ESTIMATION OF THE MEAN FROM CENSORED INCOME DATA

Sandra A. West, Bureau of Labor Statistics

1. INTRODUCTION

In recent years economists and sociologists have increasingly become interested in the study of income attainment and income inequality as is reflected by the growing numbers of articles which treat an individual's income as the phenomenum to be explained. Some research on income have included the investigation of the effects of status group membership on income attainment and inequality: the determinants of racial or ethnic inequality; the determinants of gender inequality; and the determinants of income and wealth attainment in old age.

Although many of the studies used census data, research on income has come increasingly to rely on data from social surveys. The use of survey data in the study of income can present a serious measurement problem since respondents' incomes most often are not measured in their true dollar amounts but in categories with the last category being open-ended. The problem with this categorical measurement is the most acute when the researcher wants to estimate income through the application of a statistical technique, such as regression, which assumes specific measurements. The most common way for changing categorical income variables into measurement variables has been to assign all incomes within a specific category to the midpoint of that category. However, this procedure still presents the problem of how to estimate the mean (or midpoint) for the upper income category which is open ended. In this paper the problem of estimating the mean of the open ended interval is considered, but the problem arose in a different manner.

The project underlying this paper began in connection with a program at the Bureau of Labor Statistics (BLS) that reports weekly earnings of wage and salary workers. The data on usual weekly earnings are collected in the Current Population Survey (CPS), a national survey of households conducted for the Bureau of Labor Statistics by the Bureau of the Census. Estimates published quarterly include median usual weekly earnings of full-time wage and salary workers by age, race, and sex. In addition comparisons are often made over time. A function of the medians often considered is the ratio of women's earnings to men's earnings at any given point in time and the change of the ratio over time. Due to problems inherent in operating with medians, see West (1985a), it was decided to also compute means. However, the exact mean could not be computed since due to operational procedures the data received by the BLS are censored at $999; that is any person making $999 or more per week is coded as making $999 .

The problem will be considered from the point of view of computing a population mean from censored data. One way to approach this problem is to fit a theoretical distribution to the upper portion of the observed income distribution and then determine the conditional mean of the upper tail. An estimator of this conditional mean will be referred to as a tail estimator. In this paper a truncated Pareto distribution is used and a modified maximum likelihood estimator is developed for the Pareto parameter. Other estimators of the Pareto parameter are also considered. In particular it is shown that the estimator most used in applications can lead to very misleading results. In order to see the effect of the tail estimator, six actual populations, consisting of income data that were not censored, were considered. Different estimates of the tail and overall mean were compared to the true values for twenty subpopulations. In Section 2 the problem is formulated. In Section 3 the Pareto distribution is discussed and methods of estimating the parameter are considered. In particular, a maximum likelihood estimator using truncated and censored data is developed. In Section 4, the two leading candidates for the overall mean, using a tail estimator, are computed on six real populations (twenty sub-populations), where the true tail is known. Section 5 contains the conclusions. In addition to the usual properties of means, there is another nice feature that is discussed in the Appendix. It is shown that under a mild assumption the percent difference between two means is bounded relative to the percent difference between subgroup means.

2. FORMULATION OF PROBLEM

True population data are: $X_1, X_2, \ldots, X_N$. Let Y_i be the ordered X's up to and including some fixed number U. Assume there are (N-t) observations of value U and over. The data actually observed are:

$$Y_1 \; Y_2, \ldots Y_t, \; (N-t) \; U\text{'s}.$$

The mean of the complete population is desired; that is,

$$\bar{X} = \sum_{i=1}^{N} X_i/N \tag{2.1}$$

An obvious estimator is:

$$\bar{Y} = \left[\sum_{i=1}^{t} Y_i + U(N-t) \right]/N, \tag{2.2}$$

which is known as the Winsorized mean. This estimator underestimates the true mean and overestimates the ratio of women's earnings to men's earnings. A natural extension of the Winsorized mean is discussed in this paper. A theoretical distribution is fit to the censored data using the observed income data and then an overall mean is determined. Since much income data available are not in exact dollar amounts but in categories, the problem will be reformulated in terms of grouped data. The observed data are in the following form:

Income Intervals	f = frequency
$U_o \le y < U_1$	f_1
$U_1 \le y < U_2$	f_2
⋮	⋮
$U_{r-2} \le y < U_{r-1}$	f_{r-1}
$y \ge U_{r-1} = U$	f_r

where $\sum_{i=1}^{r} f_i = N$, and $f_r = N-t$.

Let $\hat{M}_i$ be the mid-point of the i-th interval, for i = 1, 2, ..., r-1, and $\hat{M}_r$ be an estimator of the mid-point of the r-th interval, then an estimator of the mean is:

$$\hat{\bar{X}} = [\sum_{i=1}^{r-1} M_i f_i + \hat{M}_r f_r]/N. \qquad (2.3)$$

One possibility for $\hat{M}_r$ is to fit a theoretical distribution to the (r-1) mid-points and then $\hat{M}_r$ would be the mean of the conditional distribution, $P(X \le x | X \ge U)$. That is,

$$\hat{M}_r = \int_U^\infty x \, dP(X < x | X > U) = \int_U^\infty x \, f(x|U) \, dx$$

where $f(x|U)$ is the conditional density of X given that X is greater than the fixed number U.

Another possibility for $\hat{M}_r$ is the median of the conditional density. Parker and Fenwick (1983) found that this estimator performed better than using the mean, but this was not the case with the method developed in this paper.

Many distributions have been proposed for income data, see Theil (1967), Arnold (1983). From the literature it seems that researchers are satisfied with the Pareto distribution as a fit to the upper portion of the income curve. In this paper only the Pareto density will be considered, but different methods of estimating the parameters will be investigated.

3. PARETO DISTRIBUTION

Consider the classical Pareto distribution

$$F(x) = P(X \le x) = 1-(K/x)^\alpha \text{ for } x \ge K \ge 0, \ \alpha > 0$$
$$= 0 \text{ for } x < K. \qquad (3.1)$$

Noting that $P(X \ge x | X \ge U) = P(X \ge x)/P(X \ge U)$ $= U^\alpha x^{-\alpha}$ then $f(x|U) = -\, d\, P(X \ge x | X \ge U)/dx =$ $\alpha U^\alpha x^{-\alpha-1}$, $x \ge U$. Thus,

$$\hat{M}_r = \int_U^\infty x \, f(x|U) dx = [\alpha/(\alpha-1)] \, U. \qquad (3.2)$$

In practice, it is frequently assumed that $\alpha > 1$, so that the distribution has a positive finite mean. Note that the sum of two independent Pareto random variables has a distribution which has a Paretian tail. Specifically, let X_1 and X_2 be independent Pareto random variables with parameters α_1 and α_2 respectively. Letting $Z = X_1 + X_2$ then it can be shown that as $x \to \infty$

$$\text{Prob } (Z>x) \approx x^{-\min(\alpha_1,\alpha_2)} H(x), \qquad (3.3)$$

where H is a slowly varying function of x. This result gives insight into the persistence of Paretian tail.

Parameter estimation for the classical Pareto distribution has been fairly well investigated from the point of view of point estimation. Interval estimation has not been extensively investigated. A new generalized Bayesian approach for interval estimation of the Pareto parameters has been developed by West (1986). In this paper three popular methods for point estimation will be considered: 1) Least squares, 2) Quantiles; and 3) Maximum likelihood.

Since the Pareto distribution is considered a good fit for the income distribution over the higher portion, the parameter will be estimated from the left truncated distribution. Let M_s, M_{s+1}, ... M_{r-1} denote the truncated and non-censored data.

For the least squares estimator, note that if

$$M_s \le M_i < U, \ F(M_i) = 1 - M_s^\alpha / M_i^\alpha \quad \text{then}$$
$$\ln [1- F(M_i)] = \alpha \ln M_s - \alpha \ln M_i. \qquad (3.4)$$

Letting $Y_i = \ln[1-F(M_i)]$ and $X_i = \ln M_i$

where $f(M_i) = \sum_{j=s}^{i} f_j \Big/ \sum_{l=s}^{r} f_l$ for i = s, s+1, ..., (r-1) then the least squares estimator for α is:

$$\hat{\alpha} = \frac{-(r-s) \sum_{i=s}^{r-1} X_i Y_i + (\sum_{i=s}^{r-1} X_i)(\sum_{i=s}^{r-1} Y_i)}{(r-s) \sum_{i=s}^{r-1} X_i^2 - (\sum_{i=s}^{r-1} X_i)^2} \qquad (3.5)$$

Equation (3.4) implies that as M_i, the level of income under consideration increases, the number of people in the population who have incomes greater than M_i decreases. However, the Pareto Curve as represented in (3.4) is linear only in the upper portions of an income distribution. In the literature the least squares method is one of the two principal methods for the estimation of α.

The second is estimation from quantiles, described below. Let M_p and M_q denote the p-th and q-th quantile respectively; that is,

$$F(M_p) = P(M \le M_p) = 1 - (K/M_p)^\alpha = p. \qquad (3.6)$$

Letting $\hat{M}_p$ and $\hat{M}_q$ be estimators of M_p and M_q respectively, leads to the following estimator of α:

$$\hat{\alpha}_{pq} = \ln((1-p)/(1-q))/\ln(\hat{M}_q/\hat{M}_p). \qquad (3.7)$$

Most reseachers seem to use this method with either the mid-points of the last two closed intervals or the last closed interval and the open interval. Specifically, if the mid-points of the last two closed intervals are used (3.7) becomes

$$\hat{a}_c = \ln(f_r/(f_r + f_{r-1}))/\ln(M_{r-2}/M_{r-1}). \quad (3.8)$$

If the lower bounds of the last closed interval and the open interval are used, then (3.7) becomes

$$\hat{\alpha}_o = \ln((f_{r-1} + f_r)/f_r)/\ln(U/L_{r-1}). \quad (3.9)$$

In the literature the estimator in (3.9) seems to be the one most recommended, for example see Shryock (1975), Parker and Fenwick (1983).

Consistency is easily verified for the quantile estimator and it is resistant to outliers. Quandt (1966) found that the performance of the quantile estimates was not much inferior to those of the maximum likelihood estimates. A Monte Carlo study reported by Koutrouvelis (1981) supports that view. However, it will be seen, empirically, in Section 4 that this method depends very much on the classification of the population. It can lead to gross errors and at best it does as well as the maximum likelihood estimator. Also, it can be shown theoretically that the quantile estimator is inferior to other available estimators. This is done by rewriting (3.7) in the following form:

$$(c + \sum_{i=1}^{m} d_i V_i)^{-1} \quad (3.10)$$

where the V_i's are independent exponential variables and c and the d_i's are constants.

This representation is arrived at by using the following well known theorem, David (1970). Let $V_1, V_2, \ldots, V_m$ be a sample of size m from the exponential distribution ($F(V) = 1 - \exp(-\mu V), V>0$) with corresponding order statistics W_i, i=1, 2...m, and spacings $Z_i = W_i - W_{i-1}$, i=1, 2, ..., m ($W_0 = 0$ by definition). The spacings are independent exponential random variables. Moreover, the random variables $\{(m-i+1)\ Z_i,\ i=1, 2, \ldots, m\}$ are independent identically distributed random variables with the common exponential distribution. The direct relationship between the Pareto distribution and the standard exponential distribution ($X_i = K \exp(V_i/\alpha)$) renders this theorem an important tool in discussing the distribution of Pareto order statistics. Moments of random variables in the form of (3.10) can be derived. With these formulas it can be demonstrated that the quantile estimators are inferior to other available estimators.

In the rest of this section the maximum likelihood estimator will be considered. First consider a random variable M that can take on the following values with corresponding probabilities.

$$\text{if } K \le M < U \quad f(M) = \alpha K^{\alpha}/M^{\alpha+1} \quad \alpha > 0,\ K > 0$$

$$\text{if } M = U \quad f(M) = 1 - F(U) = (K/U)^{\alpha} \quad (3.11)$$

where U is a fixed constant.

Suppose we observe r distinct values for M:
$M = M_i$ with frequency f_i where $K < M_i < U$, for i = 1, 2, ...(r-1).
$M = U$ with frequency f_r.

where $\sum_{i=1}^{r} f_i = N$.

The likelihood function L is:

$$L = \binom{N}{f_r} \prod_{i=1}^{r-1} [\alpha K^{\alpha}/M_i^{\alpha+1}]^{f_i} [(K/U)^{\alpha}]^{f_r} \quad (3.12)$$

which leads to the maximum likelihood estimator of α:

$$\hat{\alpha} = (N - f_r)/[(\sum_{i=1}^{r-1} f_i \ln M_i) - (N - f_r)\ln \hat{K} - f_r \ln \hat{K}/U]. \quad (3.13)$$

Note that the maximum likelihood estimator of K is the minimum of M_i for i = 1, ... r.

If the distribution is truncated on the left at M_s, the estimator becomes:

$$\hat{\alpha} = \sum_{i=s}^{r-1} f_i \Big/ [\sum_{i=s}^{r-1} f_i \ln M_i - \ln M_s \,(\sum_{i=s}^{r-1} f_i) - f_r \ln(M_s/U)]. \quad (3.14)$$

Note that in the case of a Pareto distribution, truncation is equivalent to rescaling, $K=M_s$.

In the next section six different truncation points are considered. It is shown that the smallest errors occur when the starting point is near the truncated mean.

In determining (3.14) the fact that there is an interval of data was not taken into account. The likelihood for grouped data leads to a maximum likelihood estimator that can only be obtained through iterative techniques. Specifically, suppose for the N observations, $X_1, X_2, \ldots, X_N$, all that is reported about a particular X_i is the interval into which it falls $[L_i, U_i]$, i=1, 2, ... r-1; and if $X_i \ge U$ all that is reported is $[U, \infty)$. Noting that $P\{L_i \le X \le U_i\} = K^{\alpha}[L_i^{-\alpha} - U_i^{-\alpha}]$, $P\{X \ge U\} = K^{\alpha} U^{-\alpha}$, and that it is observed that the interval (L_i, U_i) contains f_i observations for i = 1, 2, ..., r-1, and the interval (U, ∞) contains f_r observations then the likelihood is:

$$L = N!\, K^{\alpha N} U^{-\alpha f_r} \prod_{i=1}^{r-1} [L_i^{-\alpha} - U_i^{-\alpha}]^{f_i} \Big/ \prod_{i=1}^{r} f_i!$$

The maximum likelihood estimator can be obtained in closed form only if $U_i = c\, L_i$, for some constant c; this is not the case in the

current situation. It will be seen in the next section that the maxiumum likelihood estimator in (3.14) leads to fairly accurate estimates.

4. EVALUATION OF MEAN ESTIMATORS ON SIX REAL POPULATIONS

In this section the two leading candidates for tail estimators will be evaluated on six real populations, where the true tail is known. The two candidates are the modified maximum likelihood estimator (3.14) derived in the previous section and the quantile estimator (3.9) that is recommended in the literature. The six populations are briefly described.

Two years of income data were obtained from the 1982 and 1983 Consumer Expenditure Interview Survey, CE, conducted by the Bureau of Labor Statistics. The Interview Survey is one component of the Consumer Expenditure Survey Program. It uses a national probability sample of households and is designed to collect data on the types of expenditures which respondents can be expected to recall for a period of 3 months. Information is collected on demographic and family characteristics and on the inventory of major durable goods of each consumer unit. In the fifth and final interview, an annual supplement is used to obtain a financial profile of the consumer unit. Only people who worked full time (35 or more hours per week and 50 or more weeks per year) were used. The yearly income recorded was transformed into average weekly earnings. Five groups are constructed: all, men, women, age 16-24, and age 25 and over.

One year of income data was obtained from the 1979 wave of the Panel Study of Income Dynamics, PSID, conducted by the Institute for Social Research. The data were divided into the same five groups as the CE data.

One year of income data was obtained from a special study conducted in 1977 to gauge the accuracy of the earnings data derived from the Current Population Survey, CPS. Wage and salary workers in one-eighth of the CPS sample were asked in January, to supply information on how they were paid, how much they earned and how many hours they worked. With their permission, the same information was then obtained by mail from their employers. Again only full time workers were used. The data were divided into three groups: all, men and women.

The last two sets of data were obtained from an article by Parker and Fenwick (1983). The authors use the 1976 and 1977 waves of the Panel Study of Income Dynamics. The PSID '76 and '77 data are joint husband and wife yearly income. All the distributions are displayed in West (1985b).

The studies performed on the three populations, CE '82, '83 and PSID '79, will be described first. For each of the five categories in the three populations the data were grouped in five different ways: \$1, \$10 centered and uncentered intervals, and \$50 centered and uncentered intervals. The uncentered intervals are the usual intervals of equal width starting with zero. The centered intervals are centered around multiples of ten. After the true means were computed on the entire set of data, the data were truncated at \$999 and overall means with estimated tails were computed for each of the possibilities. Of the two estimating techniques, the maximum likelihood, (3.14), has a second parameter, which will be referred to as the starting point. In order to determine how much of the income data should be used in estimating the Pareto parameter α, six starting points were considered. The parameter α was estimated using data starting with \$200, \$300, \$400, \$500, \$600 and \$700. Empirically it was found that if the earnings distribution was truncated at the mid-point, M_s, of the interval containing the truncated mean then the resulting estimate of α led to the estimate of the mean that was the closest to the true mean.

The overall mean reported in the tables is the mean computed on \$1 intervals for data up to \$999 combined with the estimated tail mean; that is, letting $S = \{i \mid X_i < 999\}$ then

$$\bar{X} = \left(\sum_{i \in S} X_i + \hat{M}_r f_r\right)/N \qquad (4.1)$$

$$= \left(\text{TRM}\,(N-f_r)/N\right) + \hat{M}_r f_r/N$$

where $\text{TRM} = \left\{\sum_{i \in S} X_i\right\}/(N - f_r)$. The percent error of the estimator is determined by

$$PE = 100(\mu - \bar{X})/\mu \qquad (4.2)$$

where μ is the mean computed on \$1 intervals for the complete population. The percent error of the tail contribution to the mean (or the tail mean) is defined as:

$$100\left(\sum_{i \in S'} X_i - \hat{M}_r f_r\right)/\sum_{i \in S'} X_i, \qquad (4.3)$$

where S' is the complement of S.

Due to space limitations, the empirical results will be summarized and four tables will be displayed as an example of the results. Additional tables are in West (1985b). In almost every case the centered intervals led to smaller errors than the uncentered intervals. For the maximum likelihood estimator, out of 30 cases 83% had absolute percent errors between 0 and 1; the remaining 17% had absolute percent errors between 1.1 and 5.2. The quantile method could not be used in several cases because it led to an $\alpha < 1$, which yields a negative mean. This method depends on the classification of the data, can lead to gross errors - such as 236% and at best does as well as the modified maximum likelihood method.

In Tables 1-3, the mean using the modified maximum likelihood method and the mean using the quantile method are recorded for the 5 subgroups, using data from the 1982 and 1983 Consumer Expenditure Survey and the 1979 Panel Study of Income Dynamics. The maximum likelihood estimates were computed by chosing M_s to be in the interval that contains the truncated mean. (The data were grouped into \$10 centered intervals.) In Table 4 the means computed by the two methods are recorded for three subgroups using 1977 CPS special data.

Table 1. Mean Weekly Earnings Computed From 1982 Consumer Expenditure Survey

	MLE $\bar{X}$	Percent Error	Quantile $\bar{X}$	Percent Error
ALL	385	-.3	434	-13.1
MEN	449	-.1	486	- 8.4
WOMEN	282	4.6	275	7.0
16-24	230	.1	227	1.5
25+	420	-.6	457	-9.3

Table 2. Mean Weekly Earnings Computed From 1983 Consumer Expenditure Survey

	MLE $\bar{X}$	Percent Error	Quantile $\bar{X}$	Percent Error
ALL	392	1.0	439	-10.9
MEN	461	-.9	*	
WOMEN	292	2.8	288	4.4
16-24	229	.2	224	2.4
25+	424	1.0	467	8.9

* Method led to a negative mean.

Table 3. Mean Weekly Earnings Computed From 1979 Panel Study of Income Dynamics

	MLE $\bar{X}$	Percent Error	Quantile $\bar{X}$	Percent Error
ALL	284	0.2	315	-10.6
MEN	343	-0.1	388	13.0
WOMEN	194	0.0	194	0.2
16-24	207	0.3	206	0.5
25+	302	0.0	336	11.1

Table 4. Mean Weekly Earnings Computed From 1977 CPS - Special Data

	MLE $\bar{X}$	Percent Error	Quantile $\bar{X}$	Percent Error
ALL	389	-1.0	*	
MEN	441	-4.0	*	
WOMEN	288	-1.0	980	-236.8

* Method led to a negative mean.

5. CONCLUSIONS

In practice, estimates of α have been used in a variety of ways to characterize the inequality of an observed income distribution. For example, Bowman (1945) used α directly for comparisons of inequality. More common is the procedure of transforming α to obtain an inequality measure. The Lorenz measure of income concentration will equal $1/(2\alpha-1)$ if the income distribution is Pareto (with $\alpha>1$). The Gini measure of income concentration will equal $\alpha/(\alpha-1)$ if the income distribution is Pareto (with $\alpha>1$). Whichever function of α is the object of investigation, its best estimate (in the sense of asymptotic efficiency) will be obtained by inserting the best estimate of α in the function.

From the theoretical and empirical investigations, it is clear that the maximum likelihood estimator of α is the best. Also the variance of a function of the maximum likelihood estimator is easily determined. The quantile estimator is inferior to the maximum likelihood estimator and can lead to misleading results.

From the empirical investigation, the estimator of the population mean which uses a Pareto tail with the modified maximum likelihood estimator for the parameter α does very well and is the mean recommended. If instead of the entire population a random sample was drawn then the estimates from the sample data would be subject to sampling variation. Confidence intervals could easily be computed for the population mean.

APPENDIX

A simple geometric argument is given for the fact that the percent difference between two means is bounded relative to the percent difference between subgroup means, if the proportion of people in each subgroup remains the same over time.

Let $\bar{X}_{t1}$ and $\bar{X}_{t2}$ denote at time t the average income of two mutually exclusive groups (say men and women) of size n_{t1} and n_{t2} respectively.

Let $\bar{X}_{tT}$ denote the average income of the combined group at time t: that is,

$$\bar{X}_{tT} = (n_{t1}\bar{X}_{t1} + n_{t2}\bar{X}_{t2})/(n_{t1} + n_{t2})$$

$$= U_t \bar{X}_{t1} + (1-U_t)\bar{X}_{t2}$$

where $U_t = n_{t1}/(n_{t1} + n_{t2})$.

For each group it is often of interest to look at the change over two time periods. That is, for group 1 the change between time periods 1 and 2 is defined as:

$$(\bar{X}_{21} - \bar{X}_{11})/\bar{X}_{11} = (\bar{X}_{21}/\bar{X}_{11}) - 1.$$

Similarly, the percent change for group 2 and the combined group are respectively,

$$(\bar{X}_{22}/\bar{X}_{12})-1 \text{ and } (\bar{X}_{2T}/\bar{X}_{1T})-1.$$

It will be shown geometrically, that if $U_1 = U_2$ then the percent change for the combined group is bounded by the percent changes for the subgroup means.

Assuming $U_1 = U_2$ and $\bar{X}_{21}/\bar{X}_{11} \leq \bar{X}_{2T}/\bar{X}_{1T}$ it will be shown that $\bar{X}_{2T}/\bar{X}_{1T} \leq \bar{X}_{22}/\bar{X}_{12}$. The reverse inequalities are shown in a similar way.

Let A be the point with coordinates $(U_1 \bar{X}_{11}, U_2 \bar{X}_{21})$ and B the point with coordinates $(-(1-U_1)\bar{X}_{12}, -(1-U_2)\bar{X}_{22})$. Letting 0 be the point through the origin (0,0) then the slopes of the lines $\overline{AB}$, $\overline{OA}$ and $\overline{OB}$ are respectively $\bar{X}_{2T}/\bar{X}_{1T}$, $\bar{X}_{21}/\bar{X}_{11}$ and $\bar{X}_{22}/\bar{X}_{12}$, since $U_1 = U_2$.

Letting ϕ, ω, and Θ be the respective angles that the lines $\overline{AB}$, $\overline{OA}$ and $\overline{OB}'$ make with X-axis, as indicated in Figure 1, then

$$\text{slope } (\overline{AB}) = \bar{X}_{2T}/\bar{X}_{1T} = \tan \phi$$

$$\text{slope } (\overline{OA}) = \bar{X}_{21}/\bar{X}_{11} = \tan \Theta$$

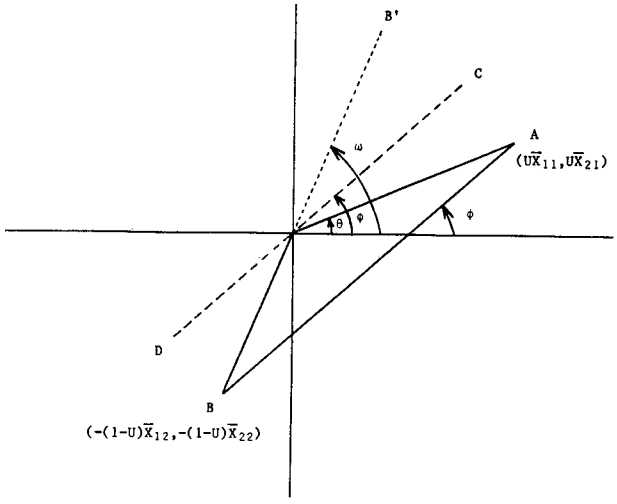
$$\text{slope } (\overline{OB}) = \bar{X}_{22}/\bar{X}_{12} = \tan \omega.$$

Given that $\Theta \leq \Phi$ it follows that $\Phi \leq \omega$. From Figure 1 it is clear that since $\Theta \leq \Phi$ then $\Phi \leq w$ for a triangle to be formed from $\overline{OA}$ and $\overline{OB}$. Thus

$$\text{slope } (\overline{AB}) \leq \text{slope } (\overline{OB}) \quad \bar{X}_{2T}/\bar{X}_{1T} \leq \bar{X}_{22}/\bar{X}_{12}$$

Figure 1

Assuming that $\theta \leq \phi$ it is shown that $\phi \leq \omega$. Geometrical proof that if $U_1 = U_2 = U$ then the percent change between group means is bounded by the percent changes for the subgroup means.



REFERENCES

Arnold, B.C. (1983). Pareto Distributions 5 International Co-operative Publishing House, Fairland, Md.

Bowman, M.J. (1945). "A Graphical Analysis of Personal Income Distribution In the United States," American Economic Review, 35, 607-628.

David, H.A. (1970). Order Statistics, Wiley, New York.

Koutrouvelis, I.A. (1981). "Large Sample Quantile Estimation in Pareto Laws," Communications In Statistics, Theory and Methods, A10, 189-201.

Parker, R.N. and Fenwick, R. (1983). "The Pareto Curve and Its Utility for Open-Ended Income Distributions in Survey Research," Social Forces, 61, 872-885.

Quandt, R.E. (1966). "Old and New Methods of Estimation and the Pareto Distribution," Metrika, 10, 55-82.

Sands, S. (1973). "Estimating Data Withheld in Grouped Size Distributions," Journal of American Statistical Association, 68, 306-311.

Shryock, H. and Siegel, J. (1975). The Methods and Materials of Demography, Washington, D.C., U.S. Government Printing Office.

Theil, H. (1967). Economics and Information Theory, Amersterdam, North Holland Publishing Company.

West, S.A. (1985a). "Standard Measures of Central Tendency for Censored Earnings Data, From the Current Population Survey", Bureau of Labor Statistics Report.

West, S.A. (1985b). "Estimation of the Mean From Censored Income Data," Bureau of Labor Statistics Report.

West, S.A. (1986). "Generalized Bayesian Inference for Pareto Parameter," (to be submitted to the Journal of the Royal Statistical Society.

ACKNOWLEDGMENTS

Thanks are due to Janice Shack-Marquez for her computer work and to Darlene King and Daphne Van Buren for typing this paper.