EVALUATION OF CONFIDENCE INTERVAL METHODOLOGY FOR THE OCCUPATIONAL COMPENSATION SURVEY PROGRAM

Christina L. Harpenau, Joan L. Coleman, Mark D. Lincoln
Christina L. Harpenau, 2 Mass. Ave., N.E., Room 3160, Washington, D.C. 20212

Key Words: Collapsing Strata, Quantiles, Relative Median Length

Introduction

A theoretical investigation and two simulation studies of the Occupational Compensation Survey Program (OCSP) data, presented by Casady, Dorfman, and Wang ("CDW"-1994) suggested that the standard 95% confidence intervals (C.I.) for domain means or totals, when based on the standard normal distribution and standard methods of variance estimation, tend to yield less than the actual 95% coverage. Even though the sample size is large enough to support standard normal estimations, the individual occupations are represented by a small number of establishments. "CDW" presented new nonstandard methods that offer an improvement, giving intervals with more accurate coverage, typically at or close to the nominal 95% coverage. These intervals tend to be longer than the standard intervals and depend mainly on the use of t-statistic having degrees of freedom dependent on the available domain data. The increase in length will vary with domain, and will depend on the particular method for C.I. construction that is used.

A related concern is the degree and type of collapsing of strata that should be used in the estimation of variances and the degrees of freedom for the purpose of confidence interval construction. In general, there will be a tradeoff: as strata are reduced in number, the estimate of variance will tend to increase, but so will the degrees of freedom.

Universe Development

A study was undertaken to evaluate the proposed methodologies. Thirty-two primary metropolitan statistical areas were classified into three categories (large, medium, small) based on the total number of workers in each area. The median area from the large category and the medium category were selected to serve as a standard for the "typical" Metropolitan Statistical Area (MSA) for the two categories. Two artificial populations were constructed to have the properties of the median area in terms of size and number of establishments using available sample data from the OCSP. The "large" MSA population was constructed by randomly selecting establishments from the samples of all areas in the large category and allowing each area to have equal representation within each stratum. Establishments are assigned to strata based on Standard Industrial Classification (SIC) and employment size. When necessary, establishments were taken from either the medium or small category to obtain enough establishments in each stratum to equal the stratum size of the "large" MSA. Similar procedures were completed for the "medium" MSA.

Sample Development

Once the universes were constructed, multiple samples were randomly selected based on the median area's sample size. For this study, twenty samples were drawn from both the large and medium universes. For each sample, establishments were randomly selected from the universe with equal probability within each stratum. Each stratum sample size corresponds to the actual median area's stratum sample size. Weights were assigned to sample establishments using the current OCSP weighting procedures. Weights were assigned to each sample member such that the total weight for each stratum equaled the universe stratum size. Since the universe was generated from previously collected data, the collection status (collected, refusal, etc.) was retained for each sample unit and nonresponse adjustment was then completed as follows: (1) refusals were given a weight of zero and the weighted employment was distributed onto other establishment(s); and (2) out of business and out of scope establishment(s) were given a weight of zero.

Collapsing Strata

Since the contribution to the variance from strata with only one usable sample establishment cannot be estimated, the collapsing of strata is necessary. Three collapse patterns were considered for this empirical study. The first collapse pattern was based on size class within an SIC division, which is the current method used in the OCSP. For example, stratum '203', where '20' represents the SIC and '3' denotes the size class, would be collapsed into stratum '204'. The

second collapse pattern was based on SIC division within a size class. For example, stratum '203' would be collapsed into stratum '213'. The final collapse pattern was based on maximal collapsing of all SICs and size classes within a major industry level; i.e., there would be one stratum per State and Local Government, Goods Producing, Manufacturing, Service Producing, and Transportation and Utilities. Each of the three collapse patterns described above were performed on each of the large and medium samples.

Variance Estimation

Within the large and medium samples, the variance for average earnings and total workers were computed for each of the three collapse patterns across all samples. The methodology for the OCSP variance components of estimation is as follows:

Variance formula for Average Occupational Earnings denoted by

$$\hat{V}_{a_b}\left(\hat{\bar{y}}_j\right)=\frac{1}{\hat{x}_j^2}\sum_{h=1}^{H}\left[\left(\frac{n'_h}{n'_h-1}\right)*\left(1-\frac{n'_h}{N_h}\right)\sum_{i=1}^{n'_h}\left[\left(y_{hij}-\hat{\bar{y}}_j*x_{hij}\right)-\left(\frac{\hat{y}_{hj}}{n'_h}-\hat{\bar{y}}_j*\frac{\hat{x}_{hj}}{n'_h}\right)\right]^2\right]$$

and the variance formula for Total Occupational workers is

$$\hat{V}_{a_b}\left(\hat{x}_j\right)=\sum_{h=1}^{H}\left(\frac{N_h^2}{n'_h\left(n'_h-1\right)}\right)*\left(1-\frac{n'_h}{N_h}\right)\left[\sum_{i=1}^{n'_h}x_{hij}^2-\frac{\left(\sum_{i=1}^{n'_h}x_{hij}\right)^2}{n'_h}\right]$$

, where

$\hat{V}_{a_b}\left(\hat{\bar{y}}_j\right)$ = variance of average earnings for occupation j, collapse pattern a (1..3), and sample b (1..m) where m is the number of samples

$\hat{V}_{a_b}\left(\hat{x}_j\right)$ = variance of total workers for occupation j collapse pattern a (1..3), and sample b (1..m) where m is the number of samples

$\hat{x}_j$ = total weighted employment of occupation j (total over all strata)

n'_h = number of usable establishments for stratum h

N_h = number of establishments in the universe for stratum h

y_{hij} = weighted aggregate earnings for occupation j in establishment i of stratum h

$\hat{\bar{y}}_j$ = estimated average earning of occupation j

x_{hij} = number of weighted workers in stratum h with occupation j in establishment i

$\hat{y}_{hj}$ = aggregate earnings for occupation j in stratum h

$$=\sum_{i=1}^{n'_h}y_{hij}$$

$\hat{x}_{hj}$ = total number of workers in stratum h with occupation j

$$=\sum_{i=1}^{n'_h}x_{hij}$$

Once variances across all samples and collapse patterns were produced, the next step was to compute the mean and median of the variance estimates for each variable $\left(\hat{\bar{y}}_j \text{ and } \hat{x}_j\right)$. The mean of the variance for average earnings was calculated by:

$$\hat{\bar{V}}_a\left(\hat{\bar{y}}_j\right)=\sum_{b=1}^{m}\frac{\hat{V}_{a_b}\left(\hat{\bar{y}}_j\right)}{c}$$

where c= the number of samples with occupation j and m is the number of sample drawn.

The mean of the variance for total workers was calculated by:

$$\hat{\bar{V}}_a\left(\hat{x}_j\right)=\sum_{b=1}^{m}\frac{\hat{V}_{a_b}\left(\hat{x}_j\right)}{m}$$

for each occupation j and each collapse pattern a for each sample m.

The median variance of each occupation and collapse pattern was generated based on the 50th percentile of the variances produced for each of the samples. This will be denoted as $\hat{V}_a\left(\hat{\bar{y}}_j\right)_{med}$ and $\hat{V}_a\left(\hat{x}_j\right)_{med}$.

Accuracy of Variance Estimator

In addition to the proposed research outlined in the introduction, the next part of our empirical study was completed in order to show the accuracy of our current variance estimator. We began by computing the true total workers, (X_j), by summing the workers across the entire universe for occupation j. The true mean earnings, $(\overline{Y}_j)$, are obtained by summing the total earnings across the entire universe divided by the number of workers in the universe with occupation j. The true total workers and the true mean earnings were calculated based on usable establishments in the universe. Nonresponse procedures were not performed on the universes as was done for each of the samples.

The true variance for occupation j across all samples within each collapse pattern was estimated using the following formula:

The estimated true variance for mean earnings is,

$$\hat{V}_{true}\left(\hat{\bar{y}}_j\right) = \sum_{b=1}^{m} (\hat{\bar{y}}_{bj} - \overline{Y}_j)^2 / c,$$

where c= number of samples with occupation j.

The estimated true variance for total workers is,

$$\hat{V}_{true}\left(\hat{x}_j\right) = \sum_{b=1}^{m} (\hat{x}_{bj} - X_j)^2 / m.$$

Next, we used each variable and collapsing pattern and computed the ratio of the mean and median to the estimated true variance such that,

$$R_a\left(\hat{\overline{V}}\right) = \frac{\hat{\overline{V}}_a\left(\hat{\bar{y}}_j\right)}{\hat{V}_{true}\left(\overline{Y}_j\right)} \text{ and } R_a\left(\hat{V}_{med}\right) = \frac{\hat{V}_a\left(\hat{\bar{y}}_j\right)_{med}}{\hat{V}_{true}\left(\overline{Y}_j\right)}$$

where a represents the collapse pattern. The same ratios were computed for the mean and median variance values of the total workers.

In order to compare the above ratios, geometric means were calculated across all occupations within each collapse pattern (size, SIC, major) and size category (large, medium). The geometric means for the ratio of the mean variance, $R_a\left(\hat{\overline{V}}\right)$ and the ratio of the median variance, $R_a\left(\hat{V}_{med}\right)$ were calculated using the following formula:

$M_{g_a} = \sqrt[d]{\prod_j^d \left(R_{a_j}\left(\hat{\overline{V}}\right)\right)}$, where d represents the total number of occupations j in collapse pattern a.

For this empirical study, different confidence interval (C.I.) methods were considered and will be presented in more detail in another section. For one C.I. method, degrees of freedom were calculated and, for consistency purposes, this study has only included occupations with degrees of freedom greater than zero. Since each collapse pattern could result in different degrees of freedom values for each occupation, the value d above could differ between collapse patterns.

To reduce the impact on the geometric mean, for occupations with very small or large ratios, limitations were set, i.e., if the ratio was less than .25 the ratio was set at .25 and if the ratio exceeded 4, the ratio was set at 4. Below are the results across the three collapse patterns and two size classes:

A-1. GEOMETRIC MEANS OF THE RATIOS FOR THE LARGE CATEGORY

	Mean Variance Earnings	Mean Variance Workers	Median Variance Earnings	Median Variance Workers
SIC Collapse	0.627	0.815	0.500	0.485
Size Collapse	0.625	0.800	0.500	0.484
Major Collapse	0.991	1.153	0.851	0.809

A-2. GEOMETRIC MEANS OF THE RATIOS FOR THE MEDIUM CATEGORY

	Mean Variance Earnings	Mean Variance Workers	Median Variance Earnings	Median Variance Workers
SIC Collapse	0.730	0.886	0.590	0.688
Size Collapse	0.733	0.865	0.594	0.669
Major Collapse	0.973	1.247	0.826	1.037

The geometric means for the SIC collapse and size collapse patterns are approximately equal to each other. Based on the geometric mean of the size collapse pattern above (charts A-1 and A-2), this indicates that current OCSP procedures may slightly underestimate the variance.

Confidence Interval

For each of the samples and their three collapse patterns, two methods were applied to generate 95% confidence intervals (C.I.s) for each occupation j. The first method produced the 95% C.I. using the standard normal quantile, such that

$C.I._{SN} = \text{estimate} \pm (1.96 * \text{standard deviation})$, where the standard deviation is estimated from the particular sample. The second method generated the 95% C.I. using unweighted degrees of freedom (d.f.) as defined in "CDW". The student's t distribution was applied based on the unweighted degrees of freedom, using the following formula:

$C.I._{St} = \text{estimate} \pm \left(t_{(d.f.,0.025)} * \text{standard deviation}\right)$. The unweighted d.f. for each occupation were calculated by $\sum_{h=1}^{H}\left[\max\left(n_{h_j} - 1, 0\right)\right]$, i.e., the total number of establishments with occupation j in stratum h minus 1, summed across all strata. For each sample, occupations with d.f.=0 were not included in the C.I. analysis.

In the next stage, we determined the proportion of C.I.s, which contained the true universe values for both variables, $\overline{Y}_j$ and X_j. For this calculation, we determined how many C.I.s, within each collapse pattern, contained the true values over the total number of samples containing the particular occupation. We performed this procedure for both confidence interval methods. The following distributions (charts B-1 through B-8) represent the proportion of occupations in the respective coverage levels:

B-1. PERCENTAGE OF PROPORTIONS ACROSS ALL OCCUPATIONS FOR EARNINGS STANDARD NORMAL C.I. FOR THE LARGE CATEGORY

	SIC Collapse	Size Collapse	Major Collapse
Less than 25%	5.10%	5.00%	1.60%
25% to less than 50%	5.90%	5.00%	4.70%
50% to less than 75%	13.60%	14.20%	10.10%
75% to less than 95%	46.60%	50.80%	32.60%
95% or more	28.80%	25.00%	51.20%

B-2. PERCENTAGE OF PROPORTIONS ACROSS ALL OCCUPATION FOR EARNINGS UNWEIGHTED DEGREES OF FREEDOM C.I. FOR THE LARGE CATEGORY

	SIC Collapse	Size Collapse	Major Collapse
Less than 25%	0.80%	0.80%	0.00%
25% to less than 50%	1.70%	2.50%	2.30%
50% to less than 75%	4.20%	2.50%	2.30%
75% to less than 95%	32.20%	31.70%	27.10%
95% or more	61.00%	62.50%	68.20%

B-3. PERCENTAGE OF PROPORTIONS ACROSS ALL OCCUPATIONS FOR WORKERS STANDARD NORMAL C.I. FOR THE LARGE CATEGORY

	SIC Collapse	Size Collapse	Major Collapse
Less than 25%	7.60%	8.30%	3.10%
25% to less than 50%	7.60%	7.50%	3.80%
50% to less than 75%	11.90%	10.80%	9.90%
75% to less than 95%	35.60%	35.00%	20.60%
95% or more	37.30%	38.30%	62.60%

B-4. PERCENTAGE OF PROPORTIONS ACROSS ALL OCCUPATIONS FOR WORKERS UNWEIGHTED DEGREES OF FREEDOM C.I. FOR THE LARGE CATEGORY

	SIC Collapse	Size Collapse	Major Collapse
Less than 25%	2.50%	2.50%	2.30%
25% to less than 50%	5.10%	6.70%	3.80%
50% to less than 75%	11.00%	8.30%	7.60%
75% to less than 95%	21.20%	20.80%	12.20%
95% or more	60.20%	61.70%	74.00%

B-5. PERCENTAGE OF PROPORTIONS ACROSS ALL OCCUPATIONS FOR EARNINGS STANDARD NORMAL C.I. FOR THE MEDIUM CATEGORY

	SIC Collapse	Size Collapse	Major Collapse
Less than 25%	0.90%	1.80%	0.00%
25% to less than 50%	4.50%	3.60%	0.80%
50% to less than 75%	12.60%	11.80%	7.40%
75% to less than 95%	49.50%	49.10%	41.00%
95% or more	32.40%	33.60%	50.80%

B-6. PERCENTAGE OF PROPORTIONS ACROSS ALL OCCUPATION FOR EARNINGS UNWEIGHTED DEGREES OF FREEDOM C.I. FOR THE MEDIUM CATEGORY

	SIC Collapse	Size Collapse	Major Collapse
Less than 25%	0.00%	0.00%	0.00%
25% to less than 50%	0.90%	0.00%	0.00%
50% to less than 75%	0.90%	1.80%	0.80%
75% to less than 95%	14.40%	15.50%	24.60%
95% or more	83.80%	82.70%	74.60%

B-7. PERCENTAGE OF PROPORTIONS ACROSS ALL OCCUPATIONS FOR WORKERS STANDARD NORMAL C.I. FOR THE MEDIUM CATEGORY

	SIC Collapse	Size Collapse	Major Collapse
Less than 25%	0.00%	0.00%	0.00%
25% to less than 50%	0.90%	0.00%	0.00%
50% to less than 75%	9.00%	10.90%	4.10%
75% to less than 95%	29.70%	33.60%	17.10%
95% or more	60.40%	55.50%	78.90%

B-8. PERCENTAGE OF PROPORTIONS ACROSS ALL OCCUPATIONS FOR WORKERS UNWEIGHTED DEGREES OF FREEDOM C.I. FOR THE MEDIUM CATEGORY

	SIC Collapse	Size Collapse	Major Collapse
Less than 25%	0.00%	0.00%	0.00%
25% to less than 50%	0.00%	0.00%	0.00%
50% to less than 75%	4.50%	4.50%	1.60%
75% to less than 95%	10.80%	11.80%	8.10%
95% or more	84.70%	83.60%	90.20%

Also, proportions across all occupations in each collapse pattern were produced and are as follows:

C-1. PERCENTAGE OF C.I.s CONTAINING TRUE VALUES FOR THE LARGE CATEGORY

	Standard Normal C.I.s for Earnings	Unwgted D.F. C.I. for Earnings	Standard Normal C.I.s for Workers	Unwgted D.F. C.I. for Workers
SIC Collapse	79.28%	90.16%	77.31%	86.00%
Size Collapse	79.26%	90.50%	77.20%	86.51%
Major Collapse	87.26%	92.67%	89.7%	89.70%

C-2. PERCENTAGE OF C.I.s CONTAINING TRUE VALUES FOR THE MEDIUM CATEGORY

	Standard Normal C.I.s for Earnings	Unwgted D.F. C.I. for Earnings	Standard Normal C.I.s for Workers	Unwgted D.F. C.I. for Workers
SIC Collapse	83.36%	95.59%	90.21%	95.13%
Size Collapse	86.21%	95.52%	89.59%	95.02%
Major Collapse	90.34%	94.71%	95.32%	97.16%

As shown in charts C-1 and C-2, the distributions of the proportions of coverage tend to be comparable for the SIC collapse and size collapse patterns. Also, it appears that the unweighted degrees of freedom confidence intervals produce, on average, closer to the desired 95% coverage. When comparing the percentages above to those on the previous pages, for the collapse patterns and C.I. methods that provide less than 95% coverage, it is mainly due to those occupations with proportion that are less than 50% (see charts B-1 through B-8).

For each category (large and medium), collapse pattern (SIC, size, and major) and variable (earnings and workers), the relative median length of the confidence intervals (standard normal and unweighted degrees of freedom) were computed for all samples. The relative median length (RML) was computed as:

$$\frac{\text{C.I. median length over all samples}}{3.92 * \text{estimated true standard deviation}}.$$

In addition, the geometric mean of all RML values within each collapse pattern was produced as follows: $M_{g_a} = \sqrt[d]{\prod_1^d (RML_{a_j})}$, where d represents the total number of occupations j in collapse pattern a. Only occupations with d.f.>0 were studied; thus, since each collapse pattern could result in different degrees of freedom values for each occupation, the value d above differs between collapse patterns.

The results of the geometric means are listed below in charts D-1 and D-2.

D-1. GEOMETRIC MEANS OF THE RELATIVE MEDIAN LENGTH FOR THE LARGE CATEGORY

	Standard Normal C.I.s for Earnings	Unwgted D.F. C.I. for Earnings	Standard Normal C.I.s for Workers	Unwgted D.F. C.I. for Workers
SIC Collapse	0.676	1.265	0.625	1.187
Size Collapse	0.677	1.274	0.620	1.154
Major Collapse	0.901	1.207	0.906	1.253

D-2. GEOMETRIC MEANS OF THE RELATIVE MEDIAN LENGTH FOR THE MEDIUM CATEGORY

	Standard Normal C.I.s for Earnings	Unwgted D.F. C.I. for Earnings	Standard Normal C.I.s for Workers	Unwgzted D.F. C.I. for Workers
SIC Collapse	0.800	1.939	0.812	1.940
Size Collapse	0.801	1.764	0.802	1.784
Major Collapse	0.801	1.327	1.032	1.542

Once again, the geometric means for the SIC collapse and size collapse patterns are about equal to each other. The geometric means appear to be low for the standard normal and high for the unweighted degrees of freedom.

Conclusion and Future Studies

From our empirical investigation on OCSP data we draw the following conclusions:

Standard 95% confidence intervals for domain means or totals, when based on the standard normal distribution and standard methods of variance estimation yield less than the actual 95% coverage.

Confidence Intervals using unweighted degrees of freedom, produce intervals with better coverage closer to the nominal 95% coverage. The intervals tend to be longer than the standard normal intervals. The increase in length will vary with occupation.

The principal effect of this research shows the loss of coverage, for purposes of C.I. construction, of the standard normal quantiles (±1.96 for 95% coverage). These are replaced by quantiles for the Student's t-distribution, with degrees of freedom determined from the sample and varying with occupation.

In the future for each variance estimate, we may compute a 95% confidence interval using weighted degrees of freedom, as proposed by "CDW". This, we expect, will yield coverage a few points higher than the unweighted degrees of freedom. Also a study consisting of a larger amount of replicated samples may be conducted to verify the results gained from our smaller study.

Any opinions expressed in this paper are those of the authors and do not constitute policy of the Bureau of Labor Statistics.