Survey Methods and Reliability Statement for the May 2018 Occupational Employment Statistics Survey

Introduction

The Occupational Employment Statistics (OES) survey measures occupational employment and wage rates for wage and salary workers in nonfarm establishments nationally, and in the 50 states and the District of Columbia, Guam, Puerto Rico, and the Virgin Islands.

About 7.7 million in-scope establishments are stratified within their respective states by substate area, industry, and ownership. Substate areas include all officially defined metropolitan areas and one or more nonmetropolitan areas. The North American Industry Classification System (NAICS) is used to stratify establishments by industry.

Probability sample panels of approximately 180,000 to 200,000 establishments are selected semiannually. Responses are obtained by mail, Internet or other electronic means, email, telephone, or personal visit. Larger employers typically report their employees' job titles or occupations and wage rates. Smaller employers typically report their employees' job titles or occupations across 12 wage ranges. The job titles and descriptions are used to classify workers into occupations in the Standard Occupational Classification (SOC) system.

Estimates of occupational employment and wage rates are based on six panels of survey data collected over a 3-year cycle. The final in-scope sample size when six panels are combined is approximately 1.2 million establishments. Total 6-panel unweighted employment covers approximately 82.6 million of the total employment of 143 million.

Occupational and industrial classification systems

The occupational classification system

The U.S. Office of Management and Budget's Standard Occupational Classification (SOC) system is used to define occupations. All panels through November 2016 were collected using the 2010 SOC system. The May 2017 through May 2018 panels used a slightly modified version of the 2010 SOC in which 21 detailed occupational codes were replaced with 10 aggregations of those occupations. In most cases, SOC detailed occupations were aggregated to the SOC broad occupation level. The remaining aggregations do not correspond to existing SOC broad occupations and use OES-specific codes and titles. The published

May 2018 estimates reflect the 2010 SOC with these 10 aggregations. The purpose of this aggregation is to achieve more robust estimates by combining SOC occupations that are similar and for which the survey does not have the information needed to distinguish between the occupations for accurate coding. More information about the 2010 SOC system can be found at www.bls.gov/soc/2010. More information on the SOC aggregations used in the May 2018 estimates is available in Table 1 at www.bls.gov/oes/changes_2017.htm.

The industrial classification system

The May 2018 OES estimates use the 2017 North American Industry Classification System (NAICS). More information about NAICS can be found at the BLS web site www.bls.gov/bls/naics.htm or in the *2017 North American Industry Classification System* manual available at www.census.gov/eos/www/naics/. Each establishment in the survey is assigned a 6-digit NAICS code based on its primary economic activity. The four panels from November 2015 to May 2017 were collected using the 2012 NAICS codes; these data were mapped to 2017 NAICS codes. The November 2017 and May 2018 panels were collected using the 2017 NAICS. Beginning with the May 2017 estimates, selected 4-digit and 5-digit NAICS industries previously published by OES are no longer published separately. More information on these NAICS aggregations is available at www.bls.gov/oes/changes_2017.htm in Tables 2 through 4.

Industrial scope and stratification

The survey covers the following NAICS industry sectors:

11	Logging (1133), support activities for crop production (1151), and support activities for animal production (1152) *only*
21	Mining
22	Utilities
23	Construction
31-33	Manufacturing
42	Wholesale trade
44-45	Retail trade
48-49	Transportation and warehousing
51	Information
52	Finance and insurance
53	Real estate and rental and leasing
54	Professional, scientific, and technical services

55	Management of companies and enterprises
56	Administrative and support and waste management and remediation services
61	Educational services
62	Healthcare and social assistance
71	Arts, entertainment, and recreation
72	Accommodation and food services
81	Other services, except public administration [private households (814) are excluded]

Federal government executive branch (assigned industry code 999100)*

State government (assigned industry code 999200)*

Local government (assigned industry code 999300)*

* These are OES-defined industry codes and not a part of the NAICS industry classification.

These sectors are stratified into 302 industry groups at the 3-, 4-, 5-, or 6-digit NAICS level of detail.

Concepts

An ***establishment*** is generally a single physical location at which economic activity occurs (e.g., store, factory, restaurant, etc.). Each establishment is assigned a 6-digit NAICS code. When a single physical location encompasses two or more distinct economic activities, it is treated as two or more separate establishments if separate payroll records are available and certain other criteria are met.

Employment refers to the number of workers who can be classified as full- or part-time employees, including workers on paid vacations or other types of paid leave; salaried officers, executives, and staff members of incorporated firms; employees temporarily assigned to other units; and noncontract employees for whom the reporting unit is their permanent duty station regardless of whether that unit prepares their paychecks.

The OES survey includes all full- and part-time wage and salary workers in nonfarm industries. Self-employed workers, owners and partners in unincorporated firms, household workers, and unpaid family workers are excluded.

Occupations are classified based on work performed and on required skills. Employees are assigned to an occupation based on the work they perform and not on their education or training. For example, an employee trained as an engineer but working as a drafter is reported as a drafter. Employees who perform the duties of two or more occupations are reported in the occupation that requires the highest level of skill or in the occupation where the most time is spent if there is no measurable difference in skill requirements. ***Working supervisors*** (those spending 20 percent or more of their time doing work similar to the workers they supervise) are classified with the workers they supervise. ***Workers receiving on-the-job training, apprentices, and trainees*** are classified with the occupations for which they are being trained.

A ***wage*** is money that is paid or received for work or services performed in a specified period. Base rate pay, cost-of-living allowances, guaranteed pay, hazardous-duty pay, incentive pay such as commissions and production bonuses, and tips are included in a wage. Back pay, jury duty pay, overtime pay, severance pay, shift differentials, nonproduction bonuses, employer costs for supplementary benefits, and tuition reimbursements are excluded. Federal government, the U.S. Postal Service (USPS), and most state governments report individual wage rates for workers. Wage rates for other employers are placed into one of the 12 wage intervals below:

Interval	Wages	
	Hourly	Annual
Range A	Under $9.25	Under $19,240
Range B	$9.25 to $11.74	$19,240 to $24,439
Range C	$11.75 to $14.74	$24,440 to $30,679
Range D	$14.75 to $18.74	$30,680 to $38,999
Range E	$18.75 to $23.99	$39,000 to $49,919
Range F	$24.00 to $30.24	$49,920 to $62,919
Range G	$30.25 to $38.49	$62,920 to $80,079
Range H	$38.50 to $48.99	$80,080 to $101,919
Range I	$49.00 to $61.99	$101,920 to $128,959
Range J	$62.00 to $78.74	$128,960 to $163,799
Range K	$78.75 to $99.99	$163,800 to $207,999
Range L	$100.00 and over	$208,000 and over

3-year survey cycle of data collection

The survey is based on a probability sample drawn from a universe of about 7.7 million in-scope establishments stratified by geography, industry, size, and ownership. The sample is designed to represent all nonfarm establishments in the United States.

Semiannual samples are referred to as *panels*. The survey is conducted over a rolling 6-panel (or 3-year) cycle. This is done in order to provide adequate geographic, industrial, and occupational coverage. Over the course of a 6-panel (or 3-year) cycle, approximately 1.2 million establishments are sampled. In this cycle, data collected in May 2018 are combined with data collected in November 2017, May 2017, November 2016, May 2016, and November 2015.

The sample was reduced in recent panels. The November 2017 and May 2018 panels of the OES survey allocated and selected samples of approximately 186,000 establishments each. The May 2017 panel sample consisted of approximately 195,000 establishments, and the 3 prior panels allocated and selected samples of approximately 200,000 establishments semiannually. To the extent possible, private sector units selected in any one panel are not sampled again in the next five panels.

For a given panel, most sampled establishments initially receive either a survey questionnaire or instructions for reporting their data electronically. Nonrespondents receive up to three additional mailings and may be contacted by phone or email.

Censuses of federal and state government are collected annually.

- A census of the executive branch of the federal government and the U.S. Postal Service (USPS) is collected annually in June from the U.S. Office of Personnel Management (OPM), the Tennessee Valley Authority, and the U.S. Postal Service. Data from only the most recent year are retained for use in OES estimates.
- In each area, a census of state government establishments, except for schools and hospitals, is collected annually every November. Data from only the most recent year are retained for use in OES estimates.

A probability sample is taken of local government establishments, private sector establishments, and state schools and hospitals.

Sampling procedures

Frame construction

The sampling frame, or universe, is a list of about 7.7 million in-scope nonfarm establishments that file unemployment insurance (UI) reports to the state workforce agencies. Employers are required by law to file these reports to the state where each establishment is located. Every quarter, BLS creates a national sampling frame by combining the administrative lists of unemployment insurance reports from all of the states into a single database called the Quarterly Census of Employment and Wages (QCEW). Every six months, OES extracts the administrative data for establishments that are in scope for the OES survey from the most current QCEW. QCEW files were supplemented with frame files covering rail transportation (NAICS 4821) and Guam because these establishments are not covered by the UI program.

Construction of the sampling frame includes a process in which establishments that are linked together into multiunit companies are assigned to either the May or November sample. This prevents BLS from contacting multiunit companies more than once per year for this survey. Furthermore, the frame is matched to the 5 prior sample panels, and units that have been previously selected in the 5 prior panels are marked as ineligible for sampling for the current panel.

Stratification

Establishments on the frame are stratified by geographic area and industry group.

- Geography—606 Metropolitan Statistical Areas (MSAs) and nonmetropolitan or balance-of-state (BOS) areas are specified. MSAs are defined and mandated by the Office of Management and Budget. The May 2018 OES estimates use the July 2015 MSA definitions.

  As of the May 2018 estimates, OES no longer publishes data for metropolitan divisions. Data for the 11 large metropolitan areas that are further broken down into divisions are now published at the metropolitan area level only. In addition, some nonmetropolitan areas were combined to form larger nonmetropolitan areas in the May 2018 estimates. Details can be found at www.bls.gov/oes/areas_2018.htm.

  Each officially defined metropolitan area within a state is specified as a substate area. Cross-state MSAs have a separate portion for each state contributing to that MSA. In addition, states

may have up to six residual nonmetropolitan areas that together cover the remaining non-MSA portion of their state.

- Industry—302 industry groups are defined at the NAICS 3-, 4-, 5-, or 6-digit level.

- Ownership—Schools are also stratified by state government, local government, or private ownership. Also, local government casinos and gambling establishments are sampled separately from the rest of local government.

- Size—Establishments are divided into certainty and noncertainty size classes.

At any given time, there are about 151,000 nonempty State/MSA-BOS/NAICS 3-, 4-, 5-, 6-digit/ownership strata on the frame. When comparing nonempty strata between frames, there may be substantial frame-to-frame differences. The differences are due primarily to normal establishment birth and death processes and normal establishment growth and shrinkage. Other differences are due to establishment NAICS reclassification and changes in geographic location.

A small number of establishments indicate the state in which their employees are located, but do not indicate the specific county in which they are located. These establishments are also sampled and used in the calculation of the statewide and national estimates. They are not included in the estimates of any substate area. Therefore, the sum of the employment in the MSAs and nonmetropolitan areas within a state may be less than the statewide employment.

Allocation of the sample to strata

The frame is stratified into approximately 151,000 nonempty State/MSA-BOS/NAICS 3-, 4-, 5-, 6-digit/ownership strata. Each time a sample is selected, a 6-panel allocation of the 1.2 million sample units among these strata is performed. The largest establishments are removed from the allocation because they will be selected with certainty once during the 6-panel cycle. For the remaining noncertainty strata, a set of minimum sample size requirements based on the number of establishments in each cell is used to ensure coverage for industries and MSAs. For each State/MSA-BOS/NAICS 3-, 4-, 5-, 6-digit/ownership stratum, a sample allocation is calculated using a power Neyman allocation. The actual 6-panel sample allocation is the larger of the minimum sample allocation and the power allocation. To determine the current single panel allocation, the 6-panel allocation is divided by 6 and the resulting quotient is randomly rounded.

Two factors influence the power Neyman allocation. One is the square root of the employment size of each stratum. With a Neyman allocation, strata with higher levels of employment generally are allocated more sample than strata with lower levels of employment. Using the square root within the Neyman allocation softens this effect. The other is a measure of the occupational variability of the industry based on prior OES survey data. The occupational variability of an industry is measured by computing the coefficient of variation (CV) for each occupation within the 90th percentile of occupational employment in a given industry, averaging those CVs, and then calculating the standard error from that average CV. Using this measure, industries that tend to have greater occupational variability will get more sample than industries that are more occupationally homogeneous.

Sample selection

Sample selection within strata is approximately proportional to size. In order to provide the most occupational coverage, establishments with higher employment are more likely to be selected than those with lower employment; some of the largest establishments are selected with certainty. The unweighted employment of sampled establishments makes up approximately 57.7 percent of total employment.

Permanent random numbers (PRNs) are used in the sample selection process. To minimize sample overlap between the OES survey and other large surveys conducted by the U.S. Bureau of Labor Statistics, each establishment is assigned a PRN. For each stratum, a specific PRN value is designated as the "starting" point to select a sample. From this "starting" point, we sequentially select the first '*n'* eligible establishments in the frame into the sample, where '*n*' denotes the number of establishments to be sampled.

Single panel weights (sampling weights)

Sampling weights are computed so that each panel will roughly represent the entire universe of establishments.

Federal government, USPS, and state government units are assigned a panel weight of 1. Other sampled establishments are assigned a design-based panel weight, which reflects the inverse of the probability of selection.

National sample counts

The combined sample for the May 2018 survey is the equivalent of six panels. The sample allocations, excluding federal government and U.S. Postal Service (USPS), for the panels in this cycle are:

186,125 establishments for May 2018
185,450 establishments for November 2017
195,117 establishments for May 2017
201,952 establishments for November 2016
201,447 establishments for May 2016
202,772 establishments for November 2015

The May 2018 data include a census of 8,063 federal and USPS units. The combined sample size for the May 2018 estimates is approximately 1.2 million establishments, which includes only the most recent data for federal and state government. Federal and state government units from older panels are deleted to avoid double counting.

Response and nonresponse

Response

Of the approximately 1.2 million establishments in the 50 states and the District of Columbia in the combined initial sample, approximately 1,097,000 were viable establishments (that is, establishments that are not outside the scope or out of business). Of the viable establishments, approximately 781,000 responded and 316,000 did not—a 71.2 percent response rate. The response rate in terms of weighted sample employment is 67.9 percent.

Nonresponse

Nonresponse is a chronic problem in virtually all large-scale surveys because it may introduce a bias in estimates if the nonrespondents tend to differ from respondents in terms of the characteristic being measured. To partially compensate for nonresponse, the missing data for each nonrespondent are imputed using plausible data from responding units with similar characteristics.

Establishments that do not report occupational employment data are called “unit” nonrespondents. Establishments that report employment data but fail to report some or all the corresponding wages are called “partial” nonrespondents. Missing data for unit nonrespondents are imputed through a two-step

imputation process. Missing data for partial nonrespondents are imputed through the second step of the process only.

Step 1, Impute an occupational employment staffing pattern

For each unit nonrespondent, a staffing pattern is imputed using a nearest-neighbor "hot deck" imputation method. The procedure links a responding donor establishment to each nonrespondent. The nearest-neighbor hot deck procedure searches within defined cells for a donor that most closely resembles the nonrespondent by geographic area, industry, and employment size. Ownership is also used in the hospital, education, gambling, and casino hotel industries. The procedure initially searches for a donor whose reported employment is approximately the same as the nonrespondent's frame employment within the same 5- or 6-digit NAICS or NAICS aggregation, state, and ownership. If more than one otherwise equally qualified donor is found, a donor from a more recent panel will be selected over a donor from an older panel. If the search is unsuccessful, the pool of donors is enlarged in incremental steps by expanding geographic area and industry until a suitable donor is found. Limits are placed on the number of times a donor can be used.

After a donor has been found, its occupational staffing pattern is used to prorate the nonrespondent's frame employment by occupation. The prorated employment is the nonrespondent's imputed occupational employment.

Step 2, Impute an employment distribution across wage intervals

For each "unit" nonrespondent in step 1 or for each "partial" nonrespondent, impute an employment distribution across wage intervals for occupations without complete wage data. This distribution, called the wage employment distribution, is imputed as follows:

- Identify the imputation cell for each of the nonrespondent's occupations. Imputation cells are initially defined by MSA/BOS, NAICS 5/6 or NAICS aggregation, and size class from the most recent panel only. For schools, hospitals, gambling establishments, and casino hotels, cells are further divided by ownership.

- Determine if the imputation cell has enough respondents to compute wage employment distributions. If not, incrementally enlarge the cell until there are enough respondents.

- Use the distributions above to prorate the nonrespondent's imputed occupational employment across wage intervals. (Or, for partial respondents, use the distributions above to prorate the reported occupational employment across wage intervals.)

Special procedures

Within the past 3-year cycle, the OES had critical nonrespondents that could not be imputed using current OES methods. The OES employed special imputation procedures which used nonrespondents' prior staffing patterns. The occupational employment was benchmarked to the current year and the wage distribution was imputed using procedures very similar to the current partial imputation method.

Estimation methodology

This section describes the weighting methodology and formulas used for making the estimates. Each semiannual sample represents roughly one-sixth of the establishments for the full 6-panel sample plan and is used in conjunction with the previous five semiannual samples in order to create a combined sample of approximately 1.2 million establishments, which includes only the most recent data for federal and state government.

Reweighting for the combined sample

Employment and wage rate estimates are computed using a rolling 6-panel (3-year) sample. The May 2018 estimates were calculated using data from the May 2018, November 2017, May 2017, November 2016, May 2016, and November 2015 samples. Establishments from each panel's sample are initially assigned weights as if one panel were being used to represent the entire population. When the samples are combined, each sampled establishment must be reweighted so that the aggregated sample across six panels represents the entire population. Establishments selected with certainty in the 6-panel cycle are given a weight equal to 1. Noncertainty units are reweighted stratum by stratum. This revised weight is called the 6-panel combined sample weight. The original single-panel sampling weights are computed so that responses in a stratum could be weighted to represent the entire stratum population. In one common scenario, six panel samples are combined, and all six panels have sample units for a particular stratum. A summation of the single-panel weights would over-represent the population by a factor of six. Because we do not want to over-represent the stratum population, the 6-panel combined sample weight of each establishment is set equal to 1/k times its single-panel sampling weight. In general, when six panel samples are combined, a count of the number of panels with at least one unit selected for a given stratum is assigned to k.

Special procedures

The 2012 NAICS was used to define sampling cells and the 2017 NAICS was used to define estimation cells for panel data from November 2015 to May 2017. Sample weights for these panels were recomputed to reflect the update to 2017 NAICS codes. Sampling and data collection for the November 2017 and May 2018 panels used the 2017 NAICS.

Benchmarking to QCEW employment

A sum of ratio-adjusted weighted reported occupational employment is used to calculate estimates of occupational employment. The auxiliary variable for the estimator is the average of the latest May and November employment totals from the Bureau's Quarterly Census of Employment and Wages (QCEW). For the May 2018 estimates, the auxiliary variable is the average of May 2018 and November 2017 employment. In order to balance the states' need for estimates at differing levels of geography and industry, the ratio estimation process is carried out through a series of four hierarchical employment ratio adjustments. The ratio adjustments are also known as benchmark factors (BMFs).

The first of the hierarchical benchmark factors is calculated for cells defined by state, MSA/BOS, NAICS 3/4/5/6, and employment size class (4 size classes: 1-19, 20-49, 50-249, 250+). For establishments in the hospital and education industries (NAICS 622 and 611), the first hierarchical factor is calculated for cells defined by state, MSA/BOS, NAICS 3/4/5/6, employment size class (4 size classes: 1-19, 20-49, 50-249, 250+), and ownership (state government, local government, or privately owned). If a first-level BMF is out of range, it is reset to a maximum (ceiling) or minimum (floor) value. First-level BMFs are calculated as follows:

h = MSA/BOS by NAICS 3/4/5/6

H = state by NAICS 3/4/5/6

s = employment size classes (1-19, 20-49, 50-249, 250+)

S = aggregated employment size classes (1-49, 50+)

o = ownership (state government, local government, or privately owned)

M = average of May and November QCEW employment

w_i = six-panel combined sample weight for establishment i

x_i = total establishment employment

BMF_{min} = a parameter, the lowest value allowed for BMF

BMF_{max} = a parameter, the highest value allowed for BMF

$$\beta_{hs} = \left(M_{hs} \Big/ \sum_{i \in hs} w_i x_i \right), \qquad \beta_{hS} = \left(M_{hS} \Big/ \sum_{i \in hS} w_i x_i \right), \qquad \beta_{h} = \left(M_{h} \Big/ \sum_{i \in h} w_i x_i \right)$$

$$\beta_{hso} = \left(M_{hso} \Big/ \sum_{i \in hso} w_i x_i \right), \quad \beta_{hSo} = \left(M_{hSo} \Big/ \sum_{i \in hSo} w_i x_i \right), \quad \beta_{ho} = \left(M_{ho} \Big/ \sum_{i \in ho} w_i x_i \right), \text{ then}$$

$$BMF_{1,\,hs} = \begin{cases} \beta_{hso}, \text{ if all } \beta_{hso} \text{ within } h \text{ are bounded by } (\text{BMF}_{\min}, \text{BMF}_{\max}), \\ \beta_{hs}, \text{ if all } \beta_{hs} \text{ within } h \text{ are bounded by } (\text{BMF}_{\min}, \text{BMF}_{\max}), \\ \beta_{hSo}, \text{ if all } \beta_{hSo} \text{ within } h \text{ are bounded by } (\text{BMF}_{\min}, \text{BMF}_{\max}), \\ \beta_{hS}, \text{ if all } \beta_{hS} \text{ within } h \text{ are bounded by } (\text{BMF}_{\min}, \text{BMF}_{\max}), \\ \beta_{ho}, \text{ if all } \beta_{ho} \text{ within } h \text{ are bounded by } (\text{BMF}_{\min}, \text{BMF}_{\max}), \\ \beta_{h}, \text{ if all } \beta_{h} \text{ within } h \text{ are bounded by } (\text{BMF}_{\min}, \text{BMF}_{\max}), \\ \text{BMF}_{\min}, \text{ if } \beta_{h} < \text{BMF}_{\min}, \\ \text{BMF}_{\max}, \text{ if } \beta_{h} > \text{BMF}_{\max} \end{cases}$$

Second-level BMFs are calculated for cells defined at the state, NAICS 3/4/5/6 level by summing the product of combined 6-panel weight and first-level BMF for each establishment in the cell. For establishments in the hospital, education, gambling, and casino hotel industries (NAICS 622, 611, 7132 and 72112), the first hierarchical of the second-level BMK factor is calculated at the state, NAICS 3/4/5/6, and ownership level. Second-level BMFs account for the portion of universe employment that is not adequately covered by weighted employment in first-level benchmarking. Inadequate coverage occurs when "MSA/BOS | NAICS 3/4/5/6 | size class" cells have no sample data *or* when a floor or ceiling is imposed on first-level BMFs. Second-level benchmarks are calculated as follows:

$$\beta_{Ho} = \left(M_{Ho} \Big/ \sum_{hs \in H} \sum_{i \in hs} w_i x_i BMF_{1,\,hs} \right)$$

$$\beta_{H} = \left(M_{H} \Big/ \sum_{hs \in H} \sum_{i \in hs} w_i x_i BMF_{1,\,hs} \right), \text{ then}$$

$$BMF_{2,\,H} = \begin{cases} \beta_{Ho}\,, \text{ if all } \beta_{Ho} \text{ within } H \text{ are bounded by } \left(\text{BMF}_{\min}\,, \text{BMF}_{\max}\right), \\ \beta_{H}\,, \text{ if all } \beta_{H} \text{ within } H \text{ are bounded by } \left(\text{BMF}_{\min}\,, \text{BMF}_{\max}\right), \\ \text{BMF}_{\min}, \text{ if } \beta_{H} < \text{BMF}_{\min}, \\ \text{BMF}_{\max}, \text{ if } \beta_{H} > \text{BMF}_{\max} \end{cases}$$

Third-level BMFs ($BMF_{3,H}$) are calculated at the "State | 3-digit NAICS" cell level by summing the product of combined 6-panel weight, first-level BMF, and second-level BMF for each establishment in the cell. The third-level BMF also benchmarks by ownership for the hospital, education, gambling, and casino hotel industries. Fourth-level BMFs ($BMF_{4,H}$) are calculated at the "State | 2-digit NAICS" cell level by summing the product of final weight, first-level BMF, second-level BMF, and third-level BMF for each establishment in the cell. The fourth-level BMK factor does not benchmark by ownership. As with second-level BMFs, third- and fourth-level BMFs are computed to account for inadequate coverage of the universe employment.

A final benchmark factor, BMF_i, is calculated for each establishment as the product of its four hierarchical benchmark factors ($BMF_i = BMF_1 * BMF_2 * BMF_3 * BMF_4$). A benchmark weight value is then calculated as the product of the establishment's six-panel combined sample weight and final benchmark factor.

Occupational employment estimates

Benchmark factors and the combined 6-panel weights are used to compute estimates of occupational employment. Estimates are produced for cells defined by geographic area and industry group. The total employment for an occupation in a cell is estimated by taking the product of the reported occupational employment, the 6-panel combined sample weight, and the final benchmark factor for each establishment in the cell, and summing the product across all establishments in the cell. This sum is the estimate of total occupational employment in the cell.

The equation below is used to calculate occupational employment estimates for an estimation cell defined by geographic area, industry group, and size class.

$$\hat{X}_{ho} = \sum_{i \in h} \left(w_i \; BMF_i \; x_{io} \right)$$

o = occupation

h = estimation cell

w_i = six-panel combined sample weight for establishment *i*

BMF_i = final benchmark factor for establishment *i*

x_{io} = employment for occupation *o* in establishment *i*

$\hat{X}_{ho}$ = estimated employment for occupation *o* in cell *h*

Wage rate estimation

Two externally derived parameters are used to calculate wage rate estimates. They are:

- the mean wage rates for each of the 12 wage intervals and
- wage updating factors (also known as aging factors)

Wage rates of workers are converted to one of 12 consecutive, nonoverlapping wage bands. Individual wage rates are used for federal government and U.S. Postal Service workers. State governments may report their data as either individual wage rates or interval wage rates.

An illustration

An establishment employs 10 secretaries at the following wage rates:

$9/hour — 1 secretary
$10/hour — 1 secretary
$12/hour — 2 secretaries
$13/hour — 2 secretaries
$14/hour — 2 secretaries
$16/hour — 1 secretary
$17/hour — 1 secretary

Wage rates for secretaries, however, are used in the OES survey as follows:

Wage interval A (under $9.25/hour) — 1 secretary
Wage interval B ($9.25-$11.74/hour) — 1 secretary
Wage interval C ($11.75-$14.74/hour) — 6 secretaries
Wage interval D ($14.75-$18.74/hour) — 2 secretaries

The remaining wage intervals have 0 secretaries.

Because wage rates are grouped into intervals, we must use grouped data formulas to calculate estimates of mean and percentile wage rates. Assumptions are made when using grouped data formulas. For the mean wage rate formula, we assume that we can calculate the average wage rate for workers in each interval. For the percentile wage rate formula, we assume that workers are evenly distributed in each interval.

Wage data from the May 2018, November 2017, May 2017, November 2016, May 2016, and November 2015 panels were used to calculate May 2018 wage rate estimates. Wage data from different panels, however, are not equivalent in real-dollar terms due to inflation and changing compensation costs. Consequently, wage data collected prior to the current survey reference period have to be updated or aged to approximate that period.

Determining a mean wage rate for each interval

The mean hourly wage rate for all workers in any given wage interval cannot be computed using grouped data collected by the OES survey. This value is calculated externally using data from the Bureau's National Compensation Survey (NCS). With the exception of the highest wage interval, mean wage rates for each panel are calculated using the most recent NCS data available. The hourly mean wage rate of the highest wage interval is calculated differently from the others. A weighted average of the previous three years' means is used, instead of just the current year's mean. Note that the mean hourly wage rate for interval L (the upper, open-ended wage interval) is calculated without wage data for pilots. This occupation is excluded because pilots work fewer hours than workers in other occupations.

Wage aging process

Aging factors are developed from the Bureau's Employment Cost Index (ECI) survey. The ECI survey measures the rate of change in wages and salaries for ten major occupational groups on a quarterly basis. Aging factors are used to adjust OES wage data from past survey reference periods to the current survey reference period (May 2018). The procedure assumes that there are no major differences by geography, industry, or detailed occupation within the occupational division. The twelfth, open-ended, interval is not aged.

Mean hourly wage rate estimates

For data from private sector, local government, and certain state government establishments, the mean hourly wage is calculated as the total weighted hourly wages for an occupation divided by its weighted

survey employment. Estimates of mean hourly wages are calculated using a standard grouped data formula that was modified to use ECI aging factors.

$$\hat{R}_o = \frac{\sum_{z=t-5}^{t}\left(\sum_{i \in z} w_i \; BMF_i \; \hat{y}_{i\,o}\right)}{\hat{X}_o}$$

$$\hat{y}_{i\,o} = u_{z\,o}\sum_{r} x_{i\,o\,r}\,c_{z\,r} \qquad (i \in z)$$

o = occupation

$\hat{R}_o$ = mean hourly wage rate for occupation o

z = panel (or year)

t = current panel

w_i = six-panel combined sample weight for establishment i

BMF_i = final benchmark factor applied to establishment i

$\hat{y}_{i\,o}$ = unweighted total hourly wage estimate for occupation o in establishment i

r = wage interval

$\hat{X}_o$ = estimated employment for occupation o

$x_{i\,o\,r}$ = reported employment for occupation o in establishment i in wage interval r (note that establishment i reports data for only one panel z or one year z)

$u_{z\,o}$ = ECI aging factor for panel (or year) z and occupation o

$c_{z\,r}$ = mean hourly wage for interval r in panel (or year) z

In this formula, $c_{z\,r}$ represents the mean hourly wage of interval r in panel (or year) z. The mean is computed externally using data from the Bureau's NCS survey.

For wage rate data from federal and certain state government establishments, the hourly wages for an occupation within an establishment are summed to get total wages. Employment for that occupation within that establishment is also summed to get total employment. The total wages and total employment across all establishments in the occupation for the estimation level of interest are summed.

$$\text{Mean Wage} = \frac{\text{Total Interval Wages + Total Individual Wages}}{\text{Total Interval Employment + Total Individual Employment}}$$

Percentile hourly wage rate estimates

The p-th percentile hourly wage rate for an occupation is the wage where p percent of all workers earn that amount or less and where (100-p) percent of all workers earn that amount or more. The wage interval containing the p-th percentile hourly wage rate is located using a cumulative frequency count of estimated employment across all wage intervals. After the targeted wage interval is identified, the p-th percentile wage rate is then estimated using a linear interpolation procedure. This statistic is calculated by first distributing federal, state, local government, and private sector workers inside each wage interval. Federal and certain state government workers are distributed throughout the wage intervals according to their wage rates, while certain state government, local government, and private sector workers are distributed uniformly within each wage interval. Next, workers are ranked from lowest paid to highest paid. Finally, the product of the total employment for the occupation and the desired percentile is calculated to determine the worker that earns the p-th percentile wage rate.

$$pR_o = L_r + \frac{j}{f_r}(U_r - L_r)$$

pR_o = p-th percentile hourly wage rate for occupation o

r = wage interval that encompasses pR_o

L_r = lower bound of wage interval r

U_r = upper bound of wage interval r

f_r = number of workers in interval r

j = difference between the number of workers needed to reach the p-th percentile wage rate and the number of workers needed to reach the L_r wage rate

Annual wage rate estimates

These estimates are calculated by multiplying mean or percentile hourly wage rate estimates by a "year-round, full time" figure of 2,080 hours (52 weeks x 40 hours) per year. These estimates, however, may not represent mean annual pay should the workers work more or less than 2,080 hours per year.

Alternatively, some workers are paid based on an annual basis but do not work the usual 2,080 hours per year. For these workers, survey respondents report annual wages. Since the survey does not collect the actual number of hours worked, hourly wage rates cannot be derived from annual wage rates with any reasonable degree of confidence. Only annual wages are reported for some occupations.

Variance estimation

Occupational employment variance estimation

A subsample replication technique called the "jackknife random group" is used to estimate variances of occupational employment. In this technique, each sampled establishment is assigned to one of G random groups. G subsamples are created from the G random groups. Each subsample is reweighted to represent the universe.

G estimates of total occupational employment ($\hat{X}_{hjog}$) (one estimate per subsample) are calculated. The variability among the G employment estimates is a good variance estimate for occupational employment. The two formulas below are used to estimate the variance of occupational employment for an estimation cell defined by geographic area and industry group.

$$v(\hat{X}_{hjo}) = \frac{\sum_{g=1}^{G} (\hat{X}_{hjog} - \hat{\bar{X}}_{hjo})^2}{G(G-1)}$$

h	=	estimation cell defined by geographic area and industry group
j	=	employment size class (1-19, 20-49, 50-249, 250+)
o	=	occupation
$v(\hat{X}_{hjo})$	=	estimated variance of $\hat{X}_{hjo}$
G	=	number of random groups
$\hat{X}_{hjo}$	=	estimated employment of occupation o in cell h and size class j
$\hat{X}_{hjog}$	=	estimated employment of occupation o in cell h, size class j, and subsample g
$\hat{\bar{X}}_{hjo}$	=	estimated mean employment for occupation o in cell h and size class j based on the G subsamples (Note: a finite population correction factor is applied to the terms $\hat{X}_{hjog}$ and $\hat{\bar{X}}_{hjo}$.)

The variance for an occupational employment estimate in cell h is obtained by summing the variances $v(\hat{X}_{hjo})$ across all size classes j in the cell.

$$v(\hat{X}_{ho}) = \sum_{j \in h} v(\hat{X}_{hjo})$$

Occupational mean wage variance estimates

Because the OES wage data are placed into intervals (grouped), the exact wage of each worker is not used. Therefore, some components of the wage variance are approximated using factors developed from NCS data. A *Taylor Series Linearization* technique is used to develop a variance estimator appropriate for OES mean wage estimates. The primary component of the mean wage variance, which accounts for the variability of the observed sample data, is estimated using the standard estimator of variance for a ratio estimate. This component is the first term in the formula given below:

$$v(\hat{R}_o) = \left(\begin{array}{l} \dfrac{1}{\hat{X}_o^2} \left(\sum_h \left\{ \dfrac{n_{ho}\left(1 - f_{ho}\right)}{n_{ho} - 1} \right\} \left\{ \sum_{i \in h} (BMK_i w_i)^2 \left(q_{io} - \overline{q}_{ho}\right)^2 \right\} \right) + \\ \sum_r \theta_{or}^2 \sigma_{cr}^2 + \dfrac{1}{\hat{X}_o^2} \sum_r \left(\sum_{i=1}^{n_o} \left(BMK_i w_i x_{ior}\right)^2 \right) \sigma_{er}^2 + \dfrac{1}{\hat{X}_o} \sum_r \theta_{or} \sigma_{\omega r}^2 \end{array} \right)$$

$\hat{R}_o$ = estimated mean wage for occupation *o*

$v(\hat{R}_o)$ = estimated variance of $\hat{R}_o$

$\hat{X}_o$ = estimated occupational employment for occupation *o*

h = stratum (area/industry/size class)

f_{ho} = sampling fraction for occupation *o* in stratum *h*

n_{ho} = number of sampled establishments that reported occupation *o* in stratum *h*

w_i = six-panel combined sample weight for establishment *i*

BMF_i = final benchmark factor applied to establishment *i*

q_{io} = $\left(\hat{y}_{io} - \hat{R}_o x_{io}\right)$ for occupation *o* in establishment *i*

$\hat{y}_{io}$ = estimated total occupational wage in establishment *i* for occupation *o*

x_{io} = reported employment in establishment *i* for occupation *o*

$\overline{q}_{ho}$ = mean of the q_{io} quantities for occupation *o* in stratum *h*

θ_{or} = proportion of employment within interval *r* for occupation *o*

x_{ior} = reported employment in establishment *i* within wage interval *r* for occupation *o*

$\left(\sigma_{cr}^2 , \sigma_{er}^2 , and\ \sigma_{\omega r}^2\right)$ Within wage interval *r*, these are estimated using the NCS and, respectively, represent the variability of the wage value imputed to each worker, the variability of wages across establishments, and the variability of wages within establishments.

Reliability of the estimates

Estimates developed from a sample will differ from the results of a census. An estimate based on a sample survey is subject to two types of error: sampling and nonsampling error. An estimate based on a census is subject only to nonsampling error.

Nonsampling error

This type of error is attributable to several causes, such as errors in the sampling frame; an inability to obtain information for all establishments in the sample; differences in respondents' interpretation of a survey question; an inability or unwillingness of the respondents to provide correct information; errors made in recording, coding, or processing the data; and errors made in imputing values for missing data. Explicit measures of the effects of nonsampling error are not available.

Sampling error

When a sample, rather than an entire population, is surveyed, estimates differ from the true population values that they represent. This difference, the sampling error, occurs by chance and its variability is measured by the variance of the estimate or the standard error of the estimate (square root of the variance). The relative standard error is the ratio of the standard error to the estimate itself.

Estimates of the sampling error for occupational employment and mean wage rates are provided for all employment and mean wage estimates to allow data users to determine if those statistics are reliable enough for their needs. Only a probability-based sample can be used to calculate estimates of sampling error. The formulas used to estimate OES variances are adaptations of formulas appropriate for the survey design used.

The particular sample used in this survey is one of a large number of many possible samples of the same size that could have been selected using the same sample design. Sample estimates from a given design are said to be unbiased when an average of the estimates from all possible samples yields the true population value. In this case, the sample estimate and its standard error can be used to construct confidence intervals, or ranges of values that include the true population value with known probabilities. To illustrate, if the process of selecting a sample from the population were repeated many times, if each sample were surveyed under essentially the same unbiased conditions, and if an estimate and a suitable estimate of its standard error were made from each sample, then:

1. Approximately 68 percent of the intervals from one standard error below to one standard error above the estimate would include the true population value. This interval is called a 68-percent confidence interval.

2. Approximately 90 percent of the intervals from 1.6 standard errors below to 1.6 standard errors above the estimate would include the true population value. This interval is called a 90-percent confidence interval.

3. Approximately 95 percent of the intervals from 2 standard errors below to 2 standard errors above the estimate would include the true population value. This interval is called the 95-percent confidence interval.

4. Almost all (99.7 percent) of the intervals from 3 standard errors below to 3 standard errors above the estimate would include the true population value.

For example, suppose that an estimated occupational employment total is 5,000, with an associated estimate of relative standard error of 2.0 percent. Based on these data, the standard error of the estimate is 100 (2 percent of 5,000). To construct a 90-percent confidence interval, add and subtract 160 (1.6 times the standard error) from the estimate: (4,840; 5,160). Approximately 90 percent of the intervals constructed in this manner will include the true occupational employment if survey methods are nearly unbiased.

Estimated standard errors should be taken to indicate the magnitude of sampling error only. They are not intended to measure nonsampling error, including any biases in the data. Particular care should be

exercised in the interpretation of small estimates or of small differences between estimates when the sampling error is relatively large or the magnitude of the bias is unknown.

Quality control measures

Several edit and quality control procedures are used to reduce nonsampling error. For example, completed survey questionnaires are checked for data consistency. Follow-up mailings, emails, and phone calls are sent out to nonresponding establishments to improve the survey response rate.

The OES survey is a federal-state cooperative effort that enables states to conduct their own surveys. A major concern with a cooperative program such as OES is to accommodate the needs of BLS and other federal agencies, as well as state-specific publication needs, with limited resources while simultaneously standardizing survey procedures across all 50 states, the District of Columbia, and the U.S. territories. Controlling sources of nonsampling error in this decentralized environment can be difficult. One important computerized quality control tool used by the OES survey is the Survey Processing and Management system. It was developed to provide a consistent and automated framework for survey processing and to reduce the workload for analysts at the state, regional, and national levels.

To ensure standardized sampling methods in all areas, the sample is drawn in the national office. Standardizing data processing activities, such as validating the sampling frame, allocating and selecting the sample, refining mailing addresses, addressing envelopes and mailers, editing and updating questionnaires, conducting electronic review, producing management reports, and calculating employment estimates, have resulted in the overall standardization of the OES survey methodology. This has reduced the number of errors on the data files as well as the time needed to review them.

Other quality control measures used in the OES survey include:

- Follow-up mail and telephone solicitations of nonrespondents, especially critical or large nonrespondents
- Review of data during collection to verify its accuracy and reasonableness
- Adjustments for atypical reporting units on the data file
- Validation of the benchmark employment figures and of the benchmark factors
- Validation of the analytical tables of estimates at the NAICS 3/4/5/6 level

Confidentiality

BLS has a strict confidentiality policy that ensures that the survey sample composition, lists of reporters, and names of respondents will be kept confidential. Additionally, the policy assures respondents that published figures will not reveal the identity of any specific respondent and will not allow the data of any specific respondent to be inferred. The most relevant statute which governs BLS confidentiality is the Confidential Information Protection and Statistical Efficiency Act (CIPSEA). Each published estimate is screened to ensure that it meets these confidentiality requirements. To further protect the confidentiality of the data, the specific screening criteria are not listed in this publication. For additional information regarding confidentiality, please visit the BLS website at www.bls.gov/bls/confidentiality.htm.

Data presentation

OES data are available in several formats from the OES home page at www.bls.gov/oes/. The OES database search tool (www.bls.gov/oes/data.htm) allows customers to create customized HTML or Excel tables using the most recent OES estimates. OES data are also published as HTML tables or can be downloaded as zipped XLS files at www.bls.gov/oes/tables.htm. Included are cross-industry data for the United States as a whole, for individual U.S. states, and for metropolitan and nonmetropolitan areas, along with U.S. industry-specific estimates by 2-, 3-, most 4-, and some 5- and 6-digit NAICS levels. A research dataset of OES estimates by state and industry is available at www.bls.gov/oes/current/oes_research_estimates.htm. BLS does not publish OES estimates by metropolitan/nonmetropolitan area and industry, but these data may be available from individual state workforce agencies: www.bls.gov/bls/ofolist.htm. Available data elements include estimates of employment, hourly and annual mean wages, and hourly and annual percentile wages by occupation, as well as relative standard errors (RSEs) for the employment and mean wage estimates.

When updated estimates become available, a BLS news release makes an announcement featuring highlights from the data. For additional information, contact the OES staff at (202) 691-6569 or send e-mail to oesinfo@bls.gov.

Uses

For many years, the OES survey has been a major source of detailed occupational employment data for the nation, states, and areas, and by industry at the national level. This survey provides information for many data users, including individuals and organizations engaged in planning vocational education programs, higher education programs, and employment and training programs. OES data also are used to prepare information for career counseling, for job placement activities performed at state workforce

agencies, and for personnel planning and market research conducted by private enterprises. OES data also are used by the Department of Labor's Foreign Labor Certification (FLC) program, which sets the rate at which workers on work visas in the United States must be paid.