EVALUATION OF CONFIDENCE INTERVAL METHODOLOGY FOR THE NATIONAL COMPENSATION SURVEY

Glenn Springer, Martha Walker, Steven Paben, Alan Dorfman
2 Massachusetts Ave., N.E., Room 3160, Washington, D.C. 20212

Key Words: Artificial MSA, Establishment Sample, Occupational Sample, Quantiles

Introduction

The National Compensation Survey (NCS) is a Bureau of Labor Statistics (BLS) program that provides data on occupational wages. An investigation by Casady, Dorfman, and Wang 1996 (CDW) 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. The estimation of means or totals within an occupation is a case of domain estimation presented in Cochran (1977, pg. 34) since the observations in the sample falling within a specified occupation are not known prior to sampling. Even though the sample size is large enough to support standard normal estimations, the individual occupations can be 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 on the use of a 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. In Harpenau, Coleman, Lincoln (HCL 1995) this was shown to be true for data from the Occupational Compensation Survey Program (OCSP). We modified this method to make it suitable to the multi-stage design of the NCS. Using NCS data, an artificial sampling frame was created and simulated samples were selected. The standard normal confidence intervals were compared to confidence intervals using the t-distribution with weighted degrees of freedom for estimates of means and quantiles. Coverage properties for confidence intervals using the non-standard approach were found to be superior to the standard normal approach.

Universe Development

A study was undertaken to evaluate the proposed methodologies. NCS production data was used in constructing the artificial Metropolitan Statistical Area (MSA). At the time of this study, 16 NCS surveys had been completed. We created a medium MSA "universe" of establishments and workers using all occupational wage records from the available medium and large MSA surveys. The following steps were carried out within each major industrial stratum by size class using data from the sixteen NCS surveys.

In order to determine the "typical" distribution of establishments by size class within industry for a medium sized MSA, we computed the mean number of establishments across areas. In calculating this average, we excluded the two largest areas, Los Angeles and New York, and the two smallest areas, Huntsville and Dayton from the 16 NCS surveys because we did not want them to influence the determination of the typical distribution of the medium size MSA.

After deleting nonrespondent establishments, we combined the establishment data from the 16 available MSA's and sorted the establishments by size class (the number of employees in the establishment) within industry cell. Based on the distribution of establishments by size class within industry in "medium sized" MSAs, we determined the appropriate employment sizes for the artificial MSA. Using the number of establishments in each industry size class cell for the typical medium size MSA, we computed the required number of employees for each cell by multiplying the number of establishments in the cell by the median number of in-scope employees in the cell for the 12 medium size MSAs. This "employment" is the reported (total) employment, excluding out of scope workers, e.g. contractors, individuals who participate in setting their own pay, student employees, and volunteers.

Because the distribution of the combined data file does not provide a sufficient number of workers in the small size classes we had to borrow workers from the larger size establishments. This was done by

splitting the larger establishments to form pseudo establishments, which were then used to provide the additional workers needed for the smaller size classes. When the number of employees in occupations for the larger establishments exceeded the number needed for the smaller establishments, we split the heavily populated occupations (larger occupations) into two occupational groups, for the contribution to a pseudo-establishment one size-class lower. Where the occupation contributed to pseudo-establishments two size-classes lower the occupation was split into 4 occupational groups, three size classes lower split the occupational group by 8. This guarded against giving a pseudo-establishment of a given size class an occupation in which there were more employees than there should be in the pseudo-establishment. In doing this splitting, the original occ was divided by randomly selecting half of its employees in order to form "sub-occ's". Once the pseudo-establishments were selected, the following steps were performed within each of the pseudo-establishments.

The first step was to ensure that the distribution of employees by occupation was typical for establishments of that size class. As mentioned earlier, we omitted non-respondent establishments and by implication, omitted non-respondent occupations in creating the universe. Also because the selection of occupations in NCS is done using probability proportional to size based on the number of workers in the occupation, our artificial MSA contained an abundance of occupations with a relatively large number of employees and a shortage of occupations with small employment. To adjust for this we "fractured" (split) occupations with a large number of employees to form occupations with a smaller number of employees.

The second step was the reassignment of major occupational group and level (MOGL) where necessary, because the distribution of workers in the artificial MSA did not reflect the distribution of workers by MOGL in the aggregate file of 16 MSA's. To reassign MOGLs we computed an ideal fraction which was equal to the total employment within a MOGL for the 16 MSA's divided by the total employees in the 16 MSA's. The ideal employment for the artificial MSA in each MOGL was calculated by multiplying the ideal fraction times the total employment in the artificial MSA. This resulted in the desired distribution of employees by MOGL in the artificial MSA. The resulting MSA had 2,880 establishments and 53,617 occupations consisting of 629,039 employees.

Sample Simulation Methodology

We selected 100 establishment samples from the artificial MSA. After selecting the establishment samples we then selected a PPS sample of occupations within each establishment. The procedure used to select the establishment sample and the occupation sample mimicked the approach used in NCS as described below.

Using the sample size and industry distribution of the Minneapolis NCS as a guide, we selected a PPS sample size of 355 establishments from the artificial universe of 2,880 establishments. A PPS sample of occupations within each establishment was then selected.

Artificial MSA Variance and Confidence Intervals

The set of domains in our analysis includes: All workers, MOGs, Levels, and MOG X Level. For each sample, we computed means, medians and other quantiles for hourly wages for these domains, using the current NCS estimation system.

Variance estimates for means were computed for each sample and each domain using the current NCS program in Tehonica, Ernst, and Ponikowski (1997). This approach uses the standard Taylor Series Method. The variance estimates for medians and other quantiles combined this program with the Woodruff method.

For each of the samples, two methods were applied to generate 95% confidence intervals (C.I.s) for each occupational group and work level. The first method produced the 95% C.I. using the standard normal quantile, such that

$C.I._{SN}$ = *estimate* $\pm$ (1.96 * standard deviation) ,

where the standard deviation is estimated from the particular sample. The second method generated the 95% C.I. using weighted degrees of freedom (d.f.) as defined by

$$df = \frac{\left[\hat{V}\left(\hat{\bar{Y}}_D\right)\right]^2}{\sum_i n_i^2 V_{iD}^2 K_{iD}^{-1} + \sum_{ik} m_{ik}^2 V_{ikD}^2 K_{ikD}^{-1}}$$

where, K_{iD} is the noncertainty establishments minus 1 and K_{ikD} is the certainty establishments minus 1, unless there is only one establishment in which case K=1.
Also, V^2_{ikD} is the certainty variance and V^2_{iD} is the noncertainty variance and is computed,

$$V_{iD} = \frac{1}{n_i - 1}\left[\sum_k \left(\hat{Z}_{ikD}\right)^2 - \frac{1}{n_i}\left(\sum \hat{Z}_{iD}\right)^2\right]$$

where,

$$\hat{Z}_{ikD} = \left(\frac{1}{\hat{X}_D}\right)\left(\hat{Y}_{ikD} - \hat{\bar{Y}}_D \hat{X}_{ikD}\right)$$

i is the sampling stratum
k is the establishment
D is the domain of interest
n is the number of establishments in the stratum
m is the number of occupational selections within an establishment
Y is the total weighted annual wages
X is the weighted hours worked

The certainty variance is computed similarly but at the quote level.

The degrees of freedom formula is similar in principle to earlier formulae for degrees of freedom in the case of stratification presented in Cochran (1977, p.96) and CDW (Section 3), but allows for the multi-level aspect of the sampling design. The second term in the denominator corresponds to within certainty establishment sampling variation of occupation wages. There is a corresponding term in $\hat{V}\left(\hat{\bar{Y}}_D\right)$.

Summary Statistics

We present in this section several summary statistics. The most important of these statistics is the proportion of C.I.s, which contained the true universe values for hourly wages, for the different domains, for both confidence interval methods. It is these proportions that are used to compare the performance, in terms of coverage, of the standard normal confidence intervals (columns labeled Normal % Cover) to that of the t-distribution with weighted degrees of freedom (columns labeled t-dist % Cover).

The second column in Table 1A shows the mean hourly wage by major occupational group (population value). This is a weighted mean estimate based on number of workers and corresponding hours worked. Table 1B shows the average hourly wage by occupational level. The occupational levels are equivalent to the federal government GS levels.

The third column of Table 1A and 1B is the root mean square error across runs of the estimated mean hourly wages, and should be an accurate estimate of the true standard error. The relatively high values for Sales in Table 1A reflect the skewed distribution of these wages as a result of commission based pay. The next column, the standard error column, is the arithmetic mean of the estimated standard errors, computed from the variance estimation program, of the 100 samples. SE/RMSE, the ratio of the previous two columns is an indication of whether the standard error computed from the variance estimation program underestimates or overestimates the true standard error. In Table 1B, the standard error is an overestimate for the six lowest occupational levels and an underestimate for the six higher levels. This may be related to the fact that the distribution of wages for the higher levels is more skewed.

The next two columns show the percentage of samples for which the estimated confidence intervals cover the true population mean. In Table 1A, of the seven occupational groupings for which coverage was not the same for both methods, five were closer to the ideal of 95% coverage for the t-distribution. In Table 1B, for the analogous comparison the t-distribution performed better in 7 out of 14 levels, with large improvements for levels 10, and 13 through 15. These levels also happen to be associated with relatively low degrees of freedom (i.e. 5 or below). This is consistent with results in Table 1A for the Technical occupational group and the Sales occupational group, the two occupational groups with degrees of freedom less than ten.

The next column, the ratio of the confidence interval lengths, is the geometric mean of the 100 interval lengths using the normal approach divided into the geometric mean of these intervals using the t-distribution, for the corresponding domain.

Generally high ratio values are associated with the lower degrees of freedom, which is given in the last column as the arithmetic mean of the degrees of freedom for the domain over the 100 samples. The cases where there were few observations for the domain of interest are those that the alternative approach for constructing confidence intervals lengthened the intervals most and had the lowest degrees of freedom.

Table 1C shows the same statistics as in Tables 1A and 1B by MOGL. These statistics are given for "all MOGLs", calculated as the arithmetic mean of the results for each of the 76 MOGLs. In cases involving ratio's (i.e., SE/RMSE and the ratio of the CI lengths) the geometric mean was used instead. These statistics are also given for the MOGL with the minimum ratio of the CI lengths (Professional MOG, Level 9) and the MOGL with the maximum ratio of the CI lengths (Machine Operators MOG, Level 2). Table 1C shows a marked improvement in coverage for the All MOGL line, 85.7 percent for the normal coverage improved to 94.6 percent for the t-distribution. This represents a greater improvement than the MOG or level breakouts. The SE/RMSE is an indication of whether the computed SE underestimates or overestimates the true standard error. In Table 1C for all MOGLs the SE/RMSE ratio is less than 1, so the sample based standard errors are underestimates. These findings are consistent with those found in Paben (1999) who showed for the same artificial population that in general the Taylor Series variance estimates obtained through replication methods, result in under-estimating the variance for small domains.

Tables 2A-E show the summary statistics for 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles for All workers, MOGs, Levels and MOGLs. In each table the average normal percent coverage and the average t-distribution percent coverage are shown, where the confidence intervals are computed using the Woodruff method in Woodruff (1952). Since the Woodruff method does not produce confidence intervals that are symmetric around the estimate two alternative confidence intervals that are symmetric were constructed for comparison purposes. For the first method the length of the interval remains the same as the Woodruff confidence interval but was "shifted" to form a symmetrical interval around the estimate. The second method adjusts the interval by extending the shorter half to equal the length of the longer half of the interval. This results in a wider symmetrical interval than the first method of adjustment.

In Tables 2A-E, using the standard Woodruff method for constructing confidence intervals, the *t*-method does not show conclusive improvement in coverage for the All Workers and MOG estimates. For levels, the t-distribution approach brings coverage closer to the ideal 95% for the standard method in all but one case (i.e. Table 2C). All but one table (Table 2A) of the five tables show a coverage closer to the ideal 95% for the t-distribution for those estimates that have a smaller average degrees of freedom, that is the MOGLs. Table 2B coverage improves from 90.9 to 97.0 percent for the MOGLs. For the median estimate in Table 2C, coverage improves from 89.0 to 95.8 percent. Table 2D shows an improvement from 84.7 to 92.1 percent and Table 2E from 74.6 to 82.4 percent. Table 2A, Summary Statistics for 10^{th} Percentile of Hourly Wage Estimate does not show as much of an improvement. The symmetrical method shows similar improvement; however, the long-half method appears to over-extend the intervals. Table 2E shows improvement in coverage across all three methods for the *t*-method. These results may point to the improved performance of the t-distribution approach for the relatively skewed portion of the wage distribution (i.e., the upper tail or 90^{th} percentile). Also included in Tables 2A-E are the average ratio of the confidence interval lengths. As expected, the ratios increase (i.e., the intervals are lengthened) as the average degrees of freedom decrease.

Conclusion

Standard 95% confidence intervals for domain means, when based on the standard normal distribution and standard methods of variance estimation yield less than the actual 95% coverage, particularly for the smaller domains, with the smaller degrees of freedom, that is the MOGLs. Confidence Intervals using the t-distribution with weighted degrees of freedom produce intervals with 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 occupational group and level and associated average degrees of freedom.

Bibliography

Casady, R.J., Dorfman, A.H. and Wang, S. (1996), "Confidence Intervals for Sub-domain Parameters

when the Sub-domain Sample Size is Random", *Survey Methodology, 24, 139-145*.

Cochran, W.C. (1977), *Sampling Techniques*, 3rd Edition, Wiley, New York.

Harpenau, C.l., Coleman, J.L., and Lincoln, M.D. (1995), "Evaluation of Confidence Intervals Methodology for the occupational Compensation Survey Program", presented at Joint Statistical Meetings, Orlando.

Paben, Steven (1999), "Comparison of Variance Estimation Methods for the National Compensation Survey", presented at Joint Statistical Meetings, Baltimore.

Tehonica, J., Ernst, L.R., and Ponikowski, C.H. (1997), "Summary of Estimation and Variance Specifications for the 1996 Albuquerque, NM COMP2000 Test Survey", January 15, 1997 Memorandum.

Woodruff, R.S. (1952). Confidence intervals for medians and other positional measures. Journal of the American Statistical Association, 47, 635-646.

TABLE 1A
Average Hourly Wage Summary Statistics

Major Occupational Group	Pop Value	Root MSE	Avg. SE	Avg. SE/RMSE	Normal %Cover	t-dist. %Cover	Ratio of CI Lengths	Mean df
All Workers	16.28	0.3404	0.3316	0.974	91	94	1.019	87.7
Professional	26.42	0.8088	0.7750	0.958	96	97	1.081	33.9
Technical	16.84	1.0669	0.8697	0.815	90	97	1.418	7.7
Exec., Admin., Mgr.	27.42	1.2458	1.3242	1.063	97	97	1.119	20.2
Sales	12.69	2.0137	2.0090	0.998	88	91	1.383	9.2
Admin. Support	11.81	0.3299	0.2983	0.904	91	94	1.051	42.7
Precision, Production	17.97	0.7579	0.7603	1.003	95	95	1.149	20.6
Machine Operators	12.30	0.6133	0.5453	0.889	93	94	1.072	24.5
Transportation & Material	14.58	0.8765	0.8647	0.987	96	98	1.135	12.6
Handlers, Equip.	10.00	0.5266	0.5629	1.069	94	94	1.110	16.3
Service	13.99	0.6301	0.6233	0.989	94	94	1.092	15.2

TABLE 1B
Average Hourly Wage Summary Statistics

Levels	Pop Value	Root MSE	Avg. SE	Avg. SE/RMSE	Normal %Cover	t-dist. %Cover	Ratio of CI Lengths	Mean df
All Workers	16.28	0.3404	0.3316	0.974	91	94	1.019	87.7
1	6.72	0.1839	0.2131	1.159	97	98	1.080	28.3
2	8.37	0.3008	0.3110	1.034	94	95	1.131	21.9
3	9.90	0.2778	0.2907	1.046	96	96	1.067	23.9
4	11.74	0.2825	0.3707	1.312	99	100	1.060	27.8
5	13.95	0.3491	0.3870	1.109	94	95	1.120	17.4
6	15.63	0.6022	0.6173	1.025	96	99	1.157	12.1
7	18.78	0.6972	0.6482	0.930	93	98	1.404	12.8
8	20.50	0.9230	0.8986	0.974	92	95	1.214	7.1
9	23.66	0.7860	0.7967	1.014	96	99	1.315	8.3
10	25.83	2.1882	1.8054	0.825	84	92	1.589	5.0
11	28.57	1.8151	1.7701	0.975	94	97	1.349	8.5
12	33.50	1.6568	1.4586	0.880	92	99	1.442	6.5
13	39.57	1.8848	1.8149	0.963	87	93	1.414	5.4
14	46.08	2.9325	2.6210	0.894	86	94	1.603	4.0
15	50.42	11.863	9.4062	0.793	69	87	2.193	2.2

TABLE 1C
Average Hourly Wage Summary Statistics

Domain	Root MSE	Avg. SE	Avg. SE/RMSE	Normal %Cover	t-dist. %Cover	Ratio of CI Lengths	Mean df
All MOGLs	1.642	1.434	0.888	85.7	94.6	1.64	30.7
Min. of MOGLs	0.698	0.807	1.150	95.0	97.0	1.14	12.5
Max. of MOGLs	1.900	1.407	0.740	82.0	93.0	2.34	9.8

TABLE 2A Summary Statistics for 10th Percentile of Hourly Wage

			Standard Method			Symmetrical Method		Long-Half Method			
Domain	Number of Estimates	Avg. SE/RMSE	Avg. Normal % Coverage	Avg. t -dist. % Coverage	Avg. Ratio of CI Lengths	Avg. Normal % Coverage	Avg. t -dist. % Coverage	Avg. Normal % Coverage	Avg. t -dist. % Coverage	Avg. Ratio of CI Lengths	Avg. df
All Workers	1	0.8702	100.0	100.0	1.01	83.0	84.0	100.0	100.0	1.01	114.0
MOGs	10	1.1039	93.8	96.0	1.20	90.4	93.1	95.3	97.1	1.22	22.5
Levels	15	1.5777	93.4	96.3	1.48	90.9	95.1	95.7	97.5	1.55	13.0
MOG x Levels	76	2.0430	92.2	97.9	1.50	92.5	97.9	96.1	98.9	1.46	4.4

TABLE 2B Summary Statistics for 1st Quartile of Hourly Wage

			Standard Method			Symmetrical Method		Long-Half Method			
Domain	Number of Estimates	Avg. SE/RMSE	Avg. Normal % Coverage	Avg. t -dist. % Coverage	Avg. Ratio of CI Lengths	Avg. Normal % Coverage	Avg. t -dist. % Coverage	Avg. Normal % Coverage	Avg. t -dist. % Coverage	Avg. Ratio of CI Lengths	Avg. df
All Workers	1	1.0443	98.0	98.0	1.01	96.0	96.0	99.0	99.0	1.01	133.3
MOGs	10	0.9819	94.0	96.0	1.14	92.0	94.0	94.8	96.8	1.14	27.2
Levels	15	1.1385	92.9	96.7	1.42	91.3	95.6	94.8	97.6	1.45	16.1
MOG x Levels	76	1.4707	90.9	97.0	1.90	89.2	96.5	93.9	98.4	1.93	5.2

TABLE 2C Summary Statistics for the Median of Hourly Wage

			Standard Method			Symmetrical Method		Long-Half Method			
Domain	Number of Estimates	Avg. SE/RMSE	Avg. Normal % Coverage	Avg. t -dist. % Coverage	Avg. Ratio of CI Lengths	Avg. Normal % Coverage	Avg. t -dist. % Coverage	Avg. Normal % Coverage	Avg. t -dist. % Coverage	Avg. Ratio of CI Lengths	Avg. df
All Workers	1	0.9218	92.0	93.0	1.02	92.0	94.0	94.0	95.0	1.01	88.7
MOGs	10	1.0909	95.9	97.0	1.12	93.9	96.4	97.2	98.4	1.12	24.0
Levels	15	1.1011	93.5	97.1	1.34	93.3	96.6	95.3	98.4	1.34	15.5
MOG x Levels	76	1.0466	89.0	95.8	1.86	85.4	94.3	91.8	97.4	1.83	5.7

TABLE 2D Summary Statistics for 3rd Quartile of Hourly Wage

			Standard Method			Symmetrical Method		Long-Half Method			
Domain	Number of Estimates	Avg. SE/RMSE	Avg. Normal % Coverage	Avg. t -dist. % Coverage	Avg. Ratio of CI Lengths	Avg. Normal % Coverage	Avg. t -dist. % Coverage	Avg. Normal % Coverage	Avg. t -dist. % Coverage	Avg. Ratio of CI Lengths	Avg. df
All Workers	1	0.9817	95.0	95.0	1.02	90.0	91.0	95.0	95.0	1.02	54.2
MOGs	10	1.0120	94.9	96.6	1.16	91.0	93.4	96.2	97.6	1.17	18.7
Levels	15	1.1733	91.8	95.3	1.37	89.4	93.6	94.0	97.3	1.41	12.9
MOG x Levels	76	1.1249	84.7	92.1	1.69	81.3	90.8	88.3	95.3	1.72	5.0

TABLE 2E Summary Statistics for 90th Percentile of Hourly Wage

			Standard Method			Symmetrical Method		Long-Half Method			
Domain	Number of Estimates	Avg. SE/RMSE	Avg. Normal % Coverage	Avg. t -dist. % Coverage	Avg. Ratio of CI Lengths	Avg. Normal % Coverage	Avg. t -dist. % Coverage	Avg. Normal % Coverage	Avg. t -dist. % Coverage	Avg. Ratio of CI Lengths	Avg. df
All Workers	1	1.1399	96.0	97.0	1.04	98.0	98.0	98.0	98.0	1.04	51.2
MOGs	10	1.1320	90.5	93.8	1.31	84.7	88.8	93.3	95.5	1.34	16.5
Levels	15	2.0942	87.2	92.4	1.46	83.3	88.9	89.8	94.7	1.50	9.1
MOG x Levels	76	0.9499	74.6	82.4	1.52	69.6	79.2	78.8	90.1	1.49	3.9

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