DEVELOPING AN ESTIMATOR TO ESTIMATE THE VARIANCE OF MEAN WAGE RATES COMPUTED FROM GROUPED DATA IN THE OCCUPATIONAL EMPLOYMENT STATISTICS SURVEY

Kenneth W. Robertson, Albert Tou, Larry Huff
K.W. Robertson, U.S. Bureau of Labor Statistics, 2 Mass. Ave. N.E., Washington, D.C. 20212

Outline.

I. Description of Problem.

The Occupational Employment Statistics (OES) survey is an annual mail survey of business establishments conducted by the U.S. Bureau of Labor Statistics. The design of the survey has recently been changed from a State / Industry based survey collecting occupational employment totals to a Metropolitan Statistical Area / Industry based survey collecting occupational employment totals and occupational wages. The respondents are asked to classify their establishments' employees into a matrix structure which includes Occupational Titles and 11 contiguous, nonoverlapping wage intervals. These wage data are obtained from payroll records by the responding establishments. Therefore, we assume that there is no memory-related recall error associated with these wage data. Estimators for both the mean and median have been developed for use with these wage data. To date, however, no variance estimators have been derived for these statistics. The impetus for this project was the need to develop a variance estimator for the mean wage estimates. The wage intervals currently used are shown in the following table.

OES Survey Wage Intervals

Interval	Hourly	Annual
A	less than $5.74	less than $11,960
B	$5.75 - $8.49	$11,960 - $17,679
C	$8.50 - $9.99	$17,680 - $20,799
D	$10.00 - $11.24	$20,800 - $23,399
E	$11.25 - $13.24	$23,400 - $27,559
F	$13.25 - $15.74	$27,560 - $32,759
G	$15.75 - $19.24	$32,760 - $40,039
H	$19.25 - $24.24	$40,040 - $50,439
I	$24.25 - $43.24	$50,440 - $89,959
J	$43.25 - $60.00	$89,960 - $124,800
K	more than $60.00	more than $124,800

Data collected in intervals are somewhat less clear than an exact data point would be. For example, a person who indicates that their salary is within wage interval F (*$27,560 - $32,759*) actually has one precise data point within that wage interval which describes their salary. Therefore, the data collected within any particular wage interval possesses some unknown, underlying distribution bounded by the endpoints of the interval.

This project is primarily concerned with one approach to variance estimation for wage data collected within intervals. An obvious choice for an estimator for the variance of these data is the usual variance estimator for frequency data. This grouped data variance estimator (GDVE) can be expressed as follows

[1] [equation 1] $$S_G^2 = \frac{\sum_r n_r (c_r - \bar{x})^2}{n-1},$$

where r indicates a class or interval, and c_r represents the midpoint of class r. This estimator has several shortcomings. First, when all of the data are in one interval this estimator gives us a variance of zero. This is because we assume that all data points within the interval are located at the midpoint. When considering the distribution of data points underlying the frequency data this is an underestimate of the true variance. An empirical investigation of this estimator (see the Table below) shows a second shortcoming. When all of the data are not in one interval it has a considerable upward bias for variance estimation of the underlying distribution.

Empirical Distribution of the Percentage Error of the Grouped Data Variance Estimator when compared to the Standard Variance Estimator*.**

Percentile	Percent Error	Percentile	Percent Error
0.01	-100.00 %	0.50	6.15 %
0.05	-74.56 %	0.75	17.27 %
0.10	-31.28 %	0.90	33.26 %
0.25	-6.20 %	0.95	45.39 %
0.50	6.15 %	0.99	83.29 %

* The Standard Variance Estimator is $$S_S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

$$** = \left(\sqrt{\frac{S_G^2}{S_S^2}} - 1 \right) * 100 ,$$

also see the description of Q later in this document. This corresponds to the percentage by which the standard errors differ for the 9,625 empirical estimates obtained from the data when applying the grouped data variance estimator and the standard variance estimator to these data. The data are described in Section II. Note that these values were derived by classifying the data by Industry and Occupation. The values provided later in the paper were classified by Geographic area, Industry, and Occupation.

The table above provides us with the percentile distribution of the percentage errors obtained when using the Group Data Variance Estimator instead of the Standard Variance Estimator. The statistics in the table above show us that 50% of the grouped data variance estimates are overestimating the true variance (the Standard Variance Estimator) by 6.15% or more. Since we would expect an unbiased distribution of errors to have a median near 0%, this median of 6.15% indicates a significant upward bias to this distribution of errors. At least 1% are underestimating by 100%. That is, the standard variance estimate is non-zero, while the grouped data variance estimate is zero. Further examination of this distribution of errors shows that it is bimodal. There is a pronounced spike at -100% and a peak at 6.15%. Therefore, the problem is that the most obvious estimator for the variance of these grouped data is biased and produces an error distribution which is bimodal.

An alternative approach to variance estimation is to utilize an auxiliary data set to derive a set of location parameters for use with the interval data set. These location parameters are referenced in this paper as Dispersion Location Statistics (DLS's). These parameters would replace $\mathbf{c_r}$ in [equation 1] above with values optimized to produce desirable variance characteristics. In this paper we explored this alternative approach. Constrained multivariate optimization is used with an auxiliary data set to develop optimized parameters for use in variance estimation for the interval data. The auxiliary data set must contain data which were not collected by interval. We make the basic assumption that the distribution of the auxiliary data is the same as the distribution underlying the interval data.

II. Data.

In this project we utilized a data set containing wage values collected via personal visit by the Office of Compensation and Working Conditions of the Bureau of Labor Statistics (BLS). These data were collected for the Occupational Compensation Survey Program (OCSP). The data set contains approximately 1.3 million observations which are classified by Industry and Occupation. Each observation contains a wage rate, and the number of employees within the organization in a specific Occupation making that wage rate. Using these data we can estimate the mean wage rate by Area, Industry, and Occupation. Each Industry / Occupation combination defines an estimation *category*. For example, one category may be defined as Warehouse Specialists within the Furniture and Fixtures industry in Area 1.

The classification values allow us to arrange the observations into 12,424 Area / Industry / Occupation *categories*. For example, we may have hundreds of observations for Warehouse Specialists in the Furniture and Fixtures industry. These observations would cover a considerable range of salaries. Therefore, we could use these observations to estimate a mean and variance for the wage data in this category. These data will also be used to simulate interval data using the intervals developed for the OES survey.

The data are reasonably classified by attributes reported by the survey respondents. The primary attributes used for classification are the category attributes, area, industry and occupation. It would make sense to use this classification in our research. Unfortunately, there is not a complete one-to-one mapping of categories from the OCSP survey to the OES survey. The OCSP survey collects data on only a subset of the occupations that the OES survey collects. Therefore, in order for this research to be of general use for the OES survey, some set of attributes which provides a complete one-to-one mapping for the surveys must be used as a classification mechanism.

We chose to use the empirically determined start/end interval as a grouping mechanism. For example, we located all Industry/Occupation categories which have data reported in both wage interval A and wage interval B, but in no other intervals, and call this group *AB*. Similarly, we could locate all categories which have data reported in wage interval A through wage interval C, but in no other intervals, and call this group *AC*. This classification does map one-to-one between the OCSP and the OES surveys. Classifying the data in this manner results in 66 groups.

Notice that within each group there are a number of categories. We emphasize that each category contains many observations. An estimate of the mean and variance must be produced for each category. Therefore, within each group there will be a number of variances to examine.

III. Research.

There are several statistics associated with a probability distribution that are in common use. The more common of these statistics can be placed in two classes, those that tell us something about location, and those that tell us something about dispersion. Imagine, if you will, that we have an additional location statistic which is also useful in describing dispersion. Let's call this statistic the *DLS*, for Dispersion Location Statistic, and denote it with **m**. For the simplest case we could define this as follows:

[equation 2]

$$\sum_{i=1}^{n}(m-\bar{x})^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$n(m-\bar{x})^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$\begin{cases} m = \bar{x} + \sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} & , \; m > \bar{x} \\ m = \bar{x} - \sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} & , \; \bar{x} > m \end{cases}$$

Therefore, $|m - \bar{x}|$ describes an average dispersion of the data about the mean. As described above, **m** could replace all of the individual x_i values in the sum of squares calculation without changing the value of the sum of squares. If we consider this in conjunction with data placed into intervals, it is easy to imagine that there is some *DLS*, or **m** within each interval. At this juncture it might be useful to take a closer look at the standard variance estimator sum of squares in an interval setting, and expand it using the usual ANOVA techniques. Additionally, we will look at what a sum of squares calculation would look like with the **m** vector in place.

[2] [equation 3]

$$\sum_r \sum_i (x_{ri} - \bar{x})^2 = \sum_r \sum_i \left((x_{ri} - c_r) + (c_r - \bar{x})\right)^2 =$$

$$\sum_r n_r (c_r - \bar{x})^2 + \sum_r \sum_i (x_{ri} - c_r)^2 + 2 \sum_r \sum_i (x_{ri} - c_r)(c_r - \bar{x})$$

[equation 4]

$$\sum_r \sum_i (m_r - \bar{x})^2 = \sum_r \sum_i \left((m_r - c_r) + (c_r - \bar{x})\right)^2 =$$

$$\sum_r n_r (c_r - \bar{x})^2 + \sum_r n_r (m_r - c_r)^2 + 2 \sum_r n_r (m_r - c_r)(c_r - \bar{x})$$

A comparison of these two equations shows that the definition of **m** described earlier is no longer completely appropriate unless r is equal to one and $c_r = \bar{x}$. The addition of multiple intervals and midpoints forces us to reevaluate what these **m** statistics mean in this setting. In [equation 3], the final term provides for some measure of cancellation within each interval r, as we would expect that the x_{ri} are dispersed on both sides of the midpoint c_r. In fact, if $c_r = \bar{x}_r$ then the final term sums to 0. That is, the dispersal to either side of the midpoint is exact by definition. In our case $c_r \neq \bar{x}_r$ at the level of estimation. However, it is probably somewhere near that statistic since it is a "global" midpoint for that interval. We would thus expect some dispersion around this midpoint, but not a complete cancellation of this term. When examining [equation 4] , however, we see that cancellation is not taking place at all in this context in the final term. Therefore, **m** must be chosen within each interval in a way which minimizes any error due to this loss of cancellation. While $|m_r - c_r|$ can still be thought of *conceptually* as indicating a measure of dispersion about a midpoint, there is now some error associated with this conception. If we take values for **m** chosen based on the dispersion within each interval as defined previously, and then make slight adjustments to **m** we could reduce this error. The only cost of this adjustment is that of deviating from our conceptual definition. Obviously, a complete derivation of an interval based vector of **m** values and an adjustment to account for the interval induced error is significantly more complex than that given so far. Given the scope and intent of this project, a complete algebraic derivation of the more complex case will not be made here.

So far, we have described a statistic which can accurately substitute for known data in a sum of squares calculation. Further, we have discussed a conceptual extension of this statistic (with some error) into a vector of statistics to represent known data within an interval structure in a sum of squares computation. Now imagine that we must use one vector of these statistics to represent data in multiple categories. Additionally, we are going to use this vectors to minimize the errors associated with a ratio variance estimator. Obviously, there will not be one vector which will provide us with an exact replacement for data in each interval of each category. In this scenario we would expect to have an error term associated with each category. What we would like to do is pick the **m** vector in such a way that the errors from all categories are minimized.

We make the assumption throughout the rest of this paper that there is an **m** vector, as described above, underlying data placed within intervals or groups. If the data are factually represented by the midpoint of the interval, then the **m** vector will equal the midpoints of the intervals. In this case, use of the **m** vector in variance estimation would provide the same sum of squares as the grouped data variance estimator. If the data are not factually represented by the midpoints of the intervals, then using the **m** vector should provide an increase in the accuracy of the variance estimates. Because of the complexity of a mathematical derivation of the vector, and because of the **m** vectors dependence on the data within each interval, we must find some other method for determining the value of the **m** vector. When attempting to locate a minimum or maximum of a complex function we often turn to numerical techniques. Numerical techniques have been developed which are useful for solving complex function minimization problems. These methods include both simple searches, and optimization techniques. Since we have a complex function we need to estimate, optimization is an appropriate technique to use. Optimization will allow us to develop values for an **m** vector which minimizes a sum of ratio variance estimate errors from multiple estimates simultaneously.

As mentioned earlier, in order to map one-to-one from the OCSP data to OES data, we had to classify the data into 66 Groups. Therefore, instead of calculating an exact **m** vector for each of the 12,424 categories, we must

calculate an **m** vector for each Group which provides us with a good approximation for each category contained in the Group. One additional constraint must be added to our discussion of the **m** vector. Since each $\mathbf{m_r}$ represents every $\mathbf{x_{ri}}$ within interval r, its value should be constrained such that $\mathbf{lb_r} \le \mathbf{m_r} \le \mathbf{ub_r}$, where $\mathbf{lb_r}$ is the lower bound of interval r, and $\mathbf{ub_r}$ is the upper bound of interval r. We could then find the values of **m** which minimize the sum

[equation 5]

$$\sum_{j \in Group} \left(\left(\frac{\sum_i w_i \left(\hat{y}_i - \hat{R} x_i \right)^2}{n \bar{x}^2 (n-2)} \right) - \left(\frac{\sum_i w_i \left(y_i - \hat{R} x_i \right)^2}{n \bar{x}^2 (n-1)} \right) \right)^2 = \sum_{j \in Group} \left(S_{Rm}^2 - S_{RS}^2 \right)^2$$

$S_{Rm}^2 = PSRVE$

$S_{RS}^2 = SRVE$

$\hat{y}_i = \sum_r x_{i,r} \hat{m}_r$

y_i = *reported wages for unit i*

where

i = establishment,
r = interval,
x = employment, and
j indicates an Industry/Occupation *category*. The summation is over all categories within the current Group. The index value j is assumed on each variable.

PSRVE denotes a "Parameter Set" Ratio Variance Estimator, and SRVE denotes a "Standard" Ratio Variance Estimator.

That is, we want the estimated sum of squares to be as close as possible to the target sum of squares.

Using the auxiliary wage data set we calculated the mean and the variance for each Area / Industry / Occupation category. We then placed these data into the OES survey wage intervals. Using these data we can then approach the problem of determining the optimal values of $\mathbf{m_r}$ to use to minimize the difference between the estimated sum of squares and the target sum of squares. Each Industry / Occupation category was further classified by the data start interval and end interval as indicated earlier. This allowed us to develop parameter sets for groups of data with similar characteristics. This also leaves us with a manageable number of parameter sets, which could later be used in the systematic production of variance estimates.

After exploring several options we decided to use the following optimization function:

[equation 6]

Minimize $\sum_{j \in Group} \left(\frac{S_{Rm}^2}{\left(S_{RS}^2\right)^\alpha} - \frac{S_{RS}^2}{\left(S_{RS}^2\right)^\alpha} \right)^2$

Subject to

$lb_r < m_r < ub_r, \quad r \in \{1, 2, \ldots, 11\}$

Where

$\alpha = \{0, 1/16, 2/16, \ldots, 1\}$

as before, j is assumed on all variables

As indicated by [equation 6], α is being allowed to range from 0 to 1 by 16^{ths}. The size of the gradation was chosen to be fine enough to allow as wide range of α values as possible without over-taxing the processing capabilities of the system used. By altering α we make adjustments to the shape of the function near the local minimums. This results in slightly different parameter sets which must then be evaluated through additional testing to determine which is best. Different parameter sets will, of course, result in variance estimates with different distributional characteristics. Therefore, a particular optimized parameter set, while providing us with the minimum of that particular optimization function, may not be as "optimal" in other regards. This is one of the primary reasons why α is being altered in the optimization equation. It provides us with an opportunity to test the distributional results of slightly different parameter sets and choose which is best.

Again, we should mention that within each group there are a number of categories. An estimate is produced for each category. Therefore, there are a number of PSRVE estimates and SRVE estimates within each group. Obviously, our goal here is to produce a set of estimates for each group which is relatively unbiased, and closely associated with their target values. Therefore, what we are hoping to have is a distribution of variance errors (ε_V = PRSVE-SRVE) which are centered at zero, and have as small a median absolute deviation as possible. In addition to this, in order to be conservative we would rather overestimate the variance than underestimate it. This last criterion is achieved by only looking at parameter sets which produce estimate sets with a group median error greater than or equal to zero. If no parameter sets fit this criterion, then the parameter set which resulted in the least negative group median error was selected for that group. If multiple parameter sets pass this test, then a Median Absolute Deviation (MAD) statistic was calculated. The final parameter set for each group was chosen by finding the parameter set which produced the smallest MAD. This value was calculated as follows

[equation 7]

$$MAD_\alpha = median \left| \varepsilon_{j,\alpha} - median(\varepsilon_{j,\alpha}) \right|$$

where $\varepsilon_{j,\alpha} = PSRVE_{j,\alpha} - SRVE_j$

Evaluating the Results.

Once these optimization functions have been minimized and we have chosen a final parameter set, we must have a way of evaluating the results. While the optimization function is reasonable, there are a few undesirable characteristics when using that function as a final tool in evaluating the results. Primarily, the resulting distribution of errors will be difficult to interpret. For example, if the distribution is centered at a numerical value of 5,000, does this mean that we did a good job of parameterization, or a bad job? A further complication is that the scale changes as α changes. For these reasons a separate function would prove useful in evaluating the results.

A function which can be used to evaluate the results of the optimization is

[equation 8] $$W = \frac{S_{Rm}^2}{S_{RS}^2}$$

We will primarily look at the distribution of

[equation 9] $$Q_P = \left(\sqrt{W} - 1\right) * 100$$

This distribution will provide us an indication of the percentages by which we are over and underestimating the relative standard error. This also gives us an indication of how accurate the associated confidence interval will be. For example, if Q_P is -10, then this tells us that the confidence interval will be 10% smaller than the target confidence interval.

In order to call this method a success, we would like to have a large percentage of the distribution of Q_P to be between -50 and +50. This will provide us with a large percentage of relative errors between -50% and +50% of the target value of the relative errors. Additionally, we can compare this distribution to the distribution generated by Q_c, which compares the modified grouped data variance estimator to the standard ratio variance estimator.

IV Results.

As a first step in our evaluation, we looked at a ratio estimator modification of the Grouped Data Variance Estimator. We then looked at the error distribution of this estimator, grouped by the number of intervals into which the data fell. A review of Equations 3 and 4 shows that we can partition the variance into three terms; the variance within an interval, the variance across intervals, and a cross product. As the number of intervals into which the data falls becomes larger, we expect that the contribution of the across-interval component to the variance would also become larger. Conversely, the within-interval component would tend to become less important. When we looked at the distribution of errors from this modified grouped data variance estimator, by the number of intervals into which the data fell, we observed exactly this. As the number of intervals rose, the error distribution became smaller, and less biased. In fact, the modified grouped data variance estimator was providing acceptable results with as few as three intervals. Because of this, and due to the amount of time the optimization routine was taking, we decided to produce m-vectors for groups where the data fell into only one or two intervals.

The following table shows the median, the mean, and the number of estimation categories by the number of intervals the data fell into.

Mean and Median of the Error Distributions by the number of intervals the data fell in.

Q_P

Intervals	Mean	Median	n
1	30.3	**6.2**	1,290
2	3.8	**-4.6**	2,976

Q_c

Intervals	Mean	Median	n
1	-100.0	**-100.0**	1,290
2	-13.8	**-19.2**	2,976
3	4.2	**2.2**	2,962
4	6.1	**4.6**	2,272
5	4.7	**4.5**	1,512
6	4.5	**3.5**	951
7	1.9	**2.1**	385
8	3.0	**4.1**	71
9	9.4	**8.7**	5

The reader can readily see the improvement when using the PSRVE as opposed to the SRVE for data falling into less than three intervals. Notice that the median of the error distribution shifts from -100% to +6.2% when the data fall in one interval, and from -19.2% to -4.6% when we use the PSRVE instead of the SRVE.

The following table indicates the proportion of the error distribution which is less than -50%, the proportion which is between -50% and +50%, and the proportion which is greater than 50%.

Proportion of Error Distributions within the bounds indicated, by the number of intervals the data fell in.

Q_P

Intervals	< -50	Between -50 & +50	> +50
1	4.9%	66.2%	28.9%
2	13.2%	70.8%	16.0%

Q_c

Intervals	< -50	Between -50 & +50	> +50
1	100.0%	0.0%	0.0%
2	23.7%	66.7%	9.6%
3	5.5%	86.7%	7.8%
4	2.0%	93.9%	4.1%
5	1.1%	97.5%	1.5%
6	0.5%	98.2%	1.3%
7	0.8%	99.0%	0.3%

8	0.0%	100.0%	0.0%
9	0.0%	100.0%	0.0%

As in the previous tables, the reader can readily see the improvement when using the PSRVE, as opposed to the SRVE when the data fall into less than three intervals. Notice that the proportion of the error distribution between -50% and +50% shifts from 0% to +66.2% when the data fall in one interval, and from 66.7% to 70.8% when we use the PSRVE instead of the SRVE.

The final table, below, provides information on the error distributions by percentiles.

Distributions of Q_c and Q_P, by percentile.

	Location of the Percentile Indicated	
Percentile	Q_c	Q_P
0.05	-100.0%	-49.8%
0.10	-100. 0%	-35.7%
0.25	-29.5%	-14.0%
0.50	-1.6%	2.0%
0.75	13.0%	17.6%
0.90	32.5%	47.6%
0.95	51.2%	83.6%

When viewed this way, we can see that the PSRVE is providing us with a larger set of errors between -50% and +50% than is the SRVE, which was our goal. We can also see that the middle 50%, as well as the middle 80% of the distribution is more centered for the PSRVE. Furthermore, we note that where once we had a noticeable bias towards underestimation with the SRVE, we now have shifted that to an upward bias with the PSRVE. While this upward bias is not desirable, it is preferred over a downward bias for variance estimation.

V. Conclusions.

The first topic explored in this paper was a construct called the *DLS*, or Dispersion Location Statistic. This was defined as a statistic useful for variance estimation. This statistic, and the mean, allow us to calculate the average dispersion of the data. This statistic can replace the individual data points in a simple variance calculation without a loss of precision. A conceptual extension of this statistic as a multivariate vector was described. This extension proves useful when data are collected in contiguous, nonoverlapping intervals. If appropriate values can be estimated for this vector, and the data are not factually represented by the midpoints of the intervals, then the usual grouped data variance estimator can be improved by using this vector. This is especially the case when the data fall in only a few intervals.

We then described an optimization procedure which allowed us to estimate values for the vector described above. These values were then used to improve the usual grouped data variance estimator. The results show that this estimator provides a better error distribution than the usual grouped data variance estimator if we assume that the data are not factually represented by the midpoint of the interval. A considerable improvement in both the bias and the dispersion of the estimator is demonstrated. The results suggest that the parameter set variance estimator developed using optimization on an auxiliary data set will prove to be useful for it's intended application in developing a wage variance estimator for the OES survey.

The results also suggest that this methodology might be useful in a more general sense when it is erroneous to assume that the data within a group are all represented by the midpoint. That is, if it is not correct to assume that all data within an interval are represented by the midpoint, then this methodology may provide the user with a considerable increase in the accuracy of the variance estimates as compared to the usual grouped data variance estimator.

VI. Future Research.

Future research might include investigating a method to include the secondary distributional testing as part of an optimization function. Additionally, we need to test the effectiveness of the parameters on a data set other than the one used to develop them. Another issue of interest would be to try different "grouping" methods than that used here. That is, we could have classified the categories by blue and white collar, or by some other classification mechanism which maps one-to-one across surveys. Each of these issues, both independently and collectively could potentially improve the results obtained here.

Disclaimer

Any opinions expressed in this paper are those of the authors and are not to be construed as policy of the Bureau of Labor Statistics.

Several of the issues in this paper are discussed in greater detail in an earlier work by Robertson[3].

Bibliography

1. Hamburg, Morris (1979), *Basic Statistics*, 2nd Edition, page 67. Harcourt Brace Jovanovich, Inc.. Note that the equation is presented here with different characters representing the variables.
2. Freund, John E., and Walpole, Ronald E., (1987), *Mathematical Statistics*, Fourth Edition, page 501. Prentice Hall.
3. Robertson, K.W. (1997), *A Problem in Constrained Multivariate Optimization: Minimizing the Sum of Errors in a Sum of Squares Calculation.* (Unpublished) BLS documentation.