Comparing Alternative Median Wage Rate Estimators for the Occupational Employment Statistics Survey

Carrae Echols, Albert Tou, and Kenneth W. Robertson, all of the U.S. Bureau of Labor Statistics

Key Words: Grouped Data, Median

I. INTRODUCTION

The Bureau of Labor Statistics' Occupational Employment Statistics (OES) program is a joint federal-state cooperative program that conducts an annual survey of nonagricultural business establishments. The program's objective is to produce annual estimates of total employment, mean wage rate, and median wage rate by occupation within industry group and geographic region. Industry groups are defined by 3-digit Standard Industrial Classification (SIC) codes, 2-digit SIC codes, and major industry divisions. Geographic regions are defined by metropolitan statistical areas (MSAs), residual areas within a state that are not part of an MSA (balance of state), the 50 states plus the District of Columbia, Guam, Puerto Rico, and the Virgin Islands, and the Nation as a whole. Because occupational employment and wage rate estimates are produced at detailed geographic (MSA) and industrial (3-digit SIC code) levels, the OES sample design calls for combining up to three consecutive years of data in order to produce useful estimates.

The questionnaire design uses a matrix format that lists occupational titles down the left margin and contiguous, non-overlapping wage bands across the top margin (see the attached sample questionnaire). To reduce respondent burden, the exact wage values of individual workers are not collected. The survey, consequently, uses a "grouped data" procedure to estimate mean and median occupational wage rates. Furthermore, since wage rate estimates are calculated using wage data from three consecutive years, data from the two previous years must be adjusted to account for inflation.

Although the OES survey produces estimates of total employment, mean wage rate, and median wage rate, the focus of this paper is on the estimation of median wage rates. Findings from an empirical study of alternative median wage rate estimators will be reported.

II. DESIGN OF THE OES SURVEY

Sample Design

To produce MSA-level estimates by 3-digit industry at the desired level of reliability, the OES survey design requires a sample size of approximately 1.2 million establishments. A sample of this magnitude places a large burden on the responding units. Distributing the sample across three consecutive years reduces that burden to a tolerable level. Therefore, approximately 400,000 establishments are sampled each year. Combining the samples across a 3-year time period makes it possible to produce reliable MSA-level estimates on an annual basis, while keeping the respondent burden to an acceptable level. Overlap among the samples drawn during any 3-year cycle is kept to a minimum.

The sampling frame for the survey is the list of business establishments that reported their employment and wage data to the state Unemployment Insurance (UI) program. The frame for an OES survey conducted in the 4th quarter of year t is the UI data collected during the 2nd quarter of year t-1. Although we prefer to use UI data collected during the 4th quarter of year t-1 as our frame, UI data from the 2nd quarter are used instead because that is the most recent time period when frame employment data are available. Seasonal disparities that arise between the sampling frame and the data collection period are partially corrected by a process called benchmarking. This process ratio adjusts the weighted sample employment to the universe employment as of the survey reference month in year t.

Establishments in the sampling frame are stratified by state-MSA, 3-digit industry code, and size of establishment (i.e., employment size class code). The sample size for a state-MSA/3-digit industry cell is determined by a Neyman allocation process that equalizes the expected relative standard errors of selected "typical" occupations across all state-MSA/3-digit industry cells. The sample selection process is based on a permanent random number procedure that allows the OES survey to minimize sample overlap with other Bureau surveys. Additionally, procedures are in place to ensure that any particular establishment is not surveyed more than once during a three year cycle.

The sample allocation process described above results in a sample size of approximately 400,000 establishments in each of years t, t-1, and t-2. Each sample unit is assigned a sampling weight that is the reciprocal of its probability of selection for that year. When combining samples for two or three years, the sampling weights are adjusted accordingly.

Target Population

The target population is all nonagricultural business establishments for the 50 states, the District of Columbia (DC), Guam, (GU), Puerto Rico (PR), and the Virgin Islands (VI). The reference period of the

survey is October, November, or December of the current survey year and is dependent on the establishment's 2-digit SIC code.

Data Collection

Questionnaires are initially mailed out to almost all sampled establishments. Some of the larger establishments, however, are contacted by personal visit. Two follow-up mailings are sent to non-responding establishments at three week intervals. After the follow-up mailings, nonrespondents are contacted by telephone.

The OES program currently defines approximately 750 detailed occupations across seven major occupational groups.

III. ESTIMATING MEDIAN WAGE RATES: A DESCRIPTION OF THE CURRENT PROCEDURE AND A CLOSELY RELATED ALTERNATIVE PROCEDURE

These two procedures are identical for the first 6 steps below. They differ in the seventh step because they make differing assumptions concerning the distribution of the wage-employment data.

Median occupational wage rate estimates for year t are calculated as follows:

(1) Calculate inflation factors for the wage rate data from years t-1 and t-2. The purpose of these factors is to inflate wage rate data from the past two years up to the level of the current year (t).

(2) Apply the inflation factors to the wage interval bounds used during years t-1 and t-2. This is analogous to applying the inflation factors to the wage data within those bounds. Note: The wage interval structures for years t, t-1, and t-2 will, in all likelihood, overlap one another.

(3) Overlay the wage interval structures of years t, t-1, and t-2 to create an extended (or universal) interval structure.
A consequence of this overlaying process is that the 11 wage intervals defined for years t, t-1, and t-2 could result in the formation of an extended interval structure composed of 33 wage subintervals.

(4) Use a linear interpolation procedure to determine the amount of employment within each of the 33 wage subintervals.

Define

$$\hat{P}_{o,q,t} = \sum_{k \in t} w_k p_{o,k,q}$$

where

w_k = combined year sampling weight for establishment k, ratio adjusted so that the summed weighted employment matches population totals, and

$p_{o,k,q}$ = reported employment for occupation o in establishment k in wage subinterval q. Note that this is within the original 11 interval structure.

This is the weighted employment within each wage interval q, for year t. Define

$$\hat{P}_{o,r} = a_1 \hat{P}_{o,q,t} + a_2 \hat{P}_{o,q,t-1} + a_3 \hat{P}_{o,q,t-2}$$

where

α_1 = the proportion of interval q from year t which overlaps in a linear manner with interval r,

α_2 = the proportion of interval q from year t-1 which overlaps in a linear manner with interval r, and

α_3 = the proportion of interval q from year t-2 which overlaps in a linear manner with interval r.

(5) Calculate a cumulative frequency distribution of estimated occupational employment across all wage subintervals. Define

$$\hat{P}_{o\,r'} = \sum_{r=1}^{r'} \hat{P}_{o\,r}$$

This is the cumulative frequency of the wage subintervals up to and including subinterval r'.

Use the distribution defined in (5) above to identify the wage subinterval that encloses the median wage rate.

(6) Define r'' as the subinterval r' where the cumulative frequency $\hat{P}_{or'}$ is defined such that

$$\begin{cases} \hat{P}_{o\,(r'-1)} < \left(0.5 * \hat{P}_o\right) \text{ and} \\ \hat{P}_{o\,r'} \geq \left(0.5 * \hat{P}_o\right) \end{cases}$$

Step (7A) is a continuation of the ***current*** procedure from Step (6) above.
Step (7B) is a continuation of the ***alternative*** procedure from Step (6) above.

(7A) In the current procedure, a uniform distribution model is applied to the appropriate wage subinterval to derive a median wage rate estimate.

Apply a linear interpolation procedure to that wage subinterval to determine the median occupational wage rate estimate. Define

$$\hat{R}_{o,0.5} = lb_{r"} + \left(\frac{0.5 * \hat{P}_o - \hat{P}_{o\,(r"-1)}}{\hat{P}_{o\,r"}} \right) \left(lb_{r"+1} - lb_{r"} \right)$$

where

0.5 is the 50th percentile,

lb_r is the lower bound of subinterval r, and

$\hat{P}$ o is the estimated employment for occupation o.

When using this procedure we assume that the data are uniformly distributed within the subintervals. Other percentile estimates can be calculated in an analogous manner.

In summary, the method described above is a linear interpolation procedure that distributes the wage-employment data from three separate interval structures into one extended interval structure. A second linear interpolation procedure which assumes a uniform distribution within the subintervals of the extended wage interval structure is then used to derive the median wage rate estimate. This method will be referred to as the "linear interpolation & linear interpolation" (LILI) approach. Note that this is the procedure currently used for the OES survey.

Note that wage data from survey years t-1 and t-2 are updated (see Robertson and Frugoli) to represent current wages using 4th quarter-to- 4th quarter wage rate changes as reported by the Bureau's Employment Cost Index (ECI) program. Two key factors are used to determine wage rate changes. They are

1. the broad ECI occupational groupings that encompass the wage rate estimate. There are nine broad ECI occupational groups at the national level, and
2. the year differential between the current survey year and the year in which the wage data were collected.

A caveat: Because wage rate changes are determined by broad occupational groups at the national level instead of by detailed occupations at the MSA level, it is likely that these wage rate changes will gauge the movement of wage rates with varying levels of success for detailed occupations at the MSA level. For example, if the wage rate change for the broad occupational group "administrators and managers" at the national level is 0.05, then we expect that there is really a distribution of rates of change for detailed occupations within this broad group that has a mean rate of change of 0.05. Similarly, we expect that there is a geographic distribution of rates of change for this group that has a mean rate of change of 0.05. Since we use the average rate of 0.05 across all geographic areas and all detailed occupations within the broad occupational group, we expect that there is some level of error and bias associated with the use of this average.

(7B) In the alternative procedure, which we will refer to as the "linear interpolation & non-uniform" (LINU) approach, a "non-uniform distribution" model is applied to the appropriate wage subinterval to derive a median wage rate estimate.

Apply an interpolation procedure to that wage subinterval to determine the median occupational wage rate estimate. This interpolation procedure utilizes a data-derived non-uniform distribution.

- Using the data-derived distribution mentioned above, locate where the $\left[0.5 * \hat{P}_o - \hat{P}_{o,(r"-1)}\right]^{th}$ person lies, in ascending order based on wage rates, in interval r''.
- Next, use the data-derived distribution to identify the wage rate of that person in interval r''. This value is the median wage rate.

IV. DESCRIPTION OF A SECOND ALTERNATIVE ESTIMATION METHOD

This section will describe a second alternative procedure that is considered to replace the current procedure.

Binary Search method

In this method, a binary search procedure is used to designate a potential median value. The procedure is calculated as follows:

(1) Calculate inflation factors for the wage rate data from years t-1 and t-2. The purpose of these factors is to inflate wage rate data from the past two years up to the level of the current year (t).

(2) Apply the inflation factors to the wage interval bounds used during years t-1 and t-2. This is analogous to applying the inflation factors to the wage data within those bounds.

(3) Estimate the employment within each of the 11 wage intervals for each year. Define

$$\hat{P}_{o,q,t} = \sum_{k \in t} w_k p_{o,k,q}$$

where

w_k = weight for establishment k, ratio adjusted so that the summed weighted employment matches population totals, and

$p_{o,k,q}$ = reported employment for occupation o in establishment k in wage subinterval q.

Note that this is within the original 11 interval structure. This value is the weighted employment within each wage interval q, for year t.

(4) Use an iterative binary search procedure to designate a test wage value for the median wage rate. On the first iteration, the test wage value is set at the midpoint of the range defined by the lower bounds of intervals 1 through 11. Note that if 50 percent or more of the employment falls in the 11th interval, we do not calculate a median wage value. Define

q' as the interval in which the test wage value resides.

Apply a linear interpolation procedure to that wage interval to determine the total weighted employment less than the test wage value. Define

$$\hat{P}'_{o,test,t} = \sum_{q=1}^{q'-1} \hat{P}_{q'-1,t} + \left[\frac{\left(test.value - lb_{q'}\right)}{\left(lb_{q'+1} - lb_{q'}\right)} \hat{P}_{q',t} \right]$$

and

$$\hat{P}'_{o,test} = \hat{P}'_{o,test,t} + \hat{P}'_{o,test,t-1} + \hat{P}'_{o,test,t-2}$$

where

test.value = the test wage value being evaluated,

$\hat{P}'_{o,test,t}$ = the weighted employment less than or equal to the test wage valuefor year t, and

$\hat{P}'_{o,test}$ = the weighted employment less than or equal to the test wage value.

This value, $\hat{P}'_{o,test}$, is compared to the value $\left(0.5 * \hat{P}_o\right)$. If the values are within a designated tolerance, then the test value is accepted as the estimated median, otherwise a new potential median value is calculated which halves the search space in the appropriate direction. The procedure iterates until a test wage value within tolerance is found.

V. COMPARISON OF ALTERNATIVE ESTIMATION METHODS

In order to conduct the empirical research we utilized data from the Bureau's National Compensation Survey. These data are collected by personal visit, and are point wage data (the OES survey collects interval based wage data. Using the NCS data, we calculated the "True" median wage value for each estimate. We then placed the NCS data into the OES interval structures to simulate OES data, and used the three estimation procedures described in this paper. Two statistics, the percent relative error (%RE) and the percent absolute relative error (%ARE), are used to evaluate the alternative estimation methods. These statistics are defined as follows:

%RE = 100 * [Estimate – "True"] / "True"
%ARE = | %RE |

These statistics were calculated for each estimate, and a frequency distribution developed. The following tables present the distribution of these statistics for each method.

Table I. – Distribution of Percent Relative Errors

	Method		
Percentile Error	**LILI (current method)**	**Binary Search**	**LINU**
95th	10.68	10.75	**7.04**
90th	6.85	6.86	**5.00**
75th	2.97	2.98	**2.21**
50th	0.49	0.50	**0.12**
25th	-1.50	**-1.49**	-2.17
10th	-3.95	**-3.94**	-5.30
5th	-5.78	**-5.72**	-8.87
Mean Error	*1.30*	*1.31*	*-0.28*

Table II. – Distribution of Percent Absolute Relative Errors

	Method		
Percentile Error	**LILI (current method)**	**Binary Search**	**LINU**
95th	12.04	12.14	**11.69**
90th	7.91	7.94	**7.68**
75th	4.53	4.52	**4.38**
50th	2.19	2.19	**2.18**
25th	**0.89**	0.91	0.90
10th	**0.32**	0.33	0.33
5th	**0.16**	**0.16**	0.17
Mean Error	*3.64*	*3.68*	*3.49*

The method which performed the best in each case has been highlighted.

The relative error statistic may be used with these data to determine if the distribution of errors is biased. The mean of the LINU distribution is slightly negative, while the means of the other two distributions are positive. The LILI and Binary Search methods are very similar. The LINU method does the best from the

median to the 95th percentile. However, it performs the poorest from the 25th percentile to the 5th percentile. This is because the other two methods tend to be slightly biased towards overestimation with these data. Most of the larger differences in these distributions are showing up in the outer parts of the distribution. It is difficult to choose a best method based on the results in this table.

The percent absolute relative error statistic is used to guage how well the method performed based on the absolute size of the errors. An examination of Table II shows that all of the methods are almost identical out to the 95th percentile. The LINU method performs the best at the 95th percentile, so that would be the method of choice based on this table.

Choosing between these methods is difficult, as they perform approximately the same when reviewing both the relative error and the relative absolute error. As mentioned previously, the primary difference in these estimators appears to be at the outer portion of the error distributions. If we chose a new estimator at this point, we would choose the LINU method. However, as the reduction in the magnitude of the errors is small, we will continue to research additional methods before altering the current procedures. The current procedure, with its operational simplicity, will be retained.

One issue that concerns us is the percentage of large errors associated with all of these methods. Given this, we attempted to isolate where the bulk of these errors are occurring. We suspected that most of them would be occurring in the two higher wage intervals with the largest width. In order to explore this we divided the estimates data set into two data sets; one with those estimates where the median fell in one of the two widest intervals, and another with all of the other estimates. We then calculated the absolute relative error distribution for these data sets. These errors are presented below for the currently used (LILI) method.

Table III. – Distribution of Percent Absolute Relative Errors for two data sets

	LILI (current method)	
Percentile Error	**Median <u>IS</u> in one of the two widest intervals**	**Median <u>IS Not</u> in one of the two widest intervals**
95th	30.08	**7.64**
90th	23.64	**6.07**
75th	14.88	**3.74**
50th	8.16	**1.92**
25th	4.17	**0.79**
10th	1.59	**0.29**
5th	0.58	**0.14**
Mean	*10.82*	***2.66***

A review of Table III indicates that the two widest intervals do, in fact, have a much worse error distribution than the rest of the intervals.

VI. CONCLUSIONS AND FUTURE RESEARCH ACTIVITIES

In this paper we have presented the results of an empirical research project. We explored several alternative grouped-data percentile procedures for use with the OES survey data. Our primary conclusion concerns the performance of the current median wage rate estimation procedure. This procedure appears to be working reasonably well; it balances a slightly larger error distribution against significantly lower processing costs.

Another conclusion we can draw from this work is that the errors in the tails of the (error) distribution are larger than we would like to see. We have recommended several ways to reduce the size of these errors. The first recommendation is to add more intervals. The use of additional intervals would allow each individual interval to have a smaller width. One additional wage interval is being implemented for the 1999 survey. Another recommendation is to select the interval bounds in a manner that distributes the potential errors evenly across all of the intervals. This recommendation is also being implemented with the 1999 survey. Another alternative to explore is the use of multiple sets of wage intervals. Each set of intervals could be targeted towards the expected wage rates of a group of occupations; however, there are workload, burden, and data quality issues associated with this suggestion that must be explored before any recommendations can be made.

In the future we plan to continue our efforts to improve the OES survey estimates. Research is being planned that will include both estimator and questionnaire design issues.

The opinions expressed in this paper are those of the authors and do not reflect the opinions or policy of the Bureau.

Bibliography

"Statistical Issues for the Redesigned Occupational Employment Statistics Survey" Kenneth W. Robertson and Pamela L. Frugoli, 1999 ASA proceedings, forthcoming.