Sample Allocation to Increase the Expected Number of Publishable Cells in the Survey of Occupational Injuries and Illnesses

October 2, 2015

Diem-Tran Kratzke[*] Daniell Toth[*]

Abstract

The Survey of Occupational Injuries and Illnesses (SOII) is an establishment survey that provides annual estimates for the incidence count and rate of employer-reported work-related injuries and illnesses. Results of the survey are published by industry for the nation and participating states. Low response rates for some industries within a state result in many of the state industry-level estimates not being published because of quality and/or confidentiality concerns. The SOII sample is stratified by state, ownership, industry, and size. The number of sample units from each sampling stratum is currently determined by the Neyman allocation, which is intended to minimize the expected sample variance of the estimator for total recordable cases given the fixed sample size. Our goal for the study is to develop a new sample allocation to increase the publishability of estimates at the state industry level while constraining the variance for the fixed sample size. In this paper, we explore a method for assigning sample allocation that aims to maximize the number of publishable cells while constraining the variance of the estimator for total recordable cases.

Key Words: Constrained optimization, establishment survey, optimum allocation, generalized variance function (GVF), stratified sample, gradient descent.

1. Survey Background

The Survey of Occupational Injuries and Illnesses (SOII) is an annual Federal/State program that collects reports of employee injury and illness along with total employment and total hours worked by all workers from about 240,000 business establishments. The SOII summary program produces estimates of incidence count and rate of nonfatal injuries and illnesses in the workplace by geographical area, ownership, and industry for workers in establishments that are in the scope of the survey in the fifty states, District of Columbia, Puerto Rico, Virgin Islands, and Guam. The survey excludes self-employed workers, workers in agricultural production businesses with less than eleven employees, private households, US postal service, and federal government. Workers in railroad and mining are in-scope but are not sampled; their data are obtained from other sources. National estimates exclude Guam, Puerto Rico, and the Virgin Islands.

The SOII uses a stratified simple random sample design. The sampling units are business establishments. We will use the terms “unit” and “establishment” interchangeably in this paper. Within each survey year, the in-scope establishments are stratified into state, ownership (whether private, state government, or local government), and industry. SOII data are published at this stratified level. In addition, each estimation stratum is divided into sampling strata by grouping establishments into five size classes (a size class is defined by the establishment’s annual average employment which is the average of employment over a twelve-month period). A fixed sample size is assigned to each state and ownership code. Within each sampling stratum, the non-certainty units are sorted by the annual average employment and units are then selected systematically with a single random start. The current SOII uses the Neyman allocation method that minimizes the variance of the estimator for the number or incidence count of Total Recordable Case (TRC) because there is a high correlation between TRC and other characteristics being measured.

[*] U.S. Bureau of Labor Statistics, 2 Massachusetts Ave, N.E. Washington DC 20212

2. Motivation

At the 2012 Occupational Safety and Health Statistics National Conference, questions of publishability arose. Some states wanted to know the reason why only a few of their Target Estimation Industries (TEIs) were published. Target Estimation Industries are specific industries that participating states request to sample for publication. The Statistical Methods Group in the Office of Compensation and Working Conditions of the Bureau of Labor Statistics conducted a review and found that some industries have unusually low response rates compared to others. In 2013, some states requested to increase the sample sizes for certain TEIs to account for low response rates. This request raised the question of how we could optimize the sample allocation by maximizing publishability while keeping the variance at or below an acceptable level.

The SOII sample is stratified by ownership and industry for participating states into H strata for estimation purposes. For sampling purposes, each estimation stratum h is stratified into five size classes h_j, for $j = 1, \ldots, 5$. Estimates at the state/ownership/industry level are provided for as many of the H strata that are deemed publishable by program office economists.

The goal of this research is to find a sample allocation method that will maximize the number of strata that are predicted to be published based on our model, while keeping the variance below an acceptable level.

The decision on whether a given stratum-level estimate is publishable takes into account the perceived quality of the estimate as well as privacy concerns of responders. However, exact criteria for making this decision can be complex and are often not available. Therefore, in order to define an allocation procedure that increases the number of TEIs that get published, we have to rely on a model of the propensity that a TEI, given an allocation, will be published. We use historical SOII data to model the probability that a TEI gets published.

3. Assumptions and Definitions

In order to obtain an allocation to maximize the number of publishable estimates, we need a model for the probability of publishing estimates for an estimation stratum h given its sample size, denoted by $\mathbf{n}_h = \big(n_{h_1}, \ldots, n_{h_5}\big)$ where n_{h_i} is the sample size of size class j in stratum h. We denote the probability of publishing stratum h by $p_h(\mathbf{n}_h)$.

Based on results of exploratory analyses on the historical SOII data, we determined that the probability that a given TEI h is published is closely associated with 1.) the relative variance of its TRC estimate, called RV_h, and 2.) the ratio of the number of usable units to the number of units in the population, called u_h. Usable units are units that provide data that are used in estimation. The relative variance RV_h and the usable ratio u_h are both random variables that depend on the sample allocated to TEI h. In other words, the model of the probability that TEI h gets published is a function of the form

$$p_h\big(RV_h, u_h\big) = p_h\big(RV_h(\mathbf{n}_h), u_h(\mathbf{n}_h)\big) = p_h(\mathbf{n}_h),$$

To be useful in practice, the model must be based on variables with known values at the time of allocation, before data are collected. Since the values of RV_h and u_h will not be known until after the sample is collected, we need to estimate them from historical data.

In estimating RV_h and u_h, we make a number of simplifying assumptions.

First, we assume that the relative variance for a state/ownership/industry/size stratum h_j is given by

$$RV_{h_j} = \sigma^2_{h_j}/n_{h_j}.$$

where σ_{h_j} is a fixed constant over time, so it can be estimated using previous survey results. Since the relative variance is defined by

$$RV_{h_j} = RSE^2_{h_j} = V_{h_j}/TRC^2_{h_j}$$

where

- V_{h_j} is the variance of the TRC estimate for stratum h, size class j
- RSE_{h_j} is the relative standard error of the TRC estimate for stratum h, size class j
- TRC_{h_j} is the Total Recordable Case for stratum h, size class j
- n_{h_j} is the sample size for stratum h, size class j,

we can then write:

$$V_{h_j} = \sigma^2_{h_j} \cdot 1/n_{h_j} \cdot TRC^2_{h_j}$$

The variance for the state/ownership/industry stratum h is the sum of the variances for each size class over all five size classes,

$$V_h = \sum_{j=1}^{5} V_{h_j} = \sum_{j=1}^{5} \sigma^2_{h_j} \cdot 1/n_{h_j} \cdot TRC^2_{h_j}. \tag{1}$$

We can compute RV_h, the relative variance for stratum h, by

$$RV_h = V_h/TRC^2_h \tag{2}$$

where $TRC_h = \sum_{j=1}^{5} TRC_{h_j}$.

Secondly, we assume that the probability that a unit responds to the survey given that it was selected in the sample is the same for all units in a given size class j in TEI h and that this response rate is fixed over time. The response rate for each stratum, r_{h_j}, can then be estimated from previous survey results.

The sample response rate for size class j in TEI h, r_{h_j}, is defined as the ratio of the number of usable units to the number of sample units in that sampling stratum. Usable units are sample units that responded and their data were used in estimation. Given the fixed response rate r_{h_j}, the ratio of the number of usable units to the number of frame units at the TEI level h is computed by

$$u_h = \frac{\sum_{j=1}^{5} r_{h_j} \cdot n_{h_j}}{\sum_{j=1}^{5} N_{h_j}} \tag{3}$$

We use the logistic regression model

$$p_h(\mathbf{n}_h) = \left(1 + \exp\{-\mathbf{X}_h(\mathbf{n}_h)\beta\}\right)^{-1}, \tag{4}$$

where β is the column vector of parameters $(\beta_1, \beta_2, \beta_3)'$, estimated from the historical SOII data and

$$\mathbf{X}(\mathbf{n}_h) = \left(RV_h(\mathbf{n}_h), u_h(\mathbf{n}_h), RV_h(\mathbf{n}_h)u_h(\mathbf{n}_h)\right)$$

to model the probability of TEI h being published.

In order to obtain a sample allocation that will result in collecting data that allows more of the H stratum-level estimates to be published, we allocate sample units in a way that maximizes

$$\sum_{h=1}^{H} \hat{p}_h(\mathbf{n}_h), \tag{5}$$

where $\hat{p}_h(\mathbf{n}_h)$ is the estimated probability that TEI h will be published given a sample allocation of $\mathbf{n}_h$, assuming the logistic model (4).

The expected number of published stratum-level estimates is $\sum_h p_h(\mathbf{n}_h)$, where $p_h(\mathbf{n}_h)$ is the true probability that the estimate for TEI h will be published, given that the sample size for each size class j is n_{h_j}. If $\hat{p}_h(n_{h_1}, \ldots, n_{h_5})$ does a good job of estimating $p_h(n_{h_1}, \ldots, n_{h_5})$, then the resulting sample allocation will come close to maximizing the expected number of published stratum-level estimates.

The maximization is done under the constraints that the total sample size n is fixed,

$$n = \sum_{h=1}^{H} \sum_{j=1}^{5} n_{h_j},$$

and the relative standard error is at or below an acceptable level,

$$\sigma_{h_j} n_{h_j}^{-1/2} \leq \sigma_M,$$

where σ_M is a fixed constant set before the allocation procedure.

4. Modeling Probability of Publishability

Since the criteria for deciding whether to publish a given stratum-level estimate is left to judgment and not exact, we model the probability $p_h(\mathbf{n}_h)$ that an estimate for a given TEI h is published given its sample size $n_h = \sum_{j=1}^{5} n_{h_j}$, using logistic regression.

Our research data include past information for all strata and size classes for participating states in four years 2009-2012. We use the GLM function of the R package to run logistic regression to determine the variables that are important in predicting the probability of publishing. We evaluate different models by using the first three years of data for modeling and the last year of data for testing our model fit. We identify two stratum-level variables that seem to drive the decision on publishability: the relative variance of the TRC estimate (RV_h) and the usable ratio (u_h), which is the ratio of number of usable units to number of frame units.

Recall that formulas (1) and (3) in section 3 show that the variance and usable ratio variables are functions of quantities that we must have estimates for at the time of allocation, except for n_{h_j} and N_{h_j}. We estimate TRC_{h_j} and r_{h_j} by taking the averages of their estimates over the four years. We estimate $\sigma^2_{h_j}$ in various ways, including by obtaining the coefficients of modeling the regression of $1/n_{h_j}$ on RV_{h_j} at different levels (ownership/industry group/size, industry group/size, and size levels). In the end, we find using the average of the estimates of $\sigma^2_{h_j} = RSE^2_{h_j} \cdot n_{h_j}$ over the four years of data works best.

We fit the following logistic regression of publishability on relative variance and usable ratio:

$$\ln \frac{p_h}{1 - p_h} = \alpha + \beta_1 RV_h + \beta_2 u_h + \beta_3 RV_h \cdot u_h$$

Relative variance and usable ratio, which are not available at the time of allocation, are estimated using historical data by:

$$\hat{RV_h} = \frac{\sum_{j=1}^{5} \hat{\sigma}^2_{h_j} \cdot 1/n_{h_j} \cdot T\hat{R}C^2_{h_j}}{(\sum_{j=1}^{5} T\hat{R}C_{h_j})^2}$$

and

$$\hat{u}_h = \frac{\sum_{j=1}^{5} \hat{r}_{h_j} \cdot n_{h_j}}{\sum_{j=1}^{5} N_{h_j}}$$

where

- p_h is the probability of being published for stratum h
- $\hat{RV}_h$ is the estimate for RV_h
- $\hat{\sigma}^2_{h_j}$ is the estimate for $\sigma^2_{h_j}$ (average of $n_{h_j} \cdot R\hat{S}E^2_{h_j}$ over four years)
- n_{h_j} is the sample size in stratum h size class j
- $T\hat{R}C_{h_j}$ is the estimate for TRC_{h_j} (average of $T\hat{R}C_{h_j}$ over four years)
- $\hat{u}_h$ is the estimate for u_h
- $\hat{r}_{h_j}$ is the estimate for r_{h_j} (average of $\hat{r}_{h_j}$ over four years)
- N_{h_j} is the population size in stratum h size class j

5. Optimum Allocation

To find the allocation $\mathbf{n} = (n_{1_1}, \ldots, n_{1_5}, \ldots, n_{H_1}, \ldots n_{H_5})$ that maximizes equation (5), we explore the partial derivative of this sum with respect to n_{h_j}. This is equal to

$$\frac{\partial p_h}{\partial n_{h_j}} = (1 - p_h) p_h \nabla_j(\mathbf{X}_h)\beta, \tag{6}$$

where p_h is the function (4) evaluated at $\mathbf{n}_h$ and

$$\nabla_j(\mathbf{X}_h) = \Big(\frac{\partial RV_h}{\partial n_{h_j}}, \frac{\partial u_h}{\partial n_{h_j}}, \frac{\partial RV_h u_h}{\partial n_{h_j}}\Big).$$

We find that under the specified constraints that $R\hat{V}_h \leq M$ and n is fixed, optimizing equation (5) is not possible by numeric optimizers or by solving directly with the Lagrange method. We implement a numerical routine to find an allocation that satisfies our constraints and increases the expected number of published cells. The routine we adopt reduces n_{h_j} for sampling cells with relatively small values of the partial derivative $\frac{\partial p_h}{\partial n_{h_j}}$ and increases n_{h_j} for sampling cells with relatively large partial derivatives. This is done using the following algorithm:

1. For a given allocation $\mathbf{n}$, compute $\frac{\partial p_h}{\partial n_{h_j}}$ for all n_{h_j}.
2. Find the vector $\mathbf{U}$ of n_{h_j} that have values of $\frac{\partial p_h}{\partial n_{h_j}}$ in the top α%, and $n_{h_j} < \min(N_{h_j}, 800)$
3. Find the vector $\mathbf{B}$ of n_{h_j} that have values of $\frac{\partial p_h}{\partial n_{h_j}}$ in the bottom α% and $n_{h_j} > \min(N_{h_j}, 2)$.
4. Randomly choose a value in $\mathbf{B}$ from which to subtract one and a value in $\mathbf{U}$ to which to add one.
5. If the new allocation $\mathbf{n}'$ satisfies the conditions, repeat the above steps 1-4 with $\mathbf{n}'$. Otherwise, repeat the above steps 1-4 with $\mathbf{n}$.

This algorithm is repeated until the value of equation (5), given $\mathbf{n}'$, begins to stabilize.

6. Empirical Results

6.1 Logistic Regression

To determine the best logistic regression model for our purposes, we use the first three years of data (2009-2011) to model, pooling data from all states and territories. We determine that the relative variance of TRC estimate, the ratio of usable units to frame units, and their interaction are significant factors in predicting the probabilities of being published. The coefficients for our logistic regression model fitting the log odds of an industry being published on these three variables are as follows:

$$\ln \frac{p_h}{1 - p_h} = 2.83 - 12.29 RV_h - 2.46 u_h + 13.53(RV_h \cdot u_h)$$

We apply the above fitted model to the 2012 data of all states and territories to measure how well the model predicts publishability. We compute the predicted probability of each TEI being published by using the estimated coefficients of our fitted model and compare the sum of probabilities of being published over all TEIs to the actual number of TEIs being published in the 2012 sample data. The evaluative statistics are the mean squared error (MSE) computed at the macro and micro levels as follows:

$$Macro\ MSE = \sqrt{\frac{(\sum_h \hat{p}_h - \sum_h publ_h)^2}{(\sum_h publ_h)^2}} \tag{7}$$

and

$$Micro\ MSE = \frac{\sum_h (\hat{p}_h - publ_h)^2}{h}, \tag{8}$$

where $publ_h$ equals 1 if stratum h was published, 0 otherwise. Our model yields a macro MSE of about 1.3% and a micro MSE of about 14%.

6.2 Optimization

We test our optimization scheme on the private ownership of five states that have expressed a desire to increase their publishability rates in the past. We use the Neyman allocation sample sizes as the initial values to our optimization program. We compute the partial derivatives of the probability of a TEI being published with respect to the sample size at each size class level on 2012 data for each state. At each iteration step of the optimization program, we change the sample sizes according to the routine described in section 5. We run the iterations until the increase in the sum of predicted probabilities is negligible. In all states except for one, the change in the sum of probabilities being published between 1,000 and 3,000 iterations is less than .5%. For the state in exception, we run into a convergence problem after 500 iterations. However, the change in the sum of probabilities being published between 400 and 500 iterations for this state is also about .5%.

Figure 1 shows the distribution of the predicted probabilities $\hat{p}_h$ over the non-published TEIs (the left-hand column labeled 0 on the horizontal axis) and over the published TEIs (the right-hand column labeled 1 on the horizontal axis) for each state when we fit the model on 2012 data. The distribution of $\hat{p}_h$ over the published TEIs concentrates more on the upper portion of the probability scale from 0 to 1 (on the vertical axis). This indicates that our model gives higher probabilities of being published to the TEIs that were actually published, which is reasonable.

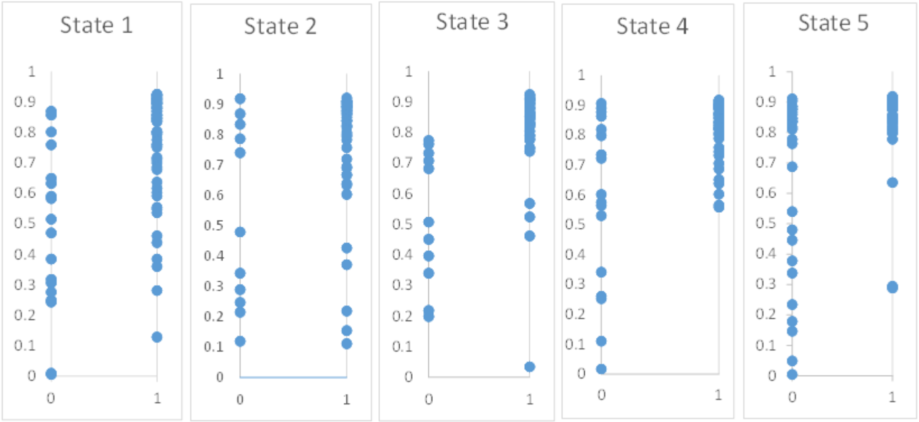
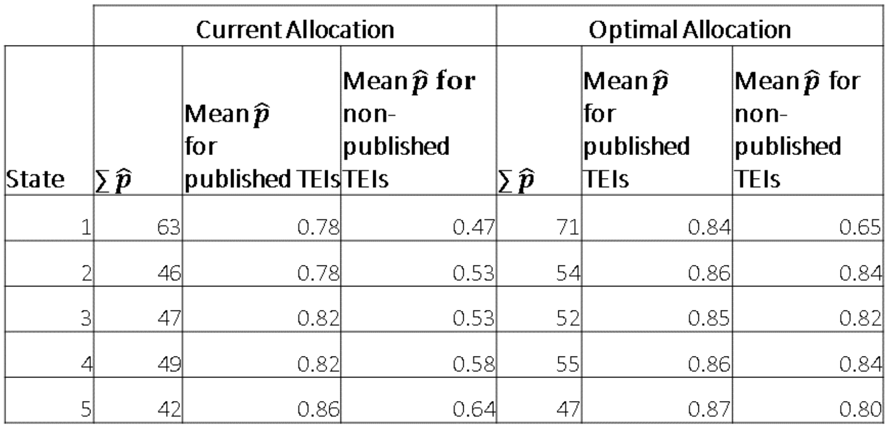


Figure 1: Distribution of $\hat{p}_h$ for unpublished and published TEIs

Figure 2 compares the sum and the mean of the predicted probabilities of being published when we apply the fitted model to the 2012 data over all TEIs in the current allocation method (which is the Neyman method to minimize variance) and the proposed optimal allocation method (which is our method to maximize publishability). The means are shown for published TEIs vs. non-published TEIs.

We see that $\sum_h \hat{p}_h$ is higher for the proposed optimal method in all five states, which confirms that our method generally increases the chances of TEIs being published. We also note that under the current allocation, the mean of $\hat{p}_h$ in published TEIs is higher than the mean of $\hat{p}_h$ in non-published TEIs, with the difference ranging from 22% to 31%, which indicates that our model works reasonably well in predicting probabilities of being published. When we compare $\hat{p}_h$ under the two methods, we see that we are successful in increasing the mean of $\hat{p}_h$ for both published and non-published TEIs. The mean of $\hat{p}_h$ for published TEIs increases from about .80 to about .85 for published TEIs, and from about .55 to about .80 for non-published TEIs.

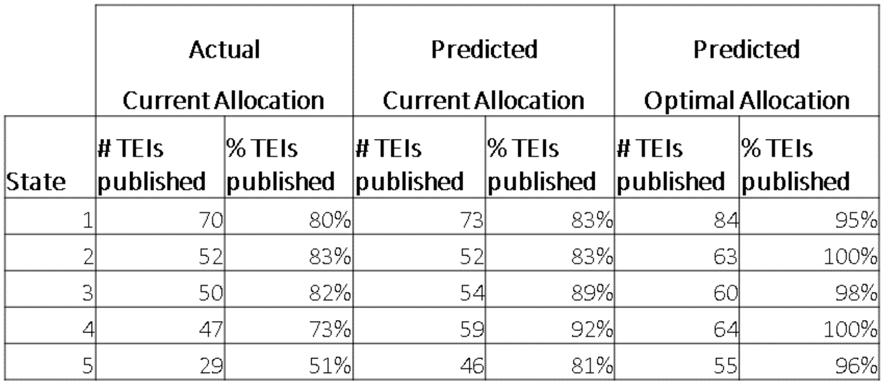


	Current Allocation			Optimal Allocation		
State	$\sum \hat{p}$	Mean $\hat{p}$ for published TEIs	Mean $\hat{p}$ for non-published TEIs	$\sum \hat{p}$	Mean $\hat{p}$ for published TEIs	Mean $\hat{p}$ for non-published TEIs
1	63	0.78	0.47	71	0.84	0.65
2	46	0.78	0.53	54	0.86	0.84
3	47	0.82	0.53	52	0.85	0.82
4	49	0.82	0.58	55	0.86	0.84
5	42	0.86	0.64	47	0.87	0.80

Figure 2: Comparing $\hat{p}_h$ of Current and Optimal Allocation

Next, we use the predicted probabilities of being published obtained from the model to make predictions on publishability with .5 as the cut-off point. That is, we consider any TEI having a predicted probability of being published greater than .5 as publishable. Thus we can compare the number of TEIs deemed publishable by the model to the number of TEIs that were actually published. The results are shown in Figure 3 under the columns "Actual Current Allocation" and "Predicted Current Allocation."

The predicted numbers of publishable TEIs are very close to the actual numbers of TEIs being published in the sample year 2012 for the first three states. In fact, it is exact for state #2. For the last two states, the predicted numbers of publishable TEIs are much higher than the actual numbers of TEIs being published, indicating that the cut-off point of .5 tends to overestimate the number of publishable TEIs.



	Actual Current Allocation		Predicted Current Allocation		Predicted Optimal Allocation	
State	# TEIs published	% TEIs published	# TEIs published	% TEIs published	# TEIs published	% TEIs published
1	70	80%	73	83%	84	95%
2	52	83%	52	83%	63	100%
3	50	82%	54	89%	60	98%
4	47	73%	59	92%	64	100%
5	29	51%	46	81%	55	96%

Figure 3: Actual vs Predicted publishability and Current vs. Optimal predictions

We can also compare the predicted numbers of publishable TEIs of the current allocation to those of the proposed allocation. These results are shown in Figure 3 under the columns "Predicted Current Allocation" and "Predicted Optimal Allocation." Here we see that the predicted number of publishable TEIs in the proposed allocation is higher than the predicted number of publishable TEIs in the current allocation for all five states. The increase in the number of publishable TEIs ranges from 5 to 11 and the percentage increase ranges from 8% to 17%.

7. Conclusions

In conclusion, the preliminary results for publishability in the five states are very promising. The predictions for publishability based on our logistic regression model seem to work reasonably well. Although we tend to overestimate the number of TEIs being published, the prediction and the actual number are positively correlated. That is, higher predicted probability tends to yield higher percentage of cells being published. We are able to increase the sum of predicted probabilities for a number of states while keeping the variance below an acceptable level. We expect that this increase will translate to increasing the number of publishable TEIs if we were to implement our optimal allocation method in production.

We need to do further research to make sure our optimal allocation method works well in practice. That is, we need to obtain results for all states and ownership types to make sure the program provides expected results under different scenarios and that the positive results are not limited to only a hanful of states. We also need to test our optimal allocation in production to address unseen problems that may arise in real situations.

Disclaimer

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

REFERENCES

Bureau of Labor Statistics (2013), *BLS Handbook of Methods*, Chapter 9: Occupational Safety and Health Statistics, `http://www.bls.gov/opub/hom/homch9.htm`

Bureau of Labor Statistics, Quarterly Census of Employment and Wages, `http://www.bls.gov/cew/cewover.htm`

Huband, E., and Bobbitt, P. (2013), “ Nonresponse bias in the survey of occupational injuries and illnesses,” in *Proceedings of the Section on Government Statistics, Joint Statistical Meetings.*

Selby, P., Burdette, T., and Huband, E. (2008), “Overview of the Survey of Occupational Injuries and Illnesses Sample Design and Estimation Methodology,” in *Proceedings of the Section on Survey Research Methods, Joint Statistical Meetings.*