Overview of the Survey of Occupational Injuries and Illnesses Sample Design and Estimation Methodology

Philip N. Selby, Terry M. Burdette, Erin M. Huband
Bureau of Labor Statistics, 2 Massachusetts Ave NE Room 3160, Washington, DC 20212

Abstract

The Survey of Occupational Injuries and Illnesses (SOII) uses a stratified sample design to produce State and National estimates for nonfatal workplace injuries and illnesses. Private industry estimates are produced separately for 42 States, the District of Columbia, and three U.S. territories (Guam, Puerto Rico and the Virgin Islands). The level of industry detail for which State estimates are available varies widely and is based on the needs determined by each State. Additionally, estimates for injuries and illnesses for State and local government workers are available for 26 of these States. Future plans will expand the coverage to national estimates for the State and local governments beginning with data released in 2009. This paper describes the frame development, stratification criteria, sample allocation, sample selection, and estimation methodologies used to produce the number and frequency (incidence rates) of nonfatal workplace injuries and illnesses.

Key Words: stratified sample, target estimation industry

1. Background of the Survey

The Bureau of Labor Statistics Survey of Occupational Injuries and Illnesses (SOII) gathers information on the number and frequency of injuries and illnesses experienced by America's workforce.

The current survey evolved from annual BLS surveys first conducted in the 1940's. For the next three decades, those nationwide surveys proved useful but with two major limitations: (1) represented employers recorded and reported worker injuries on a voluntary basis; (2) injuries were limited to death, permanent impairment, or temporary disability. Congress addressed these limitations and passed a landmark piece of legislation, *The Occupational Safety and Health Act of 1970*, implementing regulations requiring most private industry employers to maintain records (logs) and prepare reports on work-related injuries and illnesses. Participation in the survey was no longer voluntary.

Since then, the survey has undergone various changes to meet the record keeping regulations set by the Occupation Safety and Health Administration (OSHA), U.S. Department of Labor. Details of these regulations, both old and new, are available from the OSHA internet site (http://www.osha.gov/recordkeeping/index.html).

The SOII is a Federal/State cooperative program, in which the Federal government and participating States share the costs of participating State data collection activities. State participation in the survey may vary by year. Sample sizes are determined by the participating States based on budget constraints and independent samples are selected for each State annually. Data are collected by BLS regional offices for non-participating States. For 2006 survey data released in calendar year 2007, private sector injury and illness data were tabulated separately for 42 States and the District of Columbia participating in the program. Estimates for State and local government were produced for 26 of these States. There were eight states which did not participate in the program. For these States, a smaller sample is selected to provide data for national estimates only. Additionally, estimates are tabulated for three U.S. territories- Guam, Puerto Rico, and the Virgin Islands- but data from these territories are not included in the tabulation of national estimates.

Establishments responding to the SOII provide two sets of information – annual counts of injury and illness cases and detailed case and demographic data for individual cases involving one or more days away from work. The latter data include information on the worker injured, the nature of the disabling condition, and the event and source producing that condition. BLS publishes selected national estimates for private industry in two news releases: (1) a summary of counts and incidence rates in the private industry; (2) a more detailed release on the characteristics of injury and illness cases that involved days away from work.

A third release from the BLS reports on workplace fatalities in the Census of Fatal Occupational Injuries. However, given that this is a census, no discussion of its methodology will be included here. For more information on workplace injuries and illnesses, please visit http://www.bls.gov/iif/.

2. Frame Development

Because the SOII is a Federal-State cooperative program and data must meet the needs of participating State agencies, an independent sample is selected for each State. The survey covers the entire private sector except self-employed persons, private households, and small farms. Establishments with as few as one employee are sampled in all industries except agriculture production, where the minimum size is eleven employees. Mining and Railroad industries are not covered as part of the sampling process. These data are furnished directly from the Mine Safety and Health Administration and the Federal Railroad Administration, respectively, and used to produce State and national level estimates. The United States Postal Service and federal government employees are out-of-scope for the survey.

Although State and local government agencies are not currently surveyed for national estimates, several States have legislation which enables them to collect these data and publish State-level estimates. Beginning with the 2008 survey, to be published in calendar year 2009, State and local government estimates will be published at the national level and for most States.

The main source for the SOII sampling frame is the BLS Quarterly Census of Employment and Wages (QCEW). The QCEW is a near quarterly census of employers collecting employment and wages by ownership, county, and six-digit North American Industry Classification System (NAICS) code. States who survey State and local government units have an option to either use the QCEW or supply public sector sampling frames.

Data are collected on injuries and illnesses for the entire calendar year. Samples are allocated and selected from the latest available frame information from the QCEW. For example, the sample for survey year 2009 was selected in July 2008 using second quarter 2007 data from the QCEW. Once the sample is selected, BLS and State staff ensure address information is correct. In December 2008, selected establishments are notified that they are in the 2009 sample and asked to keep records.

3. Stratification Criteria

The SOII utilizes a stratified sample design with varying degrees of industry stratification levels within each State. This is desirable because some industries are more prevalent in some States compared to others. Also, some industries can be relatively small in employment but have high injury and illness rates which make them likely to be designated for estimation. Thus, States determine which industries are most important in terms of publication and the extent of industry stratification is set independently within each State. This was deemed a very important part of the SOII design because the industry structure can vary widely from State to State. The term Target Estimation Industries (TEIs) is used to describe the specific industries that States request for publication. Thus, a TEI is a NAICS industry, or group of NAICS industries, for which a State plans to produce an estimate. A State may set TEIs at different levels than other States. The TEIs are used to set the level of stratification within the States. A State's ability to actually publish these estimates depends on the reliability of the estimates and confidentiality requirements.

There are two types of TEIs: Publishable TEIs as mentioned above and Residual TEIs. Residual TEIs are groupings of industries for which the State does not wish to publish estimates but will be grouped with other TEIs for aggregate State estimates.

One distinct rule for setting TEIs is that each NAICS industry must be accounted for in one and only one TEI. As an example, within NAICS 11 (Agriculture, Forestry, Fishing and Hunting), a State may want to publish estimates for NAICS 111 (Crop Production), and NAICS 112 (Animal Production) but not for the remaining NAICS 113 (Forestry and Logging), 114 (Fishing, Hunting and Trapping), and 115 (Support Activities for Agriculture and Forestry). In this case a separate stratum would be formed to represent all of the units in NAICS 113, 114, and 115. The final product of the TEI setting process is that specific industries are identified by each State they desire to publish separately.

3.1 Additional Rules for Setting TEIs

There are additional rules in place for setting TEIs in order to produce desirable industry level estimates at both the National and State levels. National level TEIs (NTEIs) are set at various NAICS levels that target a desirable industry level estimate. Two additional constraints are that States must select a minimum number of TEIs and cannot exceed a maximum number of TEIs. The minimum constraint is required to support national estimates to which State data contribute. In general, minimum TEIs are set at the two-digit industry level for all major industries. As an example, at minimum, a State would need to set a TEI at the NAICS 11 industry level in order to support the national estimate at the same level.

States cannot exceed a predetermined maximum number of TEIs for private industry. This maximum limit is determined using the State's sample size. Establishing a maximum number of TEIs helps to ensure that the sample will be sufficient to provide reliable estimates for each TEI. The maximum threshold includes both publishable TEIs and residual TEIs.

3.2 Importance of TEI selection

Targeting appropriate State industry levels for publication is a vital step in producing accurate and reliable estimates. Setting State TEIs too broadly may not provide the necessary industry detail a data user is looking for or needs. Setting TEIs at a more detailed level in one industry that doesn't support publication will subtract from sample in other industries where more detailed levels could be published.

Some considerations that States take into account when setting TEI levels are: (1) What industries are important to their State? (2) Who are the main data users and what level of detail is required to meet their needs? (3) Is one industry more prone to injuries and illnesses than other industries? States also look at what TEI levels have been published in previous years. If detailed levels have not been published due to reliability or confidentiality issues in the past, setting the TEI at a less detailed level may allow the estimate to be published. It also improves the efficiency of the sample design. Decreasing the number of TEIs increases the number of sample units in the remaining TEIs. A higher sampling rate within a TEI results in more reliable estimates. States are encouraged to review TEI settings each year to improve reliability and maximize publication of their estimates.

TEI selection also impacts the survey sample allocation. TEIs are the cornerstone to the allocation process as the sampling cells are based on the TEIs set by the States. A sampling cell is defined by ownership x TEI x employment size class. Ownership is made up of three divisions: State Government, Local Government and private industry. Size classes are based on an establishment's average annual employment and defined as follows:

Size Class	Average Annual Employment
1	10 or less
2	11 – 49
3	50 – 249
4	250 – 999
5	1,000 or greater

4. Sample Allocation and Selection

The sample allocation and selection system was completely overhauled in 2003 along with the conversion from Standard Industrial Classification (SIC) coding to NAICS coding based estimates. Three major goals were achieved with the new system: (1) moving from the mainframe to Unix environment; (2) greater flexibility to handle sample design changes where States may move from a non-participating State to a participating State and vice-versa; and (3) simpler and more efficient allocation module.

The former system used a complex allocation method which took into account many variables including lost workday cases (LWDC) data, target sampling errors, and frame sizes. A Neyman's formula for a fixed variance was used to

calculate sample sizes for each sampling cell. Although this worked fairly well, it created inefficiently large samples in low hazard industries.

The need to move from SIC to NAICS based estimates provided an opportunity to improve the allocation module. While the new process was being developed, the old processing system was modified to handle NAICS-based strata for the 2003 and 2004 samples. Initial NAICS-based SOII estimates were first available in October 2004 from the 2003 survey. Thus, the implementation of the allocation module was delayed until NAICS-based data was available for application to the survey design. The new allocation module was introduced beginning with the 2006 survey which was selected in October 2005.

Because historical NAICS data was unavailable, an optimal allocation procedure was proposed which attempts to distribute the sample to the industries in a manner that minimizes the variance of the estimates. This method provides a smaller sample size in cells where units have similar incident rates and a larger sample size in cells where units have more variable rates.

Research was done to determine what measure of size was most appropriate for the allocation module. With the trend of establishments in the SOII going to more restricted work activity for employees that are injured on the job, the choices were narrowed down to the following: (1) Days Away from Work cases (DAFW); (2) Days of job transfer or restriction (DJTR); and (3) Total Recordable Cases (TRC). Rates from the 2003 SOII were studied for all 1251 TEIs for each of the above three categories. The average case rate, standard deviation, and coefficient of variation (CV) for each set of rates were calculated. The CV is the standard deviation divided by the estimate and is commonly used to compare estimates in relative terms. The CV for the TRC was lowest which led to the recommendation of using the TRC rate by size class for the measure of size input for the allocation module. Because TRC includes both DAFW and DJTR cases, it is the most prevalent estimate per establishment.

The important feature of the sample design is its use of stratified random sampling with a Neyman allocation. Because industry (TEI) and employment (size class) groups is highly correlated with an establishment's number and rate of reported injuries and illnesses (TRC), stratified sampling provides greater precision and results in a smaller sample size than simple random sampling would require. With Neyman allocation, the overall variance of the TRC rate is minimized for a fixed total sample size.

4.1 Allocation Process

The optimum allocation method used is an iterative process. Certainty cells are removed from the calculation after each iteration. Any cell allocated more sample than there are units in the cell are designated as certainty cells. Certainty cells can also occur as a result of ensuring that an adequate number of units are sampled to produce accurate and reliable estimates for the cell. The methodology also ensures that each sampling cell has at least two units selected (where there are at least two units in the cell) and that the maximum weight of any sampled unit is less than 250. The end product is a probability sample where, for a fixed overall sample size, the sample is distributed across the strata such that the sampling variance of the TRC rate is minimized.

The allocation module is a multi-step process applied to each State's sampling frame created as described above. The highlights of the major steps in the process are as follows:

- The Measure of Size (MOS) for sampling cell h is derived using the following formula:

$$\mathrm{MOS_h} = \sqrt{\mathrm{p_h}(1-\mathrm{p_h})} * \mathrm{E_h}$$

Where E_h = employment in stratum h and p_h = TRC rate / 100

- Certainty and minimum sample criteria are applied. The number of units in certainty cells and the minimum number of units allocated to non-certainty cells are excluded from the allocation process.

- Sum the MOS_h of the non-certainty cells. The first iteration of optimum allocation is performed by computing the sample size for each remaining non-certainty cell:

$$n_h = \frac{MOS_h}{\sum_{h=1}^{k} MOS_h} * n'$$

Where MOS_h = measure of size for sampling cell h not assigned a sample size
k = number of sampling cells not assigned a sample size
n' = total sample size for the State minus sample size already fixed
n_h = sample size in sampling cell h rounded to the nearest integer

- Check for new certainties, that is, where the stratum sample size exceeds the number of frame units and recalculate the sum of the *MOS* and *n'*.

- Re-allocate *n'* to the remaining non-certainty sampling cells and check for new certainty cells. If certainty cells exist, then complete another iteration of allocation. If no new certainty cells exist, then n_h is the sample size for the stratum. Finally, apply minimum sample sizes where there are at least two frame units in a cell. Also check to see if frame units divided by the allocated sample size is less than 250. If not, divide the frame units by 250 and round up to the nearest integer to determine the minimum sample size. These additions are made to the final sample. No further adjustments to other sampling cells are made. Thus, the final sample may be slightly larger due to these adjustments.

4.2 Sample Selection

The sample selection is done using the SAS survey select procedure. The procedure uses two inputs files: (1) the final State frame file and (2) the final allocation file. A systematic selection with equal probability is used to select a sample from each sampling cell (stratum). As mentioned earlier, a sampling cell is defined as State/ownership/TEI/size class. Prior to sample selection, units within a sampling cell are sorted by employment and then by LDB number (unique identifier assigned to each reporting unit on the QCEW) to ensure a consistent representation of all employments in each stratum. The output from the sample selection includes a sample weight assigned to each sample member. Sampling weights are calculated by dividing the number of frame units in the sampling cell by the number of sample units in that cell.

After sample selection is complete and prior to releasing the sample to the State, case subsampling codes are added to help reduce respondent burden for establishments where a large number of cases requiring days away from work are expected. As mentioned earlier, more detailed data are gathered for these cases to support the Case and Demographics estimates. During estimation, case subsampling factors are calculated to reduce the number of cases submitted from the expected number of cases recorded by the establishment.

4.3 Final Summary Weight Calculation

Nonfatal workplace injury and illness data collected for the SOII are used to tabulate estimates for two separate data series: (1) summary, or industry-level, estimates and (2) more detailed case and demographic estimates for cases that involve days away from work (DAWF). By means of a weighting procedure, sample units represent all units from the universe or sampling frame. A final summary weight (FSW) is calculated for usable units to represent all units in their size class. This final summary weight is a product of the original sample weight and four adjustment factors: (1) unit nonresponse or Nonresponse Adjustment Factor (NRAF); (2) reaggregation factor (REAG); (3) Outlier Adjustment Factor (OAF); and (4) benchmarking factor (BMF).

$$FSW = \text{Original sample weight x NRAF x REAG x OAF x BMF}$$

The unit nonresponse (NRAF) factor adjusts for the small proportion (less than 9%) of establishments that do not respond to the survey. This factor is calculated by dividing the sum of the weighted employment of viable units in a sampling cell by the sum of the weighted employment of usable units in that sampling cell. Viable units are units deemed to be in-scope during collection (i.e., collected units and units from which no response was obtained).

$$NRAF = \frac{\sum \text{weighted viable employment}}{\sum \text{weighted usable employment}}$$

The reaggregation factor (REAG) is applied to adjust for those instances where a sample unit may be unable to report data for the unit that was sampled. For example, company XYZ reports to the sampling frame as several single locations. One single location is sampled and data are requested for only that location but the company maintains one log for all locations and cannot separate the data for the sampled unit. The reaggregation factor is simply a ratio of the assigned employment to the reported employment.

$$\text{REAG} = \frac{\text{assigned employment}}{\text{reported employment}}$$

The outlier adjustment factor (OAF) is applied to units where an unusually extreme case count or hours worked has an undue influence on the estimates. For example, a unit in the health care industry reports an unusually high number of illness cases. The documentation noted a severe scabies outbreak that led to the high number of reported cases. Possible outliers are identified on a report generated by the estimation system after preliminary estimates are run. This report is first reviewed to verify that the reported data are correct (i.e. no data entry errors). State offices review this report and identify units they feel should be considered an outlier. The National Office uses several review tools to determine if the outlier request will be granted. An outlier adjustment factor is calculated to make the unit self representing, effectively changing its final weight to one before applying the benchmark factor. The formula for the OAF is:

$$\text{OAF} = \frac{1}{\text{original sample weight x NRAF x REAG}}$$

The remaining units in a sampling cell that contains an outlier unit need to have an OAF calculated that accounts for the outlier now being self representing. This factor essentially equally distributes the remaining weighted employment of the outlier unit to the other usable units in the sampling cell.

The final factor to be calculated to produce the final summary weight is the benchmark factor (BMF). Benchmarking adjusts the reported summary data for an industry to account for employment changes in the universe between the time the sample was selected and the reference period of the collected injury and illness data. The sample for a given current survey is selected from QCEW data that is approximately two years old. During this two year lag, establishments may close, start up, or change employment size. The reported employment is provided by the responding establishment. The TEI target employment is obtained from the most recent employment data from the QCEW. The SOII uses this employment as a benchmark to adjust injury and illness estimates. Benchmarking is performed at the lowest estimating level, the individual TEI level.

$$\text{BMF} = \frac{\text{TEI target emploment}}{\sum \text{TEI weighted reported employment}}$$

Industry benchmark factor ratios are produced and reviewed by the States. The BMF ratio for aggregate industry levels higher than the TEI level is simply the ratio of the sum of the weighted employments of the next lower levels to the target employment of the aggregate TEI level. Industry out-of-range BMFs have been defined and are usually caused by a change in size class of an establishment. That is, there is a change in the assigned employment versus the reported employment that causes a unit to change size classes. Estimates with BMFs out of range are not published unless a waiver is requested by the State and approved by the national office.

For the more detailed case and demographic estimates for DAWF cases, the final summary weight applied to each case is adjusted by additional factors to ensure that the number of usable cases that have been submitted represent the total DAWF cases reported by the establishment used in the tabulation of summary estimates. More information on the case and demographic estimates can be found in Chapter 9 of the BLS Handbook of Methods.

5. Estimation

The final summary weight is used to weight the cases of individual establishments to produce counts of injuries and illnesses by various characteristics as ownership, industry and size class. For example, to produce an estimate for the manufacturing industry in the private sector, weighted cases are summed for establishments with private ownership within NAICS codes 310000 to 339999.

In addition to injury and illness counts, the SOII also reports the frequency (incidence rate) of such cases in terms of the number of injuries and illnesses per 100 full-time workers. Incidence rates permit comparison among industries and establishments of varying sizes. They express various measures of injuries and illnesses in terms of a constant reflecting exposure hours in the work environment. The formula used for calculating incidence rates across injuries and illnesses and for injury cases only is:

$$\text{Incidence Rate} = \frac{(\text{Sum of characteristic reported}) * 200{,}000}{\text{Sum of number of hours worked}}$$

where the 200,000 represents 40 hours per week x 50 weeks x 100 full-time employees within the calendar year.

In this way, a firm with five cases recorded for 70 employees can compare its injury and illness experience to that of an entire industry with 150,000 employees and 12,000 cases. To view all the estimates produced by the SOII, please visit http://www.bls.gov/iif/.

Incidence rates for illnesses and for case and worker characteristic categories are published per 10,000 fulltime employees. The above incidence rate calculation would then use 20,000,000 hours instead of the 200,000 hours to represent the 10,000 full-time employees working 40 hours per week, 50 weeks per year.

6. Reliability of Estimates

Since a probability sample was used to produce estimates for occupational injuries and illnesses, these estimates probably differ from estimates that would be obtained from a census. All estimates derived from a sample are subject to sampling and non-sampling error. Sampling errors occur because observations (estimates) are made on a sample, not the entire population (census). Standard errors are calculated to determine precision or error for each estimate in the survey. The survey does not adequately report some long-term latent illnesses. The inability to obtain information about all cases in the sample, mistakes in recording or coding the data, and definition difficulties are other examples of non-sampling error in the survey. Beginning with the 2006 SOII, a quality assurance program was implemented to address these potential non-sampling errors. This program is evaluated annually and modifications are made when deemed necessary by program management.

The SOII uses a Taylor series linearization methodology to calculate estimates of standard errors for published estimates. This method is flexible with the survey design and is relatively easy to program. Due to the number of estimates produced for individual State and national estimates, other variance estimators are time consuming and require ample computer storage.

Standard errors are used to determine if estimates meet publishable criteria defined by the OSHS program office. Standard errors are also used in validation of statistical comparisons made within a publication. Relative standard errors, standard errors divided by the estimate, are calculated for each estimate and are available on the BLS website mentioned above.

References

BLS Handbook of Methods, Chapter 9, September 2008,
http://www.bls.gov/opub/hom/pdf/homch9.pdf

Cochran, Willam G. (1953), *Sampling Techniques*, New York: John Wiley & Sons, Inc.

Wolter, Kirk M. (1985), *Introduction to Variance Estimation*, New York: Springer-Verlag, Inc.

The Occupational Safety and Health Act of 1970,
www.osha.gov

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