volume Title: Changes in Income Distribution During the Great Depression

Volume Author/Editor: Horst Mendershausen

Volume Publisher: NBER

Volume ISBN: 0-870-14162-7

Volume URL: http://www.nber.org/books/mend46-1

Publication Date: 1946

Chapter Title: Appendix C: Measures of Income Inequality

Chapter Author: Horst Mendershausen

Chapter URL: http://www.nber.org/chapters/c5313

Chapter pages in book: (p. 159 - 167)

APPENDIX C

Measures of Income Inequality

1 *Standard Deviation and Coefficient of Variation*

In a given year there are N family incomes $x_1, x_2, \ldots x_N$ with a mean income

(1;1)$\bar{x} = (\sum_1^N x) / N$, a standard deviation

(1;2) $\sigma = \frac{\sum_1^N (x - \bar{x})^2}{N - 1}$, and a coefficient of variation

(1;3) $v = \frac{\sigma}{x}$

The total variance in the distribution can be attributed to any number of components, e.g., two. Let us order family incomes by size and put the lower incomes $(x_1, \ldots, x_k)$ in the lower income group, the higher incomes $(x_{k+1}, x_{k+2}, \ldots, x_N)$ in the upper income group, and call the number in the lower group N_l, the number in the upper group N_u, so that we have $\bar{x}_l = (\sum_1^k x)/N_l$, $\bar{x}_u = (\sum_{k+1}^N x)/N_u$, and $N_l + N_u = N$, where $N_l = k$.

It can be shown that

(1;4) $\sum_1^N (x - \bar{x})^2 = \sum_1^k (x - \bar{x}_l)^2 + \sum_{k+1}^N (x - \bar{x}_u)^2 + N_u (\bar{x}_u - \bar{x})^2 + N_l (\bar{x} - \bar{x}_l)^2$.

We know that

(1;5) $\sum_1^N (x - \bar{x})^2 = \sum_1^N x^2 - N\bar{x}^2$. Correspondingly

(1;6) $\sum_1^k (x - \bar{x}_l)^2 = \sum_1^k x^2 - N_l \bar{x}_l^2$, and

(1;7) $\sum_{k+1}^N (x - \bar{x}_u)^2 = \sum_{k+1}^N x^2 - N_u \bar{x}_u^2$, and

(1;8) $N_u (\bar{x}_u - \bar{x})^2 + N_l (\bar{x} - \bar{x}_l)^2 = N_u \bar{x}_u^2 + N_l \bar{x}_l^2 - N\bar{x}^2$.

Adding equations (1;6), (1;7), (1;8) and making use of equation (1;5) we obtain (1;4). This can be simplified further, for it is easy to see that

(1;9) $N_u (\bar{x}_u - \bar{x})^2 = \frac{N_l^2 N_u}{N^2} (\bar{x}_u - \bar{x}_l)^2$ and

(1;10) $N_l (\bar{x} - \bar{x}_l)^2 = \frac{N_u^2 N_l}{N^2} (\bar{x}_u - \bar{x}_l)^2$ so that

(1;11) $N_u (\bar{x}_u - \bar{x})^2 + N_l (\bar{x} - \bar{x}_l)^2 = \frac{N_u N_l}{N} (\bar{x}_u - \bar{x}_l)^2$. Therefore (1;4) becomes

(1;12) $\sum_1^N (x - \bar{x})^2 = \sum_1^k (x - \bar{x}_l)^2 + \sum_{k+1}^N (x - \bar{x}_u)^2 + \frac{N_u N_l}{N} (\bar{x}_u - \bar{x}_l)^2$.

Thus it is proved that the total variance can be conceived of as the sum of (a) the total variance within the lower group, (b) the total variance within the upper group, and (c) the weighted square difference between the mean incomes of the two groups. Defining

$$\sigma_l = \sqrt{\frac{\sum_{1}^{k} (x - \bar{x}_l)^2}{N_l - 1}} \text{ and } \sigma_u = \sqrt{\frac{\sum_{k+1}^{N} (x - \bar{x}_u)^2}{N_u - 1}} \text{ , we obtain}$$

$$(1;13)\ \sigma = \sqrt{\frac{(N_l - 1)\ \sigma_l^2 + (N_u - 1)\ \sigma_u^2 + \frac{N_u N_l}{N} (\bar{x}_u - \bar{x}_l)^2}{N - 1}} \text{ and}$$

$$(1;14)\ v = \frac{1}{\bar{x}} \sqrt{\frac{(N_l - 1)\ \bar{x}_l^2\ v_l^2 + (N_u - 1)\ \bar{x}_u^2\ v_u^2 + \frac{N_u N_l \bar{x}^2}{N} \left(\frac{\bar{x}_u - \bar{x}_l}{\bar{x}}\right)^2}{N - 1}}$$

Formula (1;13) shows the general standard deviation as a function of the separate standard deviations of the two groups and the difference between the group means. Formula (1;14) expresses the general coefficient of variation in terms of relative income dispersion within and between groups.

All the preceding demonstrations hold also for grouped data where the σ and v are replaced by their approximations σ' and v'. There are s income classes with mean incomes $\bar{x}_i$ ($i = 1, 2, \ldots, s$) and absolute frequencies f_i,

$$(1;15)\ \sum_{i=1}^{s} f_i = N, \qquad (1;16)\ \bar{x}_i = (\Sigma_i\ x)/f_i$$

$$(1;17)\ \sigma' = \sqrt{\frac{\sum_{1}^{s} f_i\ (\bar{x}_i - \bar{x})^2}{N - 1}} \qquad (1;18)\ v' = \frac{\sigma'}{\bar{x}}\ .$$

Throughout this monograph the $\bar{x}_i$'s are used instead of the unknown individual incomes x. Since intraclass variation is neglected, all our estimates of dispersion, absolute or relative, have some downward bias.

The squared coefficient of variation for the entire distribution computed from grouped data (v') is the sum of the following three components:

(1;19) weighted inequality within the lower income group:

$$J_l = \frac{(N_l - 1)\ \bar{x}_l^2\, v'^2_l}{(N - 1)\ \bar{x}^2}$$

(1;20) weighted inequality within the upper income group:

$$J_u = \frac{(N_u - 1)\ \bar{x}_u^2 v'^2_u}{(N - 1)\ \bar{x}^2}$$

(1;21) weighted inequality between the lower and upper income groups:

$$J_{lu} = \frac{N_l\,N_u}{N\,(N-1)}\quad\frac{(\bar{x}_u - x_l)^2}{\bar{x}}$$

2 *Mean Difference, Coefficient of Concentration, and the Lorenz Chart*

N family incomes are ordered by size from low (x_1) to high (x_N). The *mean interindividual difference,* a measure developed by Corrado Gini,[1] (Δ) is the sum of all possible differences between the family incomes in the distribution, regardless of sign, divided by the number of such differences:

$$(2;1)\ \Delta = \frac{2\sum_{m=2}^{N}\sum_{j=1}^{m-1}(x_m - x_j)}{N\,(N-1)}\,,\ (j = 1, 2, \ldots, N;\ m > j).$$

For our purposes it is convenient to use the measure t

$$(2;2)\ t = \frac{N-1}{N}\,\Delta$$

which is the mean interindividual difference including 'auto-comparisons' of incomes, i.e., comparisons of each income with itself.[2] The inclusion of 'auto-differences' obviously does not affect the numerator of (Δ) since each is zero; but it raises the number of counted differences to N^2. Henceforth we shall call t the *mean difference.*

The coefficient of concentration (R') as used in this study is the ratio of two areas in a Lorenz diagram: (1) the area of the polygon between the line of perfect equality (diagonal) and the bits of straight lines linking the plotted points and (2) the total area under the line of perfect equality (see Chart C 1 for illustration of the Lorenz curve and the graphic development of the R' formula).

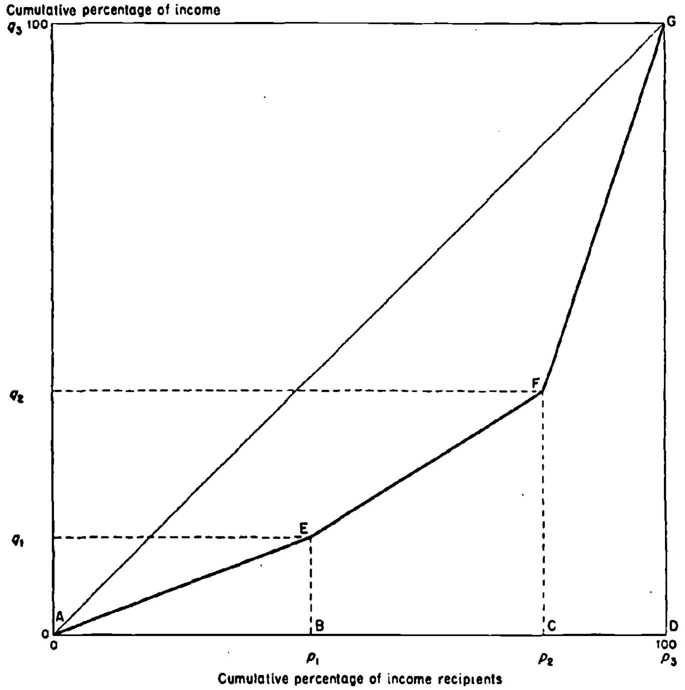
The ordered family incomes are grouped into s classes. We establish the cumulative percentage p_i ($i = 1, 2, \ldots, s$) formed by all families having incomes below the upper limit of class i and the cumulative percentage their incomes form of the aggregate income: q_i. The plotted points have the coordinates p_i, q_i. The percentage of families within class i is called r_i.

[1] Variabilità e Mutabilità, *Studi Economico-giuridici pubblicati per cura della Facoltà di giurisprudenza della R. Università di Cagliari* (Bologna, 1912), III, Part 2.

[2] Gini proved that t can also be interpreted as the mean weighted difference between the observations and their median, the *rank difference* between observations and median serving as weight (*ibid.*, pp. 32-3).

CHART C1

Illustration of Lorenz Curve


Cumulative percentage of income
q_3 100
q_2
q_1
0
A
E
F
G
B
C
D
0
p_1
p_2
100
p_3
Cumulative percentage of income recipients

It can easily be shown that the area of the polygon *AEFG* is:

$\frac{1}{2}\Sigma r_i\,(q_i + q_{i-1})$

Area of triangle ADG = $\frac{1}{2} \cdot 10{,}000$

$$\text{Area of polygon ADGFE} = \tfrac{1}{2} p_1 q_1 + \tfrac{1}{2}(p_2 - p_1)(q_2 + q_1) + \dots + \\ + \tfrac{1}{2}(p_s - p_{s-1})(q_s + q_{s-1})$$

Since $\left.\begin{array}{rl} p_1 & = r_1 \\ p_2 - p_1 & = r_2 \\ p_s - p_{s-1} & = r_s \end{array}\right\}$ the area of the polygon ADGFE is $\frac{1}{2}\sum_{i=1}^{s} r_i\,(q_i + q_{i-1})$.

Area AEFG = Area ADG − Area ADGFE, and

$$(2;3)\quad R' = \frac{\text{Area AEFG}}{\text{Area ADG}} = 1 - \frac{\sum_{i=1}^{s} r_i\,(q_i + q_{i-1})}{10{,}000}$$

Gini proved[3]

(2;4) $R' = t'/2\bar{x}$,

where t' represents an approximation to t. The measure t is based on individual incomes, t' on income groups within which there is no income variation (by assumption). Because of the neglect of intragroup variation, $t' < t$, the more so the smaller the ratio s/N. Similarly $R' < R$, which is computed from ungrouped data.

Apart from the ratio s/N the comparative size of the relative frequencies within the various classes plays a role. The difference $R - R'$ will be larger the more unequal the class frequencies. In comparing different distributions of equal or similar N's by their R's it is advisable to employ an identical or similar system of classes furnishing a similar distribution of class frequencies over the individual classes.

In our study of identical samples, s/N is the same for each sample in both years. Beginning with about 10 groups, even a sizable increase in the number of groups has only a moderate effect on the value of R'. For instance, the R' computed for the usable sample of 1929 incomes of Minneapolis tenants is .338 with 11 income groups; .349 with 39 income groups. Chart C 2 shows the Lorenz curves in the two cases.

As aggregate variance can be split into components, so the sum of interindividual differences can be subdivided. Returning to the case of ungrouped incomes we have from (2;1) and (2;2):

$$(2;5)\ t = \frac{2 \sum_{m=2}^{N} \sum_{j=1}^{m-1} (x_m - x_j)}{N^2},\ (j = 1, 2, \ldots N; m > j)$$

Calling the N_l lower (lower income group) incomes: $x_1, x_2, \ldots x_k$, and the N_u higher (upper income group) incomes: $x_{k+1}, x_{k+2}, \ldots, x_N$, where $N_l + N_u = N$, and designating the mean difference within the lower income group as t_l, the mean difference within the upper income group as t_u, and the mean incomes of the lower group, the upper group, and the aggregate $\bar{x}_l$, $\bar{x}_u$, and $\bar{x}$ respectively, we shall prove

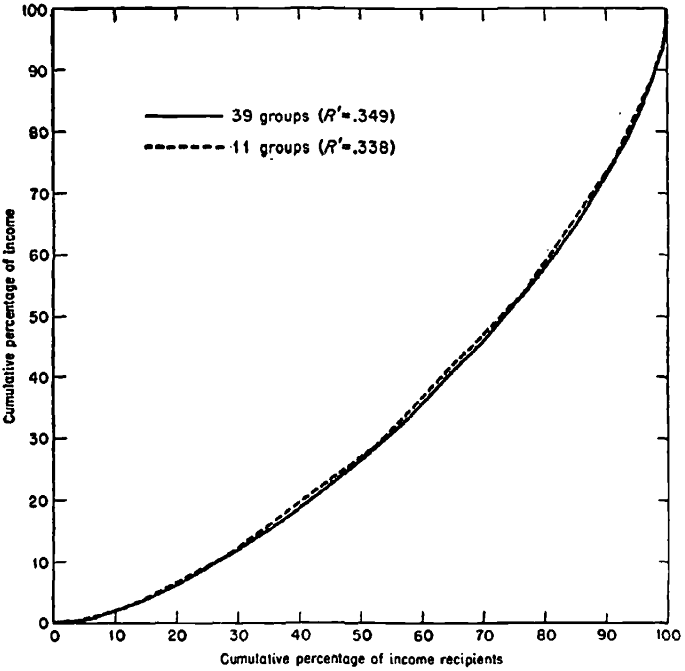
$$(2;6)\ \sum_{m=2}^{N} \sum_{j=1}^{m-1} (x_m - x_j) = \frac{N_l^2}{2} t_l + \frac{N_u^2}{2} t_u + N_l N_u (\bar{x}_u - \bar{x}_l), \text{ where}$$

$$(2;7)\ t_l = 2\bar{x}_l R_l = \frac{2 \sum_{m=2}^{k} \sum_{j=1}^{m-1} (x_m - x_j)}{N_l^2}, \text{ and}$$

[3] Sulla Misura della Concentrazione e della Variabilità dei Càratteri, *Atti del R. Instituto Veneto di Scienze, Lettere ed Arti,* 1913–14, Vol. 73, Part 2, pp. 1208–39. The present calculation of the area ratio, suggested by Professor L. Hersch of the University of Geneva, Switzerland, is much simpler than the one shown by Gini.

CHART C2

Lorenz Curves for Minneapolis Tenants in 1929
11 and 39 Income Groups



$$t_u = 2\bar{x}_u \, R_u = \frac{2 \sum\limits_{m=k+2}^{N} \sum\limits_{j=k+1}^{m-1} (x_m - x_j)}{N_u^2} .$$

We know that:

$$\begin{aligned}(2;8) \quad \sum_{m=2}^{N} \sum_{j=1}^{m-1} (x_m - x_j) = {} & (x_2 - x_1) + (x_3 - x_2) + (x_4 - x_3) + \ldots + (x_N - x_{N-1}) + \\ & + (x_3 - x_1) + (x_4 - x_2) + \ldots + (x_N - x_{N-2}) + \\ & + (x_4 - x_1) + \ldots\ldots\ldots\ldots + \\ & \ldots\ldots\ldots\ldots \\ & + (x_N - x_1).\end{aligned}$$

This sum can be considered as the sum of three items, *A, B,* and *C*:

$$\begin{aligned}A = {} & (x_2 - x_1) + (x_3 - x_2) + \ldots + (x_k - x_{k-1}) + \\ & + (x_3 - x_1) + \ldots\ldots\ldots\ldots + \\ & \ldots\ldots\ldots\ldots \\ & + (x_k - x_1).\end{aligned}$$

$$\begin{aligned} B = (x_{k+2} - x_{k+1}) + (x_{k+3} - x_{k+2}) + \ldots + (x_N - x_{N-1}) + \\ + (x_{k+3} - x_{k+1}) + \ldots\ldots\ldots\ldots\ldots + \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ + (x_N - x_{k+1}). \end{aligned}$$

$$\begin{aligned} C = {} & (x_{k+1} - x_1) + (x_{k+2} - x_1) + \ldots + (x_N - x_1) + \\ & (x_{k+1} - x_2) + (x_{k+2} - x_2) + \ldots + (x_N - x_2) + \\ & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ & (x_{k+1} - x_k) + (x_{k+2} - x_k) + \ldots + (x_N - x_k) \,. \end{aligned}$$

It is easy to see that A (B) is the sum of the interindividual differences within the lower group (upper group), i.e.,

$$(2;9)\ A = \sum_{m=2}^{k} \sum_{j=1}^{m-1} (x_m - x_j) = \frac{N_l^2}{2} t_l, \text{ and}$$

$$(2;10)\ B = \sum_{m=k+2}^{N} \sum_{j=k+1}^{m-1} (x_m - x_j) = \frac{N_u^2}{2} t_u \,.$$

The character of expression C becomes apparent when we add the first items in the parentheses over the N_l columns, the second items over the N_u rows.

$$\begin{aligned} C &= N_l\,(x_{k+1}) + N_l\,(x_{k+2}) + \ldots + N_l\,x_N - N_u\,x_1 - N_u\,x_2 - \ldots - N_u\,x_k \\ &= N_l \sum_{k=1}^{N} x - N_u \sum_{1}^{k} x. \end{aligned}$$

$$(2;11)\ C = N_l\,N_u\,(\bar{x}_u - \bar{x}_l)$$

Summing (2;9), (2;10), and (2;11), we obtain (2;6). Q.E.D.
Therefore

$$(2;12)\ t = \frac{N_l^2\,t_l + N_u^2\,t_u + 2N_l\,N_u\,(\bar{x}_u - \bar{x}_l)}{N^2}$$

and making use of (2;4)

$$(2;13)\ R = \frac{1}{2N^2\bar{x}}\left[N_l^2\,t_l + N_u^2\,t_u + 2N_l\,N_u\,(\bar{x}_u - \bar{x}_l)\right]$$

For grouped incomes, the situation is identical, except that instead of t and R we have their approximations t' and R'.

$$(2;14)\ R' = \frac{1}{2N^2\bar{x}}\left[N_l^2\,t_l' + N_u^2\,t_u' + 2N_l\,N_u\,(\bar{x}_u - \bar{x}_l)\right].$$

Introducing the coefficients of concentration for the lower and upper income groups we obtain by reason of (2;4)

$$(2;15)\quad R' = \frac{1}{N^2\bar{x}} \left[N_l^2\, \bar{x}_l\, R_l' + N_u^2\, \bar{x}_u\, R_u' + N_l\, N_u\, (\bar{x}_u - \bar{x}_l) \right].$$

The coefficient of concentration for the entire distribution may be considered as the sum of three components:

(2;17) weighted inequality within the lower income group:

$$I_l = \frac{N_l^2\, \bar{x}_l}{N^2\bar{x}}\, R_l',$$

(2;18) weighted inequality within the upper income group:

$$I_u = \frac{N_u^2\, \bar{x}_l}{N^2\bar{x}}\, R_u',$$

(2;19) weighted inequality between the lower and upper income groups:

$$I_{lu} = \frac{N_l\, N_u}{N^2}\; \frac{\bar{\ }_u - \bar{x}_l}{\bar{x}}$$