MEASURES OF CENTRAL TENDENCY FOR CENSORED WAGE DATA

Sandra West, Diem-Tran Kratzke, and Shail Butani, Bureau of Labor Statistics
Sandra West, 2 Massachusetts Ave. N.E. Washington, D.C. 20212

Key Words: Mean, Median, Pareto Distribution

I. INTRODUCTION

The research for this paper began in connection with the need for measuring the central tendency of hourly wage data from the Occupational Employment Statistics (OES) survey at the Bureau of Labor Statistics (BLS). The OES survey is a Federal-State establishment survey of wage and salary workers designed to produce data on occupational employment by industry for the Nation, each State, and selected areas within States. The OES survey provides employment data for approximately 700 detailed occupations by surveyed industries. Until recently the survey did not provide any wage information. The wage data that are produced by other Federal programs are limited in the level of occupational, industrial, and geographic coverage. In order to address this critical void in the Federal statistical effort, the OES program conducted pilot studies in 1989 and 1990 to test the feasibility of incorporating wage questions into the OES survey.

The 1992 OES survey collects data in 15 States on occupational hourly wage by industry in nonagricultural establishments. The data are collected in eleven intervals, rather than in exact dollar amounts, with the lowest and uppermost intervals open.

Research was conducted by the Office of Research and Evaluation and the Statistical Methods Division of the Office of Employment and Unemployment Statistics to find a suitable estimate of central tendency for the occupational wage data of the OES survey. It was determined that both mean and median would be measured. Each has advantages and disadvantages which will be discussed.

West (1986) discussed problems with using medians in household survey data where the distribution of earnings has many peaks and estimates are compared over time periods. In this paper, we analyze the medians of occupational wages from establishment survey data where the resulting distributions are not as multi-peaked as distributions of data from household surveys. For the most part of this paper, we investigate alternative methods of estimating occupational wage means from grouped data with open intervals.

The first part of the research focused on the problem of estimating the overall occupational wage mean for each industry, according to the OES survey's objective. The second part of the research explores the best method for estimating the wage mean of an upper open interval. This would be useful for analyses such as regression where interval wages are used as dependent variables.

The two measures of central tendency are described in Section II. The empirical studies and results are given in Section III. The conclusions are presented in Section IV. Section V contains plans for future research.

II. MEASURES OF CENTRAL TENDENCY

A. MEDIANS

In elementary theory the median has considerable claims to be used as a measure of location for unimodal distributions. It is readily interpretable in terms of ordinary ideas. What gives the arithmetic mean the greater importance in advanced theory is its superior mathematical tractability and certain sampling properties. The median has a compensating advantage in that it is less sensitive to the configuration of the outlying parts of the frequency distribution than is the mean. This is especially important with earnings data and in particular, with the censored data, the median is a logical choice. However, the median is sensitive to the way the data are grouped and operating with medians can lead to misleading results. The latter is especially true in the case of the many-peaked earnings distribution, such as arises in data collected from households, where the respondent often has to approximate the data requested. The earnings data considered in this paper are collected from establishments and are, for the most part, obtained from payroll records. Thus, the resulting distribution should not be multi peaked, and indeed, are not. This paper concentrates on finding the best estimator for the mean rather than for the median. Only the linear interpolation method for the median is tested. The method first determines the interval that contains the median, and then linear interpolation is used to estimate the median. The occupational minimum wage is used as the lower limit of the lower open interval. If the median falls in the uppermost open interval, all that can be said is that the median is equal to or above the upper limit of hourly wage.

B. MEANS

A measure of central tendency with desirable properties is the mean. In addition to the usual desirable properties of means, there is another nice feature that is proven by West (1985). It is that the percent difference between two means is bounded relative to the percent difference between subgroup means, if the proportion of units in each subgroup remains the same for the two groups. Since the problem considered in this paper deals with grouped data that have lower and upper open intervals, it is not possible to compute an exact mean. The problem will be considered from the point of view of computing a population mean from right and left censored data. First the problem will be formulated and two methods for computing the mean will be discussed. One method results in the Winsorized mean and the other method uses a classical Pareto distribution.

FORMULATION OF PROBLEM

The population of true earnings data are denoted by :

$$X_1, X_2, \ldots, X_N,$$

and $X_{(1)}, X_{(2)}, \ldots, X_{(N)}$ denote the ordered $X's$. The mean of the population is desired; that is,

$$\overline{X} \;=\; \sum_{i=1}^{N} X_i \Big/ N.$$

The data actually observed are the frequencies of each of the intervals:

$$I_1, I_2, \ldots, I_r,$$

which are mutually exclusive, and exhaustive of the earnings scale. I_1 and I_r are open intervals, where

I_1 contains all $x_{(j)}$ less than some fixed number U_1, and

I_r contains all $x_{(j)}$ greater than or equal to some fixed number $U_{(r-1)}$.

For $i = 2,3,\ldots,(r-1)$, I_i are bounded intervals containing all $x_{(j)}$ between some fixed numbers $U_{(i-1)}$ and $U_{(i)}$.

Let f_i denote the observed frequency of interval I_i, $for\ i = 1,2,\ldots,r$. Note that $\sum_{i=1}^{r} f_i = N$.

Letting M_i denote the mid-point of the i-th interval, then the usual estimator of $\overline{X}$ is the grouped mean, $\overline{X}_g$::

$$\overline{X}_g = \sum_{i=1}^{r} M_i f_i \Big/ N \ .$$

Since intervals I_1 and I_r are not bounded intervals, M_1 and M_r will need to be estimated.

I_1 has a natural lower bound, either 0 or the minimum wage; for this study the minimum occupational wage for each State, $W_{(1)}$, was used. Thus the estimate for M_1 is :

$$\hat{M}_1 = (W_{(1)} + U_1)/2 .$$

An obvious estimate for M_r is:

$$\hat{M}_{r,w} = U_{(r-1),}$$

which would lead to the Winsorized group mean:

$$\hat{\overline{X}}_{g,w} = \{\hat{M}_1 f_1 + \sum_{i=2}^{r-1} M_i f_i + U_{(r-1)} f_r\} \Big/ N .$$

Clearly, $\hat{M}_{r,w}$ will underestimate M_r . With the Winsorized mean a straight line is used for the missing values; a natural extension, now to be considered, is to fit a curve for the missing values.

For the estimator of M_r , consider fitting a theoretical distribution to the (r-1) mid-points, and take $\hat{M}_r$ as the mean of the conditional distribution, $P(X \le x | X \ge U_{(r-1)})$. That is,

$\hat{M}_r = \int_{U_{(r-1)}}^{\infty} x \, dP(X \le x | X \ge U_{(r-1)}) = \int_{U_{(r-1)}}^{\infty} x \ f(x|U_{(r-1)}) \, dx$ where $f(x|U_{(r-1)})$ is the conditional density of X, given that X is greater than or equal to the fixed number $U_{(r-1)}$.

Another possibility for $\hat{M}_r$ is the median of the conditional density. Parker and Fenwick (1983) found that this estimator performed better than the mean, but this was not the case with the new method and data considered in West (1985).

Many distributions have been proposed for earnings data, but it is clear from the literature that the researchers are satisfied with the Pareto distribution as a fit to the upper portion of the earnings curve. Consider the Pareto distribution:

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

Noting that,

$$P(X \ge x | X \ge U_{(r-1)}) = P(X \ge x)/P(X \ge U_{(r-1)}) = (U_{(r-1)})^{\alpha} x^{-\alpha}$$

then

$$f(x|U_{(r-1)}) = -dP(X \ge x | X \ge U_{(r-1)})/dx$$
$$= \alpha (U_{(r-1)})^{\alpha} x^{-\alpha-1}, \qquad for\ x \ge U_{(r-1)}.$$

Thus,

$$\hat{M}_r = \int_{U(r-1)}^{\infty} x \ f(x| U_{(r-1)})dx = U_{(r-1)} \alpha/(\alpha - 1).$$

Note that for the mean to be positive, $\alpha > 1$.

Many methods exist for estimating the parameter α. The method for estimating α most used and recommended in the literature is the quantile method, which is described next.

Let M_p and M_q denote the $p-$th and $q-$th quantile respectively; that is,

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

Similarly, for M_q.

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

$$\hat{\alpha}_{pq} = \ln[(1-p)/(1-q)] \Big/ \ln[\hat{M}_q \big/ \hat{M}_p].$$

Most researchers seem to use this method with either the mid-points of the last two bounded intervals or the last bounded interval and the open interval. Specifically, if the mid-points of the last two bounded intervals are used, the estimator of α becomes:

$$\hat{\alpha}_c = \ln[(f_r)/(f_r + f_{r-1})]/\ln[M_{r-2}/M_{r-1}].$$

This will lead to $\hat{M}_{r,q,2}$ and $\hat{\overline{X}}_{g,q,2}$ as estimators for M_r and $\overline{X}_g$, respectively. That is,

$$\hat{\overline{X}}_{g,q,2} = \{\hat{M}_1 f_1 + \sum_{i=2}^{r-1} M_i f_i + \hat{M}_{r,q,2} f_r\} \Big/ N .$$

where

$$\hat{M}_{r,q,2} = U_{(r-1)} \ \hat{\alpha}_c / (\hat{\alpha}_c - 1).$$

The method is referred to as the quantile II method, and $\hat{\overline{X}}_{g,q,2}$ is referred to as the Qnt II estimator in this paper. If the lower bounds of the last bounded interval and the open interval are used, then the estimator of α becomes:

$$\hat{\alpha}_o = \ln[(f_{r-1} + f_r)/(f_r)]/\ln[U_{(r-1)}/U_{(r-2)}].$$

This will lead to $\hat{M}_{r,q,1}$ and $\hat{\overline{X}}_{g,q,1}$ as estimators for M_r and $\overline{X}_g$, respectively. The method is referred to as the quantile I method and $\hat{\overline{X}}_{g,q,1}$ is referred to as the Qnt I estimator in this paper.

In the literature the estimator, ,$\hat{\alpha}_o$ seems to be the one most recommended, for example, see Shryock (1975), Parker and Fenwick (1983).

An alternative estimator for α is a modified maximum likelihood estimator developed in West (1985). A brief description of the estimator will be given here. Since the Pareto distribution is considered a good fit for the distribution of higher earnings, the parameter will be estimated from the left truncated distribution. Letting

$$M_s, M_{s+1}, \ldots, M_{r-1}$$

denote the left truncated mid-points of the bounded intervals, then the modified maximum likelihood estimator is:

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

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

Regarding the selection of the truncated point, M_s, it was found in West (1985) that if the earnings distribution was truncated at the mid-point of the interval containing the truncated mean, then the resulting estimate of α led to the estimate of the mean that came the closest to the true mean. The truncated mean is defined as the mean of the data below $U_{(r-1)}$.

Consistency is easily verified for the quantile estimator and it is resistant to outliers. Quandt (1966) found that the performance of the quantile estimators was not much inferior to those of the maximum likelihood estimators. A Monte Carlo study reported by Koutrouvelis (1981) supported that view. However in West (1985), it was shown theoretically and empirically that the quantile method depends heavily on the classification of the population, can lead to gross errors, and at best does as well as the modified maximum likelihood estimator. The data used in the empirical studies were relatively small populations, based on household data. In West (1986), the quantile estimator and the modified maximum likelihood estimator were compared over time on data collected from households in the Current Population Survey (CPS). The results were similar to the ones in the 1985 study. In this paper the estimators will be compared on occupational earnings data collected from establishments. The Winsorized estimator, the two quantile estimators, and the modified maximum likelihood estimators, using four different rules for choosing M_s will be compared. The four rules are to choose M_s to be the:

1. mid-point of the interval that contains the truncated mean.
2 mid-point of the interval following the one that contains the truncated mean.
3. mid-point of the interval that contains the median.
4. mid-point of the interval following the one that contains the median.

The modified maximum likelihood estimator of α will be denoted by $\hat{\alpha}_{m,k}$, and the corresponding mid-point and grouped mean estimators by $\hat{M}_{r,m,k}$ and $\hat{\bar{X}}_{g,m,k}$, respectively, where k=1, 2, 3, or 4, corresponding to the above rule of truncating. The grouped mean estimators are also referred to as Max I, Max II, Max III, or Max IV, corresponding to the rule of truncating.

III. EMPIRICAL STUDIES

In this section, the methods and results of two empirical studies are described. The first study concentrated on the overall industry/occupational wage mean and the second study concentrated on the upper open interval mean. The same measures of evaluation as described below were used for each study.

We define the error of estimation to be the difference between the estimated value and the true value, which is assumed to be known. The relative error is defined as the ratio of the error of estimation to the true value. We compare different estimators by looking at the absolute values of the relative errors of their estimates. The relative errors may be expressed in percentages and be called percent errors. Absolute percent errors are the absolute values of percent errors. For example, if the true mean is 50, the estimated mean is 48, then the relative error is -.04 and the absolute percent error is 4 percent.

Note that for the median, the error in estimation is due to grouping the data; whereas, for the mean, the error is due to grouping and to the upper open interval estimation. Sampling error was not considered in the studies since wage data in exact dollar amounts were used and regarded as the population.

A. OES RESEARCH

This part of the research is called the OES research because it is designed to meet the objective of the Occupational Employment Statistics (OES) survey, that is, to find the best estimators for the median and for the mean of occupational hourly wage for each industry. Estimates, therefore, were computed for each industry/occupation level.

1. Data

The data used for the research are from the 1989 and 1990 White-Collar Pay (WCP) surveys. The WCP survey collects wage data from establishments employing 50 or more workers in industries throughout the United States, except Alaska and Hawaii. The WCP survey collects actual dollar amounts for wages.

We selected nine industries from different major industry groups to study. Each industry is classified by a standard industrial classification (SIC) code. The nine industries and their SIC codes were: oil and gas extraction (SIC 13); food and kindred products (SIC 20); chemicals and allied products (SIC 28); stone, clay, and glass products (SIC 32); transportation by air (SIC 45); miscellaneous retail (SIC 59); security and commodity brokers (SIC 62); hotels and other lodging places (SIC 70); and educational services (SIC 82).

2. Method

We adjusted the collected WCP data by their weights and the weighted data were considered the true population, from which we computed the true mean and median. We then grouped the data into wage interval categories by industry and occupation. For each set of grouped data, we estimated the median by the linear interpolation method and estimated the mean by the methods outlined in Section II. For the modified maximum likelihood method, we only used the interval that contained the truncated mean as the left truncated point. That is, the only estimator examined was $\hat{\bar{X}}_{g,m,1}$ (Max I).

For the research, the 1989 Federal minimum wage of \$3.35 was used as the lower limit of the lower open interval.

3. Selecting the Interval Categories

The pilot OES interval categories specified \$35.00 as the lower limit of the upper open interval. It was found that this figure was too low to provide good estimates. Out of a total of 792 occupations over the nine industries chosen, 81 percent (639) did not have records in the upper open interval. Of the 153 occupations that did have some records in this interval, as many as 53 did not have records in any other interval. For these 53 occupations, there were no data to fit any distribution; the Winsorized mean was the only alternative for computing a mean. For these cases, the Winsorized mean could underestimate the true mean as much as 53 percent. Many other occupations had only a few observations in the previous bounded interval and the rest in the upper open-interval. For these occupations, there were not enough data to fit a Pareto distribution. In order to use the Pareto distribution, the wage distribution should be a decreasing function after a peak and should have a small "tail." The "tail" for our purpose is the upper open-interval itself. Of the 153 occupations that have records in the upper open-interval, only 29 (19%) occupations have "tails" less than ten percent; 8 (.5%) occupations have "tails" of ten to twenty percent; 24 (16%) occupations have "tails" of twenty to fifty percent; and 92 (60%) occupations have "tails" over fifty percent. Attempts to fit a Pareto distribution to these occupations of "large tails" lead to gross errors. In addition, there could be no median estimates for the occupations with "tails" over fifty percent.

Based on the above observations and on the distribution of the true mean of the upper open interval over industries, a modified version for the interval categories was proposed and accepted for future OES surveys. The modified interval categories increased the lower limit of the upper open interval from \$35 to \$60 and widened the middle ranges, while leaving the two lowest intervals the same as before. The number of interval categories was kept at eleven for administrative purposes. Now more

Pareto distributions could be fitted, allowing different methods for computing the mean for the upper open interval. The pilot versus the future interval categories are given below.

Interval Category	Pilot	Future
A	<5	<5
B	[5-6.5)	[5-6.5)
C	[6.5-8)	[6.5-9)
D	[8-10)	[9-12)
E	[10-12)	[12-16)
F	[12-14)	[16-20)
G	[14-17)	[20-25)
H	[17-21)	[25-35)
I	[21-25)	[35-45)
J	[25-35]	[45-60)
K	>35	≥ 60

4. Results

The modified interval categories gave us 23 occupations with observations in the upper open interval. Of these, twelve had "tails" between ten and fifty percent, and six had "tails" of fifty percent or more. There were no median estimates for these six occupations. Four of these six occupations only had observations in the upper open interval and did not have alternative measures of the mean besides the Winsorized method.

The estimates for overall means improved significantly from the pilot interval categories as the following tables show. We were able to estimate more medians using the new interval categories.

An error profile for all occupations using pilot interval categories is displayed in Table I. The numbers in the body of the table denote the number of times a specific method resulted in errors that fell in a specified error range. The error ranges are absolute percent errors (percent errors in absolute values).

A similar error profile using the future interval categories are displayed in Table II.

TABLE I

Error profile for all occupations using pilot interval categories

Error Range	Qnt I	Qnt II	Max I	Winsorized	Median
0 - 4.99	591	596	608	602	570
5 - 9.99	84	81	83	108	110
10 -14.99	21	24	20	24	18
≥15.00	43	38	28	58	3
N/A*	53	53	53	0	91

TABLE II

Error profile for all occupations using future interval categories

Error Range	Qnt I	Qnt II	Max I	Winsorized	Median
0 - 4.99	626	626	634	631	566
5 - 9.99	112	112	113	115	156
10 -14.99	41	41	38	40	56
≥15.00	9	9	3	6	8
N/A*	4	4	4	0	6

* Estimates could not be computed.

The quantile I and quantile II methods performed similarly. These estimates were not as good as the maximum likelihood or the Winsorized estimates. They had more errors in the "≥15%" range. Additionally, the errors by the quantile methods in this range were extremely high. The Winsorized estimate had more errors in the "≥15%" range than the maximum likelihood estimate, however most of these errors resulted in the cases in which the maximum likelihood estimator could not be computed. We recommended the Winsorized mean for the OES survey since it is easy to understand and to implement.

The median also performed well. Under the pilot interval categories, the median for 91 occupations could not be computed compared to six occupations under the future interval categories.

Empirical results also suggested that the absolute percent error tends to be high when the number of unweighted workers in an occupation is small. For the median, 19 percent of the occupations that have less than ten unweighted workers have percent errors of the magnitude ten percent or higher compared to three percent of the occupations that have at least ten unweighted workers. For the mean, the comparison is 15 percent versus .8 percent. Based on these results, we recommended that publishability criteria include the provision that all published figures come from occupations with at least ten unweighted workers.

5. Validation of other data sets

The White Collar Pay surveys are not completely representative of the OES survey. The survey does not cover Hawaii and Alaska and does not include small establishments of less than 50 workers. It does not target production workers and therefore its wage distribution is not the same as the wage distribution of the OES survey which includes all occupations.

However, we feel that the mean and median estimators are robust and should work as well on OES data. We validated the recommended estimators on two additional sources of wage data: the Alaska data and the Industry Wage Survey data.

In the State of Alaska a Wage Rate survey was conducted. The hourly wage data from this survey was sent to the Bureau of Labor Statistics to be tested with the recommended procedures. The staff in Alaska indicated that the following industries should be tested: metal mining (SIC 10); general building contractors (SIC 15); food and kindred products (SIC 20); depository institutions (SIC 60); hotels and other lodging places (SIC 70); and engineering and management services (SIC 87). The minimum wage of $4.75 was given to us by the staff in Alaska and was used in the research.

Since we recommended that all publishable means and medians have at least ten unweighted workers, we only considered the percent errors of the occupations with at least ten workers. Out of 89 occupations, five (6%) had absolute percent errors for the median exceeding ten percent and two (2%) had absolute percent errors for the mean exceeding ten percent. These numbers are thought to be small enough to be acceptable.

The Industry Wage Survey (IWS) collects wage data for production workers only. Wage data from this survey do not have as many observations in the upper interval categories. In order to see the effect of the future interval categories on data sets with lower wages, we tested the procedures on two IWS data sets using both sets of interval categories. The data sets available were the 1987 Men's and Boys' Shirts (SIC 2321) and the 1984 Millwork (SIC 2431). The Federal minimum wage of $3.35 at that time was used as the lower limit of the lower open interval.

There were 31 occupations in the Millwork data and 30 occupations in the Shirts data to be tested. For each data set, the two sets of interval categories basically gave the same error distribution.

6. Compare Percent Change Across Time or SICs

Of interest to the OES survey is the comparison of occupational hourly wages across time. As mentioned in Section II, the median or functions of median should not be used for these purposes. The difference in means of a characteristic of a population was found to be bounded relative to the means of its subpopulations when compared across time periods, provided the proportional size of the subpopulations tend to stay the same over time.

Also of interest to the OES survey is the comparison of occupational hourly wages across populations and across their subpopulations. The population could be a major occupational group such as "Engineers." The subpopulations could be different pay levels (which reflect the different expertise and experience levels within one major occupational group) or different detailed occupations (which reflect the many different but related jobs, for example, different kinds of engineers). The size of the subpopulations relative to the size of their parent population may not be the same in different industries.

We looked at the "bounded property" of true and estimated means and of true and estimated medians in the case of comparing different occupational pay levels. We found that true values or their estimates may or may not be bounded. Furthermore, the existence or lack of "bounded property" of true values does not carry over to their estimates. That is, bounded true means or bounded medians do not lead to bounded estimated means or bounded estimated medians, and vice versa. For these comparisons, the median as well as the mean should be used with caution.

B. INTERVAL MEAN RESEARCH

This part of the research expands on the OES research by concentrating on estimating the mean of the upper open interval.

1. Data

Data from the same 1989 and 1990 White-Collar Pay surveys were used. The following nine industries were added: coal mining (SIC 12), special trade contractors (SIC 17), apparel and other textile products (SIC 23), electronic and other electric equipment (SIC 36), communications (SIC 48), wholesale trade (SIC 50), automotive dealers and service stations (SIC 55), real estate (SIC 65), and legal services (SIC 81). These additional industries allowed additional coverage of industry groups.

2. Method

As discussed in the OES research, the Pareto distribution is not always suitable for industry/occupational wage data. With occupational data by industry, using either set of interval categories, there is a high percentage of occupations with large frequency in the upper open interval ("large tails" for Pareto distributions). In order to conduct research on estimating the mean for the upper open interval, we need data that are more suitable to fitting the Pareto distribution. To bring smaller "tails," we grouped all the industrial occupations into professional, technical, and clerical occupational types as described in the White-Collar Pay survey bulletin published by the Bureau of Labor Statistics. We estimated the median and the mean by industry/occupational type with the methods mentioned in Section II. When using the maximum likelihood method on these data sets, the four rules for choosing the left truncated point were applied. Since there were 18 industries, each with three different occupational types, a total of 54 populations were considered.

3. Selecting the Interval Categories

When we grouped the occupational type data according to the future OES interval categories, there were not enough data in the upper open interval (in percentages) to do meaningful research. Based on the frequency distribution of workers in each occupational type, we decided to use the pilot OES interval categories with different lower bounds for the upper open intervals: $35 for the professional type data, $21 for the technical type, and $17 for the clerical type. These lower bounds were not chosen arbitrarily. They were lower bounds of the pilot OES intervals, but not necessarily of the upper open interval. These lower bounds were chosen with the aim for "tails" of less than ten percent. It was found from the research that the Pareto distribution gives best estimates with "tails" of this size. This gave the percentage of workers in the upper open interval from .4 to 6.5 percent in the professional occupations (with one exception of 18 percent), from .05 to 7.6 percent in the technical occupations, and from .04 to 3.4 percent in the clerical occupations.

4. Results

The estimated median and estimated truncated mean usually fell in the same interval. Therefore, the maximum likelihood estimator derived from using either the interval containing the median or the interval containing the truncated mean as the left truncated point usually was the same. For this reason, we will only discuss the left truncated point as in the interval containing the estimated truncated mean (Max I) or in the interval following the one that contains the estimated truncated mean (Max II). For the professional occupations, the quantile II method performed well, followed by the maximum II method. For the technical occupations, the Winsorized performed well, followed by the quantile II and maximum II methods. For the clerical occupations, the maximum II method is the best. Although the Max II estimate is not always the best over the occupational types, it is consistently one of the best.

A third of the 54 occupational type populations did not have records in the upper open interval. An error profile for the two thirds that had upper open interval estimates is shown in Table III.

TABLE III

Error Profile for Estimating the Mean of the Upper open Interval

Error Range	Qnt I	Qnt II	Max I	Max II	Winsorized
0-4.99	17	23	14	18	4
5-9.99	12	7	12	15	10
10-14.99	3	5	5	2	13
≥ 15	4	1	5	1	9

As the table indicates, the Winsorized was not the best method. The quantile II method could be good, but had more large errors than the maximum II method. The quantile I, maximum I, and Winsorized methods were not recommended. They seemed to be sensitive to different data sets and gave more errors in the "≥ 15%" range. The Max II is the one recommended.

5. Effects of Grouping Data

The quantile II method uses the two bounded intervals preceding the upper open interval in estimation. When the boundary point of these two closed intervals was changed from $25 to $28 for the professional type data, the quantile II estimator did considerably worse. The other estimators were not affected much by the change. This is in agreement with West's findings (1986) that the quantile method is sensitive to the way the data are grouped.

When we lowered the lower bound of the upper open interval of the technical type data from $21 to $17, allowing larger "tails," there was a uniform decrease in the percent errors in the "0-4.99%" range and an increase in the higher percent errors across industries. The quantile II method was the best this time.

The quantile I method was affected most when Pareto distributions were fitted to "large tails." For example, when data of the gas and oil industry were grouped according to the pilot OES interval categories, the largest absolute percent error was over 400 percent for the quantile I method and just over 100 percent for the quantile II and for the maximum I methods. Most of these occupations have their truncated means fall in the bounded interval preceding the upper open interval and could not have the Max II estimates computed.

IV. CONCLUSIONS

When data are grouped such that there are no suitable "tails" to the Pareto distribution, such as when the SIC/occupation wage data were grouped according to the OES pilot interval categories, errors of the overall mean and of the upper open interval were large. In particular, the maximum likelihood and quantile methods could lead to gross errors. Furthermore, these estimates could not be computed when the upper open interval was the lone interval category. In such cases, the Winsorized estimate may be the best (although technically, all that can be said is that the mean is $\geq U_{(r-1)}$). In our research data, while the other estimates could have absolute percent errors in the hundreds, the highest error for the Winsorized estimate was just over 50 percent with the pilot OES interval categories.

When the objective is to compute an overall mean, then all estimators produce large errors if there are "large" (10% or more) tails, and all estimators perform satisfactorily if there are "small" (less than 10%) tails. When "tails" are large, the Pareto distribution is not suitable for estimation purposes. When "tails" are small, the overall error is not affected much by large errors in the small "tails." This was the case when the wage data were grouped according to occupational types: professional, technical, or clerical. Given the nature of the occupational data by industry and the OES's objective of computing an overall occupational wage for each industry, the Winsorized estimator was the one recommended with the modified interval categories.

For the upper open interval, the Winsorized estimate is biased downward. Depending on the lower limit of the upper open interval, this bias could be large or small. The advantage of this method is that it is simple and the direction of its bias is known. For "small-tailed" distributions, the quantile II method performed well most of the time. However, when it did not perform well, it had much larger errors than the maximum likelihood method. Although the maximum II estimator is not always the best over the occupational types, it is consistently one of the best. This estimator is recommended because it is more robust than either the quantile I or the quantile II method. The Winsorized estimator does not perform well for upper open interval estimation in "small-tailed" distributions.

V. FUTURE RESEARCH

Variance estimators for the mean, using the Pareto tail with the modified maximum likelihood estimator for the parameter, will be derived and evaluated in the next study. Also, in this study it had been planned to fit theoretical distributions (mixture distributions) to the occupational wage data. Unfortunately the data were too thin to accomplish this. We hope a larger appropriate data set will be available for future research.

REFERENCES

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 Distribution 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.

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

West, Sandra A. (1986), "Measures of Central Tendency for Censored Earnings Data from the Current Population Survey," *Proceedings of the Section on Business and Economic Statistics*, American Statistical Association, 751-756.

West, Sandra A. (1985), "Estimation of the Mean From Censored Earnings Data," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 665-670.