An Estimation Procedure for the New Public Employment Survey Design

Yang Cheng, Casey Corcoran, Joe Barth, Carma Hogue
Governments Division
U.S. Census Bureau[1], Washington, D.C. 20233-0001

Abstract
The 2009 Public Employment Survey uses a new multi-stage sample method which combines cut-off sampling based on unit size with stratified sampling to reduce the sample size, save resources, and improve the precision of the estimates. In this paper, we propose fitting either two separate linear models within size-based strata or one overall, based on the results of a hypothesis test of equality of the model coefficients. We will study the properties and variance of this estimation method. Data from the 2007 Census of Government Employment will be used to compare our method to previous regression method.

Key Words: survey design, model-based estimator, regression, stratification, and optimum allocation

1. Introduction

The Annual Public Employment Survey provides current estimates for full-time and part-time state and local government employment and payroll by government function (i.e., elementary and secondary education, higher education, police protection, fire protection, financial administration, judicial and legal, etc.). This survey covers all state and local governments in the United States, which include counties, cities, townships, special districts, and school districts. The first three types of governments are referred to as general-purpose governments as they generally cover several governmental functions. School districts cover only the education function. Special districts cover generally one, but sometimes two functions. These are the only sources of public employment data by program function and selected job category. Data on employment include number of full-time and part-time employees, gross pay, and hours paid for part-time employees. Reported data are for each government's pay period that includes March 12. Data collection began in March and continued for about seven months.

The methodology, questionnaires, full set of governmental functions, and classification manual are available on www.census.gov/govs/index.html.

In 2007, the Committee on National Statistics, National Research Council, released the findings of a two-year study of the U.S. Census Bureau's surveys of state and local governments. In response to the recommendations given in the study and also to concerns about small units expressed by Census Bureau survey analysts, we decided to look into ways to modify the sampling method.

Currently, a stratified, modified probability proportional-to-size sample method is used to obtain annual national and state level estimates. The current sample design yields a large number of small townships and special districts. These units only account for a very small part of the final

[1] *This report is released to inform interested parties of research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.*

estimates, and have a poor response rate. Within a geographic area, there is usually very little variability in the responses from small units for the same type of government. The objective of this research was to design a sample that would reduce the number of small units in certain problematic areas of the country.

After exploring possible cut-off sample methods for the Annual Public Employment Survey, we suggested an alternative sample method based on the current stratified sample design to reduce the sample size, save resources, and improve the precision of the estimates. We introduced a modified cut-off sample method, which was achieved in two stages. We first stratified the sample by state and government type. Later, we applied a cumulative square root frequency method to determine the cut-off point with respect to the size of unit in the problematic special districts and sub-counties (cities and townships) with two constraints: 1) sample size in the stratum is more than 50; and 2) sample size below the cut-off point is more than 20. The cut-off point serves as a decision point for distinguishing small and large governmental units in the stratum. In the second stage, we applied a sub-sampling method with a fixed rate for these special districts and sub-counties satisfying the two conditions given above. For more information on the methodology, see Barth et al. (2009).

2. Standard Approaches to Estimation

Before we investigate a new estimation procedure that corresponds to the new modified cut-off sample method, let's first introduce some notation and definitions. Let *U* be a finite population of all local governments. There are 89,526 units in our universe, $U = \bigcup_{g=1}^{G} \bigcup_{h=1}^{H} U_{gh}$, where g is the state index, and h represents the government type (h=1 for county governments, h=2 for sub-county governments – includes municipal governments and township governments, h=3 for special districts, and h=4 for school districts).

Let *X* denote the variables of interest from the 2002 Census of Government Employment, such as full-time employment, full-time payroll, part-time employment, and part-time payroll. Let *Y* denote the corresponding variables we measured in 2007. Also, we define a new variable, Total Pay, for each government unit by combining the full-time and part-time payrolls. We used Total Pay as the size of each government unit when we applied the proportional-to-size sampling method.

Before we introduce our proposal for estimation, we demonstrate how the standard estimation approaches work for the modified cut-off sample method.

2.1 Design-Based Approach

The first standard approach is a design-based method. We apply the Horvitz-Thompson (H-T) estimator (Cochran, 1977, p. 259). For given state g and government type h, a sample of n_{gh} units is selected, without replacement. Let π_{ghi} be the first-order inclusion probability for the ith unit in the sample, and $\pi_{ghi,j}$ be the second-order inclusion probability when the ith and jth units are both in the sample. The Horvitz-Thompson (1952) estimator of the population total, $Y = \sum_{g=1}^{G} \sum_{h=1}^{H} \sum_{i=1}^{N_{gh}} y_{ghi}$, is

$$\hat{Y}_{HT} = \sum_{g=1}^{G}\sum_{h=1}^{H}\sum_{i=1}^{n_{gh}} \frac{y_{ghi}}{\pi_{ghi}}$$

where y_{ghi} is the measurement of the variable of interest for the *i*th sampled unit in state *g* and government type *h*.

The H-T estimator, $\hat{Y}_{HT}$, is an unbiased estimator of *Y*, with theoretical variance

$$V(\hat{Y}_{HT}) = \sum_{g=1}^{G}\sum_{h=1}^{H}\sum_{i=1}^{N_{gh}} \frac{(1-\pi_{ghi})}{\pi_{ghi}} y_{ghi}^2 + 2\sum_{g=1}^{G}\sum_{h=1}^{H}\sum_{i=1}^{N_{gh}}\sum_{j \rangle i}^{N_{gh}} \frac{(\pi_{ghi,j} - \pi_{ghi}\pi_{ghj})}{\pi_{ghi}\pi_{ghj}} y_{ghi} y_{ghj}$$

An unbiased sample estimator of $V(\hat{Y}_{HT})$ is

$$\hat{V}(\hat{Y}_{HT}) = \sum_{g=1}^{G}\sum_{h=1}^{H}\sum_{i=1}^{n_{gh}} \frac{(1-\pi_{ghi})}{\pi_{ghi}^2} y_{ghi}^2 + 2\sum_{g=1}^{G}\sum_{h=1}^{H}\sum_{i=1}^{n_{gh}}\sum_{j \rangle i}^{n_{gh}} \frac{(\pi_{ghi,j} - \pi_{ghi}\pi_{ghj})}{\pi_{ghi}\pi_{ghj}\pi_{ghi,j}} y_{ghi} y_{ghj}$$

For this paper, we tested the estimator and variance estimator only for selected states: Alabama, California, Pennsylvania, and Wisconsin. These states were selected because they represent small states, large states, township states with fully functioning township governments (Pennsylvania) and townships that have limited government functions (Wisconsin). Also, we reduced the sample size in sub-county government units in Alabama, reduced the sample size in special districts in California, and reduced the sample size in both sub-counties and special districts in Pennsylvania and Wisconsin. Thus, we end with having five strata in Alabama (counties, small sub-counties, large sub-counties, special districts, and school districts), five strata in California (counties, sub-counties, small special districts, large special districts, and school districts), and six strata in Pennsylvania and Wisconsin (counties, small sub-counties, large sub-counties, small special districts, large special districts, and school districts). **Table 1** displays actual numbers from the 2007 Census of Government Employment, compared with the H-T estimator, difference between the true value and H-T estimator, relative difference, and Coefficient of Variation (CV) of the H-T estimator.

Table 1: Comparison of the H-T Estimates based on PES to 2007 Census of Government Employment

Source: U.S. Census Bureau, 2007 Census of Government Employment. *Payroll is in $1,000s.*

	2007 Census of Government Employment	H-T Estimate	Difference	Relative Difference (%)	CV (%)
Alabama					
Full-time Employment	183,506	188,539	5,033	2.74	1.97
Full-time Payroll	552,926	563,886	10,960	1.98	1.90
Part-time Employment	28,281	25,538	-2,743	-9.70	4.63
Part-time Payroll	24,747	24,198	-549	-2.22	2.84
California					
Full-time Employment	1,228,513	1,214,564	-13,949	-1.14	0.68

Full-time Payroll	6,626,856	6,583,673	-43,183	-0.65	0.48
Part-time Employment	509,494	499,394	-10,100	-1.98	1.97
Part-time Payroll	715,268	699,111	-16,157	-2.26	1.42
Pennsylvania					
Full-time Employment	384,145	388,250	4,105	1.07	1.96
Full-time Payroll	1,493,150	1,525,592	32,443	2.17	1.66
Part-time Employment	111,050	128,262	17,212	15.50	8.69
Part-time Payroll	98,620	108,772	10,152	10.29	8.53
Wisconsin					
Full-time Employment	181,370	177,861	-3,509	-1.93	1.76
Full-time Payroll	702,900	706,111	3,212	0.46	1.06
Part-time Employment	91,103	103,087	11,984	13.15	15.04
Part-time Payroll	82,044	82,554	510	0.62	6.01

In examining these data, we conclude that the weights among the sample units are properly distributed. For a given 95 percent significant level, all true values are falling into the confidence interval of the H-T estimator. The CVs for full-time employment and full-time payroll by state are very small. CVs for part-time employment and part-time payroll by state are reasonable and stable, except for part-time employment in Wisconsin. These CVs show that the variability of full-time variables is much smaller than that of part-time variables.

2.2 Model-Based Approach

The second standard approach is a model-dependent method. When using a probability sampling design, some still prefer inferences that are model-dependent. The linear regression estimation is a method used to increase precision by the use of an auxiliary variable, x_i, the information from 2002 Census of Government Employment, which is correlated with the same information, y_i, in 2007. In model-dependent inference, no matter how the sampling plan and estimator are obtained, inference is made on the basis of the model. Those using model-dependent methods have asserted that once the sample is drawn, the probabilities of selection are irrelevant. They regard the assumption of a model and the use of a best estimator under the model as essential. Model-dependent design and inference may have substantial advantages if the model is appropriate. Based on prior experience or from scatter plots on the variables of interest from the individual sampled local government units, the relationship observed between 2002 and 2007 could be represented approximately by a straight line, and the variability around the regression line increases as *x* increases. When the relation between x_i and y_i is examined, it may be found that although the relation is approximately linear, it does not need to go through the origin. Thus, we propose a simple linear regression model:

$$y_{ghi} = a_{gh} + b_{gh} x_{ghi} + \varepsilon_{ghi}\ ,\ \forall g = 1,\ldots,G; h = 1,\ldots,H; i = 1,\ldots,n_{gh}$$

where y_{ghi} and x_{ghi} are obtained for every government sample unit in state *g* and type of government *h*. We can estimate a_{gh} and b_{gh} using only the sample data. Therefore, the linear regression estimate of the total population, *Y*, is

$$\hat{y}_r = \sum_{g=1}^{G} \sum_{h=1}^{H} \sum_{i=1}^{N_{gh}} \left(\hat{a}_{gh} + \hat{b}_{gh} x_{ghi} \right)$$

or

$$\hat{y}_r = \sum_{g=1}^{G} \sum_{h=1}^{H} N_{gh} \left(\overline{y}_{gh} + \hat{b}_{gh} \left(\overline{X}_{gh} - \overline{x}_{gh} \right) \right)$$

where $\hat{b}_{gh}$ is called the linear regression coefficient of y_{ghi} on x_{ghi} in the finite population and is computed from the sample. The least square estimate of b_{gh} is

$$\hat{b}_{gh} = \frac{\sum_{i=1}^{n_{gh}} \left(x_{ghi} - \overline{x}_{gh} \right)\left(y_{ghi} - \overline{y}_{gh} \right)}{\sum_{i=1}^{n_{gh}} \left(x_{ghi} - \overline{x}_{ghi} \right)^2},$$

where $\hat{a}_{gh} = \overline{y}_{gh} - \hat{b}_{gh} \overline{x}_{gh}$, $\overline{x}_{gh} = \frac{1}{n_{gh}} \sum_{i=1}^{n_{gh}} x_{ghi}$ and $\overline{y}_{gh} = \frac{1}{n_{gh}} \sum_{i=1}^{n_{gh}} y_{ghi}$

If the model holds, the variance of the model-dependent estimator is reduced regardless of the procedure used for sample selection. We only need to estimate the parameters for slope and intercept from the sample. In most cases, the regression line goes through the origin when we observe the sample scatter plots. If the straight line goes through the origin, then the best estimator of slope, b, is simply the ratio of the sample means, i.e., $\hat{b}_{gh} = \overline{y}_{gh} / \overline{x}_{gh}$.

Based on the assumptions of the model-based approach, we can obtain the approximate theoretical variance of the regression estimator $\hat{y}_r$ as

$$V(\hat{y}_r) \approx \sum_{g=1}^{G} \sum_{h=1}^{H} \frac{N_{gh} \left(N_{gh} - n_{gh} \right)}{n_{gh}} S_{gh,y}^2 \left(1 - \rho_{gh}^2 \right)$$

where $\rho_{gh} = S_{gh,xy} / \left(S_{gh,x} * S_{gh,y} \right)$ is the population correlation between variables x_{gh} and y_{gh}, $S_{gh,x}^2 = \frac{1}{N_{gh} - 1} \sum_{i=1}^{N_{gh}} \left(x_{gh,i} - \overline{x}_{gh} \right)^2$ and $S_{gh,y}^2 = \frac{1}{N_{gh} - 1} \sum_{i=1}^{N_{gh}} \left(y_{gh,i} - \overline{y}_{gh} \right)^2$ are the population variances, and $S_{gh,xy} = \frac{1}{N_{gh} - 1} \sum_{i=1}^{N_{gh}} \left(x_{ghi} - \overline{x}_{gh} \right)\left(y_{ghi} - \overline{y}_{gh} \right)$ is the population covariance. The unbiased sample estimator of $V\left(\hat{y}_y \right)$ is

$$\hat{V}(\hat{y}_r) \approx \sum_{g=1}^{G} \sum_{h=1}^{H} \frac{N_{gh} \left(N_{gh} - n_{gh} \right)}{n_{gh} \left(n_{gh} - 2 \right)} \sum_{i=1}^{n_{gh}} \left[\left(y_{ghi} - \overline{y}_{gh} \right) - \hat{b}_{gh} \left(x_{ghi} - \overline{x}_{gh} \right) \right]^2 .$$

Again, we apply the linear regression model and calculate the model-dependent estimator and its variance estimator based on the same sample data from Alabama, California, Pennsylvania, and Wisconsin for full-time employment, full-time payroll, part-time employment, and part-time payroll. The model-dependent estimators for full-time employment and full-time payroll are as good as the H-T estimators when compared to the marginal survey total from 2007 Census of Government Employment. Also, the CVs of the survey total estimate from the model-dependent approach are smaller than from the design-based approach. For part-time employment and part-time payroll, model-based estimates give us better estimators and smaller variation. Therefore, we conclude that the estimates from the model-dependent method have better precision than those from the Horvitz-Thompson method.

2.3 Model-Assisted Approach

The model-assisted methodology is commonly used in the Census Bureau, and is a method between the design-based approach and the model-based approach. We assume the model fits the population reasonably well. However, we cannot make the assumption that the population was really generated by the model. Ultimately, the model serves as the vehicle for finding an appropriate regression coefficient to put into the regression estimator formula. The efficiency of the regression estimator, as compared to the design-based estimator, will depend on the goodness of the fit. The basic properties (approximate unbiasedness, validity of the variance formula, etc.) are not dependent on whether the model holds or not. This procedure is called model-assisted, not model-dependent.

For a simple linear model, we can estimate the parameters, a_{gh} and b_{gh} from the whole population. Because we do not know the whole population, we use data from the sample to calculate the statistics that are used to estimate the slopes and intercepts in the universe. Therefore, model-assisted estimates are determined by both model selection and sample design.

Given state *g* and government type *h* when $g = 1,..., G; h = 1,..., H$, we can write the population least squares line relating to x_{ghi} and y_{ghi} as

$$y_{ghi} = a_{gh} + b_{gh} x_{ghi} + \varepsilon_{ghi}\,,\ \forall g = 1,..., G; h = 1,..., H; i = 1,..., N_{gh}$$

The parameters a_{gh} and b_{gh} are defined in terms of population first and second moments as follows:

$$\hat{a}_{gh} = \overline{Y}_{gh} - \hat{b}_{gh}\overline{X}_{gh} \text{ and } \hat{b}_{gh} = S^2_{gh,xy} \big/ S^2_{gh,x}$$

where $S_{gh,xy}$ is the population covariance, $S^2_{gh,x}$ is the population variance, and $\overline{X}_{gh}$ and $\overline{Y}_{gh}$ are the population means. To obtain estimators, we replace population moments in the above formulas with design-weighted sample moments. We have

$$\hat{b}_{gh} = \frac{\sum_{i=1}^{n_{gh}} \left(x_{ghi} - \overline{x}_{gh}\right)\left(y_{ghi} - \overline{y}_{gh}\right) \big/ \pi_{ghi}}{\sum_{i=1}^{n_{gh}} \left(x_{ghi} - \overline{x}_{gh}\right)^2 \big/ \pi_{ghi}}$$

where $\overline{x}_{gh}$ and $\overline{y}_{gh}$ are the Horvitz-Thompson estimators of $\overline{X}_{gh}$ and $\overline{Y}_{gh}$.

The approximate unbiased sample variance estimator is

$$\hat{V}(\hat{y}_r) \approx \sum_{g=1}^{G}\sum_{h=1}^{H}\frac{1}{(n_{gh}-2)}\sum_{i=1}^{n_{gh}}\left[(y_{ghi}-\bar{y}_{gh})-\hat{b}_{gh}(x_{ghi}-\bar{x}_{gh})\right]^2 \Big/ \pi_{ghi}$$

From the same data set in Alabama, California, Pennsylvania, and Wisconsin, we can calculate the survey total estimates using a model-assisted approach, and compare the results to the 2007 Census of Government Employment. We then calculate the variance estimators of survey total and CVs for four states and four variables. All results are listed in **Table 2** below. We conclude the following significant results: 1) all CVs in the model-assisted approach are much smaller than the H-T estimates; 2) the ranges of relative differences between the model-assisted estimates and the totals of 2007 Census of Government Employment is from 0.02 percent to 7.65 percent, which can be compared with the ranges of relative differences between the H-T estimates and the totals of 2007 Census of Government Employment (from 0.62 percent to 15.50 percent), and in most cases, the relative difference improves when applying the model-assisted method; and 3) the model-assisted estimation significantly improves the precision of survey total estimates for part-time employment and part-time payroll. When the data follows the model, we have very similar estimates for model-assisted and model-dependent methods. Estimates are much better using a model-assisted approach than the model-dependent when the models do not fit. Compared with **Table 1**, we find that the model-assisted methods are much better than the H-T estimator for this sample design.

Table 2: Comparison of Model-Assisted Estimates on PES to 2007 Census of Government Employment

Source: U.S. Census Bureau, 2007 Census of Government Employment. *Payroll is in $1,000s.*

	2007 Census of Government Employment	H-T Estimate	Difference	Relative Difference (%)	CV (%)
Alabama					
Full-time Employment	183,506	194,894	11,388	6.21	0.16
Full-time Payroll	552,926	578,307	25,381	4.59	0.16
Part-time Employment	28,281	30,445	2,164	7.65	0.56
Part-time Payroll	24,747	25,909	1,162	4.70	0.39
California					
Full-time Employment	1,228,513	1,220,421	-8,092	-0.66	0.07
Full-time Payroll	6,626,856	6,594,102	-32,754	-0.49	0.09
Part-time Employment	509,494	509,377	-117	-0.02	0.10
Part-time Payroll	715,268	716,422	1,154	0.16	0.14
Pennsylvania					
Full-time Employment	384,145	391,407	7,262	1.89	0.11
Full-time Payroll	1,493,150	1,542,324	49,174	3.29	0.10

Part-time Employment	111,050	117,341	6,291	5.66	0.18
Part-time Payroll	98,620	98,568	-52	-0.05	2.81
Wisconsin					
Full-time Employment	181,370	182,707	1,337	0.74	0.10
Full-time Payroll	702,900	716,115	13,215	1.88	0.10
Part-time Employment	91,103	97,070	6,272	6.91	0.22
Part-time Payroll	82,044	81,328	-716	-0.87	0.30

2.4 Summary

Using the 2002 Census of Government Employment as the new universe sample frame, we applied the modified cut-off sample method to select mock samples from the 2007 Census of Government Employment. We calculate the survey total estimates for four variables of interest: full-time employment, full-time payroll, part-time employment, and part-time payroll in Alabama, California, Pennsylvania and Wisconsin. Later, we compare these estimates using three standard methods (the design-based approach, the model-dependent approach, and the model-assisted approach) with the true values we get from the 2007 Census of Government Employment. For full-time employment and full-time payroll, the estimates from design-based, model-dependent, and model-assisted all look good. Estimates from the design-based and the model-assisted approaches are slightly better. For part-time employment and part-time payroll, we conclude that the estimates from the model-assisted method and the model-dependent approach are better than the design-based approach. Additionally, we find that when the data fit a model very well, the model-dependent approach appears to have better estimates and variance estimators. Because we sampled with probability proportion-to-size and the size of government is the total pay, we find that the design-based estimates work well for full-time payroll and full-time employment cases. But, if the model fit is not perfect and the sample design is problematic, the model-assisted method works better than the design-based and model-dependent methods. In most of the cases, we find that the model-assisted estimators are better than H-T estimators and model-based estimators for part-time employment and part-time payroll.

3. Decision-Based Approach

Now, we introduce a decision-based approach in order to improve the precision of estimates and reduce the mean square error for the survey total estimate. The idea is to test the equality of linear regression lines to determine whether we can combine data in different strata. Let us start with the following ***Lemma*** .

Lemma: When we fit two linear models for two separate data sets, if $a_1 = a_2$ and $b_1 = b_2$, then the variance of the coefficient estimates is smaller for the combined model fit than for two separate stratum models when the combined model is correct.

For some sub-counties and special districts that satisfy the sample size described in Section 1, we apply a cumulative square root frequency method and create two strata within the same type of government: small units group and large units group. Data from small and large government units are drawn from the same government type. Should we estimate the survey total of key variables by combining small and large unit data or should we keep them separately? To answer this question, we test the equality of two linear regression lines in small versus large government units within sub-counties and special districts (where sub-sampling has occurred in the small government units stratum). Secondly, we evaluate the linear regression among all four types of governments within any given state to determine whether we can combine data with different government types.

3.1 Test of Two Regression Lines

We have the following procedure to test two linear regression lines. First, we compare the slopes by testing the null hypothesis that the slopes are identical (the lines are parallel). The test statistic is $t_{gh} = \left(\hat{b}_{gh,1} - \hat{b}_{gh,2}\right) / s_{b_{gh,1}-b_{gh,2}}$, where the standard error of the difference between the regression coefficients is

$$s_{b_{gh,1}-b_{gh,2}} = \sqrt{\frac{\left(s^2_{gh,xy}\right)_p}{\left(\sum_{i\in S_{gh,1}} x^2_{gh,i}\right)_1} + \frac{\left(s^2_{gh,xy}\right)_p}{\left(\sum_{i\in S_{gh,2}} x^2_{gh,i}\right)_2}}$$

and the pooled residual mean square is calculated as

$$\left(s^2_{gh,xy}\right)_p = \frac{\sum_{i\in S_{gh,1}}\left(y_{gh,i} - \hat{y}_{gh,i}\right)^2 + \sum_{i\in S_{gh,2}}\left(y_{gh,i} - \hat{y}_{gh,i}\right)^2}{n_1 + n_2 - 4},$$

where the subscripts 1 and 2 refer to the two regression lines being compared. The critical value of t_{gh} for the test has $(n_1 - 2) + (n_2 - 2)$ degrees of freedom, namely, $v_{gh} = n_{gh,1} + n_{gh,2} - 4$.

If the P value is less than 0.05, we reject the null hypothesis and conclude that the regression lines are significantly different. In this case, there is no reason to compare the intercepts. If the P value for comparing slopes is greater than 0.05, we can't reject the null hypothesis. Therefore, we conclude that the slopes are not significantly different. Now, we calculate a single slope from combining two data sets. Our next question is whether two regression lines are either parallel or identical.

We test whether two regression lines are parallel or identical by checking whether the two regression lines have the same intercept. To do this, we need to calculate the slope and intercept for the two combined data sets. Also, we need to develop an appropriate test statistic as follows:

$$t_{gh} = \frac{\left(\bar{y}_{gh,1} - \bar{y}_{gh,2}\right) - \hat{b}_{gh,c}\left(\bar{x}_{gh,1} - \bar{x}_{gh,2}\right)}{\sqrt{\left(s^2_{gh,xy}\right)_c\left[1/n_{gh,1} + 1/n_{gh,2} + \left(\bar{x}_{gh,1} - \bar{x}_{gh,2}\right)^2 \Big/ \left(\sum_{i\in S_{gh}} x^2_{gh,i}\right)\right]}}$$

where $\hat{b}_{gh,c}$ is a slope for the combined two data sets, $\left(s^2_{gh,xy}\right)_c$ is the mean square of residual for the combined regression, which equals to $\dfrac{\sum_{i\in S_{gh}} y^2_{gh,i} - \left(\sum_{i\in S_{gh}} x_{gh,i}\, y_{gh,i}\right) \Big/ \left(\sum_{i\in S_{gh}} x^2_{gh,i}\right)}{n_{gh} - 3}$. If the P value is low, we reject the null hypotheses, and conclude that the regression lines are not identical. If the P value is high, we can't reject the null hypothesis, and must conclude that there is no compelling evidence that the regression lines are different.

3.2 Test of More Than Two Regression Lines

Similarly, we test slopes first, and then test intercepts. To compare more than two slopes, we can test $H_0 : b_1 = b_2 = ... = b_k$ with $k > 2$ against the alternative hypothesis that the *k* regression lines were not derived from samples estimating populations among which the slopes were all equal. To compare *k* regression lines, we need to calculate the sample variance of *x* and *y*, the sample covariance of *x* and *y*, and the sums of squares of the residuals and the degrees of freedom for each regression line. The pooled residual sum of square (SS_p) is the sum of *k* individual sum of squares of the residual. The common residual sum of squares (SS_c) is $\sum_{i=1}^{k}(y_i - \bar{y})^2 - \sum_{i=1}^{k}(x_i - \bar{x})(y_i - \bar{y}) \Big/ \sum_{i=1}^{k}(x_i - \bar{x})^2$. To test $H_0 : b_1 = b_2 = ... = b_k$ we calculate the F statistic

$$F = \frac{\left(\frac{SS_c - SS_p}{k-1}\right)}{\frac{SS_p}{\sum_{i=1}^{k} n_i - 2k}}$$

with the numerator and denominator degrees of freedom of *k*-1 and $\sum_{i=1}^{k} n_i - 2k$, respectively.

If we reject the null hypothesis H_0, it means *k* regressions do not have similar slopes. Next, we can test *k* groups of (*k*-1) regression lines. If *k*-1=2, then we can apply the method in Section 3.1. If we cannot reject the null hypothesis, we conclude that all population slopes underlying our k samples of data are equal. In this situation, it is reasonable to ask whether all k population regression lines are identical.

To test the null hypothesis of equality of intercepts, we combine the data from all k samples, and compute a residual sum of squares, SS_t. The null hypothesis is tested with the test statistic

$$F = \frac{\left(\frac{SS_t - SS_c}{k-1}\right)}{\frac{SS_c}{\sum_{i=1}^{k} n_i - k - 1}}$$

with k-1 and $\sum_{i=1}^{k} n_i - k - 1$ degrees of freedom.

If the null hypothesis is rejected, we can employ multiple comparisons to determine the location of significant differences among the elevations. If it is not rejected, then all k sample regression lines are an approximation of the same population regression line, and the best estimate of underlying population regression is given by the ***Lemma***.

In the modified cut-off sample design, we plan to test the relationships of up to four different government types within the state. If we cannot reject all hypotheses, we will combine some government types to have a better estimator with a lower variation for the purpose of improving the precision of estimators.

3.3 Decision-Based Method

The decision-based method first combines the data from different strata by the sample design through the hypothesis test of the equality of the model coefficients, and then applies the model-assisted method to estimate the annual survey totals and their related variances. When we apply the decision-based method for the Public Employment Survey (PES), we have the following specific steps: 1) apply a simple linear regression model for each stratum based on our new two-stage sampling method; 2) perform a hypothesis test on small versus large government units, and determine whether we can combine or keep the two strata separate; 3) perform a hypothesis test on different government types for any given state; 4) fit a simple linear model for the new defined data group; and 5) apply the model-assisted method to compute the survey totals and their CVs. In the next section, we will demonstrate some numerical results from applying the decision-based method.

4. Numerical Results

Based on the 2002 Census of Government Employment, Alabama, California, Pennsylvania, and Wisconsin were ranked as the twenty-eighth, fourth, second, and eleventh, respectively, among the states with respect to the number of local governments. **Table 3** summarizes the overall frame by government type for those states.

Table 3: Government Organization for Studied States in 2002
Source: U.S. Census Bureau, 2002 Census of Government Organization

	Type of Government				
State	**Counties (1)**	**Cities and Townships (3)**	**Special Districts (4)**	**School Districts (5)**	**Subtotal**
Alabama	67	458	529	131	1,185
California	57	478	2,765	1,044	4,344
Pennsylvania	66	2,562	1,728	515	4,871
Wisconsin	72	1,851	756	441	3,120
Subtotal	262	5,349	5,778	2,131	13,520

Our modified two-stage cut-off sample design is equivalent to a stratified sampling with four to six strata for each state. In Alabama, we have five strata: county (labeled as 1), small sub-county (labeled as 31), large sub-county (labeled as 32), special district (labeled as 4), and school district (labeled as 5). In California, we also have five strata: county (labeled as 1), sub-county (labeled as 3), small special district (labeled as 41), large special district (labeled as 42), and school district (labeled as 5). For Pennsylvania and Wisconsin, we have six strata each: county (labeled as 1), small sub-county (labeled as 31), large sub-county (labeled as 32), small special district (labeled as 41), large special district (labeled as 42), and school district (labeled as 5).

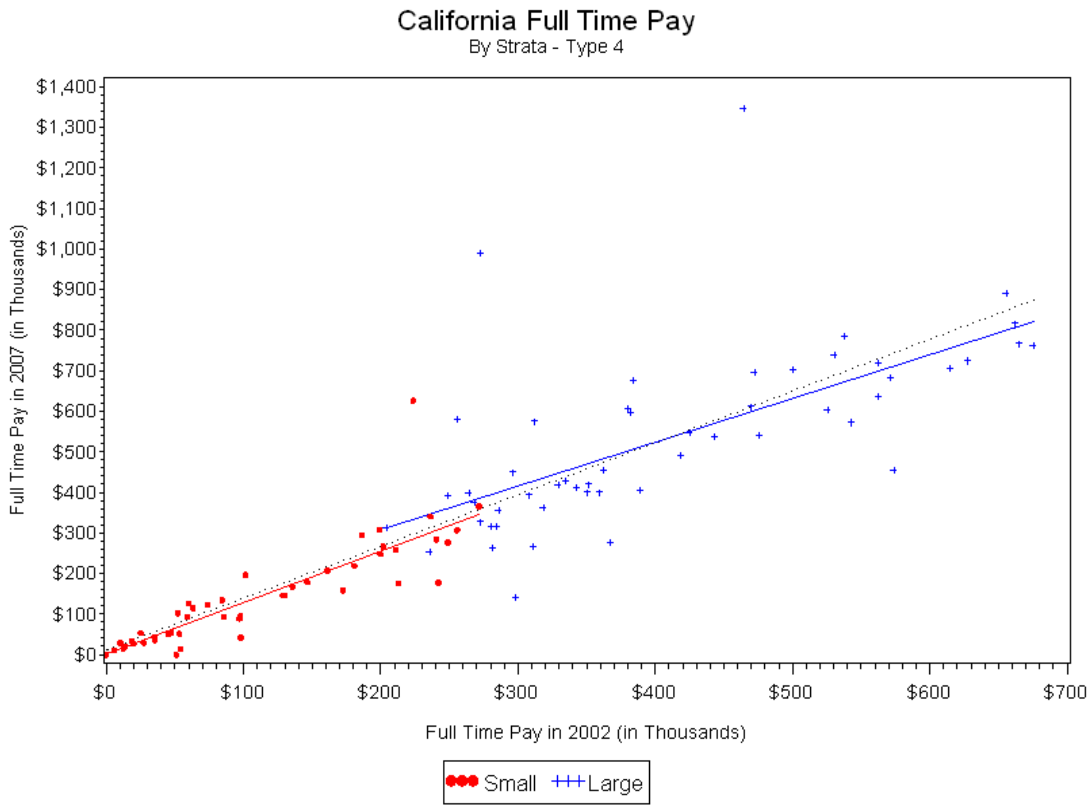
The first step of our decision-based approach estimation procedure is to test small sub-counties (31) versus large sub-counties (32) in Alabama, Pennsylvania, and Wisconsin as well as to test small special districts (41) versus large special districts (42) in California, Pennsylvania, and Wisconsin regarding variables: full-time employment, full-time payroll, part-time employment, and part-time payroll. Applying the method described in Section 3.1, we reject the null hypothesis for full-time employment in California's small special districts as compared with large special districts. For Wisconsin, we reject all null hypotheses except from the part-time employment in special districts. We cannot reject any other categories of small government units and large government units. Thus, we combine these sample data from small government units and large government units together for a better estimate.

Table 4: Combine strata based on the results of hypothesis tests of equality of the model coefficient

	Full-Time Employment	Full-Time Payroll	Part-Time Employment	Part-Time Payroll
Alabama	**(1,3,4), (5)**	(1,3,4,5)	(1,3,4), (5)	(1,3,4), (5)
California	(1,3,5), (40), (41)	(1,3,4, 5)	(1), (3,4,5)	(1), (3,4,5)
Pennsylvania	(1), (3,4,5)	(1,5), (3,4)	(1), (3,4), (5)	(1), (3,4,5)
Wisconsin	(1), (30), (31), (40), (41), (5)	(1,5), (30), (31), (40), (41)	(1,4), (30), (31), (5)	(1,5), (30), (31), (40), (41)

If we cannot combine the small and large government units in the category of sub-county or special district, then we only test three government types: county, special district and school district or county, sub-county and school district. Otherwise, we will test four government types: county, sub-county, special district, and school district. After a series of tests, we conclude the following: 1) we should combine all data for full-time payroll in Alabama and California; 2) for full-time employment, part-time employment, and part-time payroll in Alabama, we should combine data from county, city and township, and special district. Thus, we will fit two separate regression lines; 3) for full-time employment and part-time payroll in Alabama, and for part-time employment and part-time payroll in California, we should combine data from cities and

Figure 1: Linear fits for small and large special districts in California regarding full-time payroll versus linear fit for data combining small and large special district

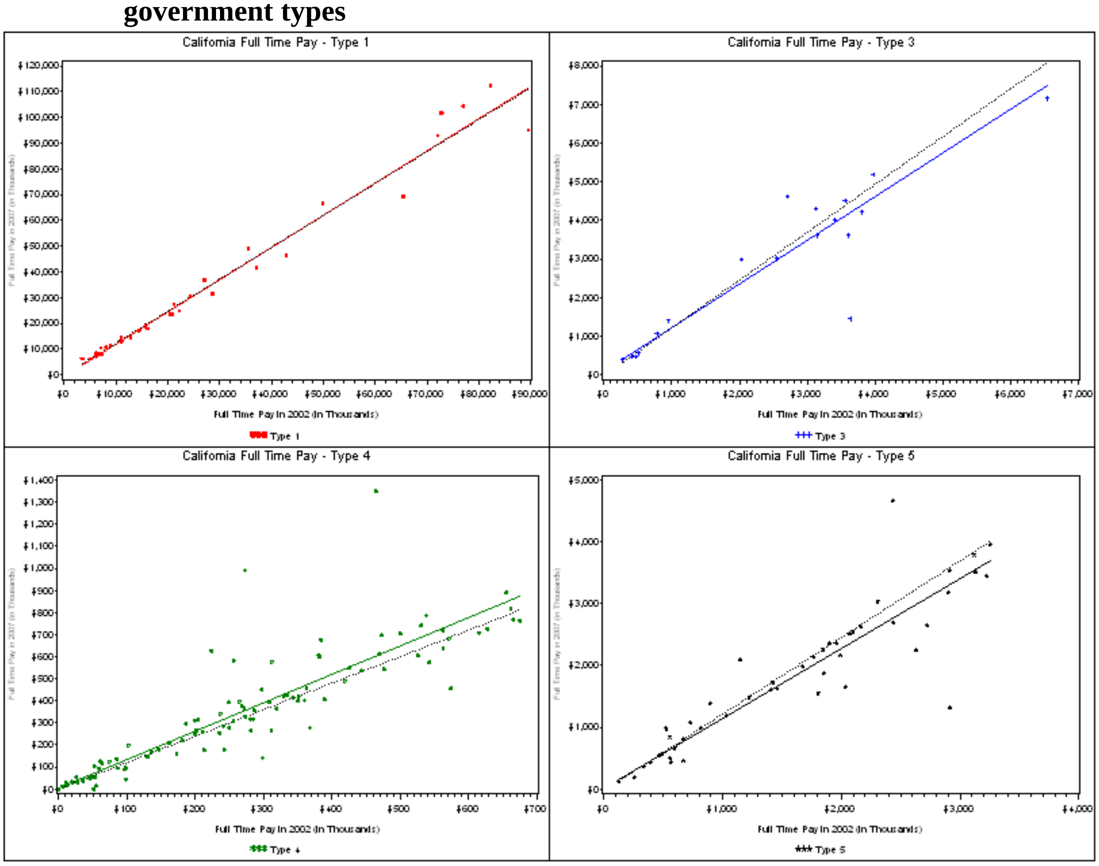


townships, special districts, and school districts. Thus, we will fit two separate regression lines; and 4) for full-time payroll and part-time payroll in Wisconsin, we are only able to combine data from counties and school districts. Therefore, we have five different regression lines.

Table 4 displays all possible combinations from the null hypothesis, which test the equality of the model coefficient. Symbol (1,3,4) means that we can group data from counties, cities and townships, and special districts, and then fit one simple linear model. Symbol (5) means that we can model by data from school districts without the other government types. For Alabama, California, and Pennsylvania, we are able to combine data from different government types. But for Wisconsin, we can't combine data in the different categories except combining data from counties and school districts for full-time payroll and part-time payroll, and combining data from counties and special districts for part-time employment.

Figure 1 displays how the decision-based approach works on small government units as compared with large government units. We use full-time payroll in California as an example. Two solid straight lines are linear regression fits for small and large special districts. They are not the same, but have very similar slopes and a small difference between two intercepts. Since we cannot reject the null hypothesis of testing equality of the model coefficients and claim model coefficients are significantly different, we combine the small and large government units to reduce model error when we apply the model fit for combining data for separate stratum models. The dotted line is the best linear fit for combining data.

Figure 2: Linear fit for individual government type vs. linear fit for data combining all government types



Again, we use full-time payment in California as an example to demonstrate how the hypothesis tests of equality work among the four different government types. In **Figure 2**, the solid line is the best linear fit for each individual government type, and the dotted line is the best linear fit for data combining all government types. We can see from **Figure 2** that the difference between two straight lines for different types of government is very small, and for the county government that are almost identical.

Finally, we compare the Coefficients of Variation (CV) among the H-T estimator, the model-assisted estimator, and our proposed decision-based estimator. We calculated the CV for full-time employment, full-time payroll, part-time employment, and part-time payroll in the states of Alabama, California, Pennsylvania, and Wisconsin.

Table 5 displays the CV comparison for three estimation methods among four variables of interest and four states. All CVs from the H-T method are less than 2 percent for full-time variables. Some CVs are pretty large for part-time variables, especially for Pennsylvania and Wisconsin. All CVs from the model-assisted approach are significantly improved over the H-T method. Only one CV is more than 0.5 percent for part-time employment in Alabama. Most of the CVs are less than 0.2 percent. All CVs in the decision-based approach are less than those from the model-assisted approach. In most cases, it seems that the improvement is small. For example, the CV for the part-time employment estimate in Wisconsin is 0.220 percent when applying the model-assisted method as compared with 0.215 percent when applying the decision-based method. However, we see some significant changes between the model-assisted and decision-based approaches. For example, CVs for the full-time payroll estimates in Alabama and California, or full-time employment estimates in California, have improved by more than 50 percent. They are from 0.159 percent, 0.086 percent, and 0.069 percent down to 0.070 percent, 0.035 percent, and 0.022 percent, respectively.

Table 5: CV comparison among H-T, model-assisted, and decision-based

		H-T	Model Assisted	Decision Based
Alabama	ft_emp	1.971%	0.158%	0.136%
	ft_pay	1.898%	0.159%	0.070%
	pt_emp	4.628%	0.564%	0.531%
	pt_pay	2.839%	0.385%	0.363%
California	ft_emp	0.675%	0.069%	0.022%
	ft_pay	0.484%	0.086%	0.035%
	pt_emp	1.966%	0.102%	0.088%
	pt_pay	1.420%	0.140%	0.129%
Pennsylvania	ft_emp	1.963%	0.111%	0.087%
	ft_pay	1.655%	0.101%	0.082%
	pt_emp	8.690%	0.179%	0.172%
	pt_pay	8.528%	0.281%	0.163%
Wisconsin	ft_emp	1.762%	0.096%	0.096%
	ft_pay	1.065%	0.100%	0.077%
	pt_emp	15.045%	0.220%	0.215%
	pt_pay	6.014%	0.301%	0.233%

5. Future Plans

In the future, we will plan to investigate more complicated models instead of simple linear regression models by exploring more variables such as population size which may affect variables of interest in: full-time employment, full-time payroll, part-time employment, and part-time payroll. We can even explore some nonlinear models or nonlinear estimators.

Secondly, we will also address the accuracy of the variance estimator in the decision-based approach. A simple standard variance formula may not be suited for our complicated survey design. Plus, there are many variations attributable to data collection such as missing data or nonresponse error. We are exploring a variance estimator based on the concept of replication methods such as random groups, balanced half-samples, and jackknife. A bias study should be conducted as well.

Finally, we need to consider a data simulation study to quantify variance due to the decision (group merging) process. The keys for a decision-based estimation are to group data in different categories through testing a series of hypotheses of equality of model coefficients. We plan to study whether the variation exists during these hypothesis tests and how much the variance increases.

Acknowledgements
We acknowledge the contributions of Dr. Eric Slud from the Statistical Research Division of the U.S. Census Bureau and from the University of Maryland at College Park, and Dr. Patrick Cantwell from the Decennial Statistical Studies Division of the U.S. Census Bureau. Also, we are indebted to our reviewers, Dr. Eric Slud, Ms. Rita Petroni, and Ms. Lisa Blumerman for their helpful suggestions, which have improved the original paper.

References
Barth, J., Cheng, Y., Hogue, C. (2009), Reducing the Public Employment Survey Sample Size, JSM 2009

Cochran, W.G. (1977), Sampling Techniques. Third Edition. New York: John Wiley & Sons, Inc.

Hansen, M.H., Madow, W.G., and Tepping, B.J. (1983), An Evaluation of Model-Dependent and Probability-Sampling Inference in Sample Surveys, Jour. Amer. Stat. Assoc., 78, 776-793.

Horvitz, D.G., and Thompson, D.J. (1952), A generalization of sampling without replacement from a finite universe, Jour. Amer. Stat. Assoc., 47, 663-685.

Pardoe, I. (2006), Applied Regression Modeling – A Business Approach. New York: John Wiley & Sons, Inc.

Sarndal, C.-E., Swensson, B., and Wretman, J. (1992), Model Assisted Survey Sampling, Springer-Verlag.

Zar, J.H. (1999), Biostatistical Analysis. Third Edition. New Jersey, Prentice-Hall