# GOVERNMENTS DIVISION REPORT SERIES

(Research Report #2012-2)

## New Technique for Modifying the Cutoff Sample and Its Application

Yang Cheng

U.S. Census Bureau
Washington, DC 20233

CITATION: Cheng, Yang. 2012. New Technique for Modifying the Cutoff Sample and Its Application. Governments Division Report Series, Research Report #2012-2

_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

Report Completed: September 23, 2011
Report Issued: February 22, 2012

*Disclaimer: This report is released to inform interested parties of research and to encourage discussion of work in progress. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.*

# New Technique for Modifying the Cutoff Sample and Its Application

Yang Cheng
Governments Division
U.S. Census Bureau[1], Washington, D.C. 20233-0001

**Abstract:** The cutoff sampling technique is sometimes used in establishment surveys in the U.S. Census Bureau. Recently, the Governments Division in the U.S. Census Bureau proposed a new method of cutoff sampling, which modified the traditional cutoff sampling techniques by constructing the cutoff points based on the size of units in the strata and then reducing the sample size in the strata with small-sized units. We also introduced a decision-based estimation method as a stratum-wise regression for strata defined first by cutoff sampling methods, and then through stratum-collapsing rules determined from the results of a hypothesis test for equality of regression slopes. Finally, we applied the modified cutoff sampling technique and decision-based estimation for two major surveys of governments: the Annual Survey of Public Employment & Payroll and the Annual Finance Survey.

**Key Words:** Cutoff sampling, modified cutoff sampling, regression estimators, decision-based estimation, governments statistics

## 1. Background

### 1.1 Cut-off sampling method

One of the most cost effective solutions for many establishment surveys is "cut-off sampling", which deliberately excludes a subset of the population. This approach involves selecting only the largest units in the population for the sample and using statistical or econometric models to extrapolate the information to the smaller units. The cutoff sampling technique is often used in establishment surveys in the U.S. Census Bureau.

The standard sampling method for establishment surveys, used in most federal statistical agencies, is stratified probability proportional-to-size (PPS) with a certainty stratum. But, it is problematic in some surveys. Cut-off sampling can be applied as an alternative to stratified PPS. Information on large firms often drives the PPS sample-based estimates. In this practical sense, cut-off sampling may not be a major departure from tradition. There is a major departure in a theoretical sense, however, because it is not probability-based sampling.

[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.*

Key reasons for using cut-off sampling method over a probability-based sampling method are to reduce the costs, reduce respondent burden, improve data quality by keeping the same large firms in the sample over time, and avoid using frames that are unsuitable for PPS sample selection due to low quality, incomplete, or outdated frame information. Reducing costs and respondent burden are the primary motivation behind the cut-off sampling idea. Survey sponsors must weigh the potential loss in estimation accuracy and the additional difficulty in calculating error estimates against the savings, in both cost and respondent burden, gained by eliminating the smaller units from the sample. In some cases, the quality of the data provided by a subset of the population is poor due to inadequate understanding of the survey questionnaire, and site visits to remedy the data quality issues are costly. The potential error caused by including this subset in the sample may outweigh the statistical advantage gained through use of a more inclusive sample. Cut-off sampling may, therefore, produce more accurate estimates, in terms of total survey error.

Cut-off sampling may be more cost effective than probability sampling. It is only probability sampling on a restricted frame. Cut-off sampling will produce accurate estimates when we are only interested in the ratio of two variables of interest as opposed to a population total. The technique could also be quite useful when we attempt to find differences or to prove the existence of a particular feature in the inconsistent members of the target population. In general, cut-off sampling is an effective sampling methodology when we have a good working knowledge of the characteristics of the total population and subpopulations of interest and are willing to take some calculated risks with regard to the lack of information from the unsampled subpopulations. Under many quite varied sets of circumstances, it can provide the required information at minimal cost.

There are a few key steps when we implement the cut-off sampling method. First, we need to determine the measure of size. Employment numbers, revenue, or key outputs of surveys are often used as the size variable when we apply the cut-off PPS sample method. Secondly, we stratify by geographic area, industrial classifications, age, and other variables of interest. To optimize the cut-off point is a further challenge. Most surveys will add some new units, for example, births, to the sample.

Sampling error is inherent to all sample surveys and reflects the natural variability in the estimates that occurs because data are only collected from a subset of the population. However, the use of cut-off sampling has the potential of increasing sampling error since the inclusion probability of some units is equal to zero. When any subset of a population has no chance of selection and estimation is based on a model, there is always a chance that substantial changes in that subset not being sampled could take place resulting in model failure. The Census Bureau makes use of administrative, census information, and prior year data to minimize the potential of sampling error. The basic standard approach is to use sample data to estimate change over time, and to apply change factors to frame data from the smaller units. The most common way of estimating data for units below a cut-off is by extrapolating from reported data using ratio estimators. The data that a unit reported in an earlier period is adjusted for the current period using growth rates calculated from similar units above the cutoff. The ratio estimation methodology can be

used to fill in components of aggregate data. Statistical uses of administrative records are more popular and important for federal agencies now. What are some options for measuring the error in estimates based on cutoff samples? In some cases a "gold standard", for example: a census is available for comparison, usually with some lag. In other cases, error bounds may be inherent in the cut-off thresholds, for example: if ninety percent of the total for the variable of interest is covered in the sample, the error potential is limited to the remaining ten percent. Model-based error estimates are also used in some surveys.

### 1.2 Annual survey of public employment and payroll

The Annual Survey of Public Employment and Payroll (ASPEP) collects data and provides statistical information on the number of federal, state, and local government civilian employees and their gross payroll. For more detailed information regarding ASPEP, see the Website: http://www.census.gov/govs/apes/. ASPEP covers all states 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 because they generally provide multiple government activities. These activities were coded into various function codes. School districts cover only the education function. Special districts usually provide one or two function codes. ASPEP is the only source of public employment data by program function and selected job category in the United States. Data on employment include the number of full-time, part-time, and total employees, full-time, part-time, and total gross pay, hours paid for part-time employees, and full-time equivalent employees. Reported data are for the government's pay period that includes March 12 and provides statistics on the number of federal, state, and local government employees and their gross payroll.

The 2007 Census of Governments yielded a total of 89,527 government units, including the federal government. Their state and locate breakdown is as follows:

**Table 1: Breakdown of 2007 Census of Government Employment**

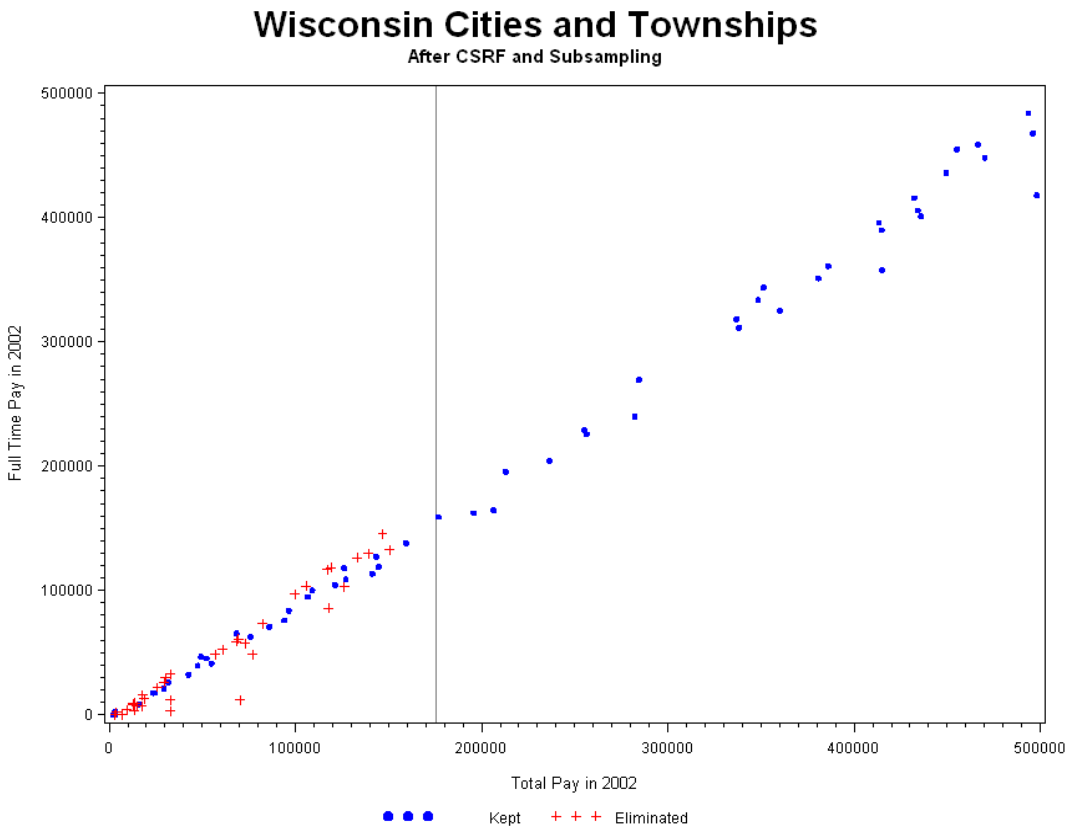
| Government Type | N | Total Employees | Total Payroll | 2008 n | 2009 n |
|---|---|---|---|---|---|
| State | 50 | 5,200,347 | $17,788,744,790 | 50 | 50 |
| County | 3,033 | 2,928,244 | $10,093,125,772 | 1,436 | 1,456 |
| Cities | 19,492 | 3,001,417 | $11,319,797,633 | 2,609 | 3,022 |
| Townships | 16,519 | 509,578 | $1,398,148,831 | 1,534 | 624 |
| Special Districts | 37,381 | 821,369 | $2,651,730,327 | 3,772 | 3,204 |
| School Districts | 13,051 | 6,925,014 | $20,904,942,336 | 2,054 | 2,108 |
| **Total** | **89,526** | **19,385,969** | **$64,156,489,693** | **11,455** | **10,464** |

Source: U.S. Census Bureau, 2007 Census of Governments: Employment.

We learn from **Table 1** that special districts and townships are larger in number and do not contribute much to payroll and employee' totals compared to their counterparts. In

addition, our census and surveys have demonstrated that smaller governments are less likely to be respondents and have less variability than larger governments. To illustrate this latter point, the scatter plot in **Figure 1** shows the concentration of smaller governments and the linear relationship between census years in Wisconsin as well as the decrease in variability of smaller governments.

**Figure 1: Results from Cumulative Square Root of the Frequency Method (using total pay as size variable) on Wisconsin City and Townships**




Source: U.S. Census Bureau, 2002 and 2007 Census of Governments: Employment

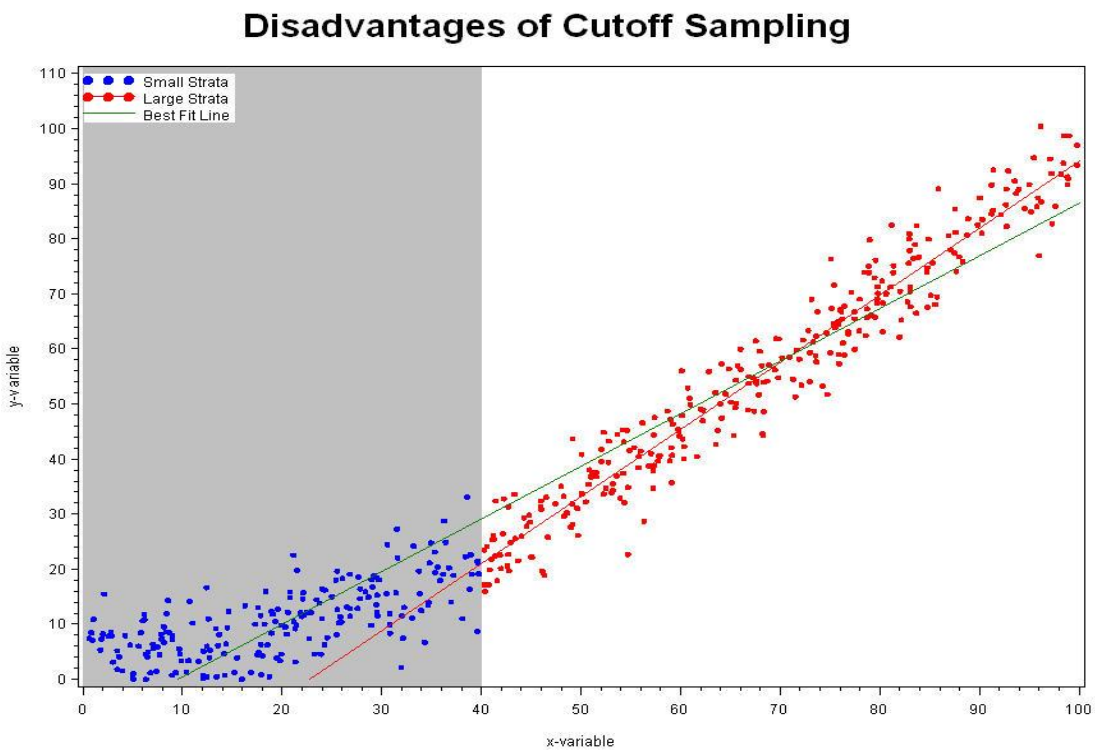
The sample methodology for ASPEP is a traditional stratified probability proportional-to-size (PPS) sample design to obtain annual national and State level estimates. ASPEP is stratified by 50 states and Washington, DC and 5 government types: counties, cities, townships, special districts, and school districts. This design yields a large number of small townships and special districts. These units only account for a very small portion of the final estimates and typically have a poor response rate.

The estimation methodology for ASPEP is a model-based estimation by using the most recent Census government data as auxiliary predictor variable.

### 1.3 Motivation

We considered applying the cut-off sampling method when we redesigned ASPEP. However, we worried about the model specification. This is one of the weaknesses of cut-off sampling when models from large units may not fit the unobserved data. We use a previous employment data to simulate a new data set and demonstrate the disadvantages of cut-off sampling in **Figure 2.** The graph shows that the large units and all units generate two different linear models.

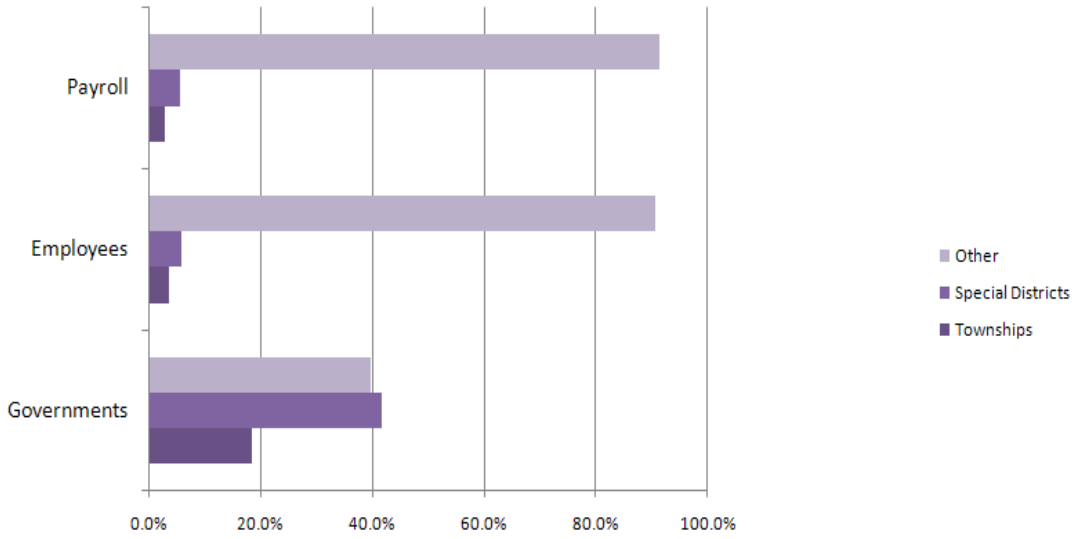
**Figure 2: Result from data simulation shows linear model from large units is different from the model from all date**




Because the smaller government units do not contribute much to the overall total and are more likely to be nonrespondents, one goal of the redesigning of the ASPEP is to reduce the number of smaller and non-contributory units and improve the precision of the estimation. In 2009, a new sample design was implemented to address this issue. It was decided that cities and townships were similar enough that they could be combined into a single stratum, called sub-counties. We developed a new two-stage cut-off sample in order to reduce the overall sample size, thereby reducing respondent burden, and improving data quality. After certainty-sampled units were removed, the first stage of the sample design was a stratified PPS sample, which was a stratified by state and government type. Total payroll from the 2007 Census of Governments: Employment part was used as the size variable for the PPS. The first stage sample yielded excessive numbers of smaller cities/townships and special districts. After the first stage sample, a cut-off point was found for each state by government type stratum for sub-counties and special districts. These cut-off points were chosen based on the cumulative square root of

frequency method and represent the decision point for distinguishing small and large governmental units in the stratum. Instead of conducting a standard cut-off sample where we would ignore the units below the cut-off point, we chose to subsample the smaller cities/townships and special districts using simple random sampling. Although the smaller units do not contribute much to the overall total, this method still allows us to estimate their share.

**Figure 3: A barplot shows the percentages on Payroll, Employees, and Government Units for the Special Districts, Townships, and other governments**




Source: U.S. Census Bureau, 2007 Census of Governments: Employment.

**Figure 3** shows that over half of government units in the United States are townships and special districts. But these townships and special districts only account for about seven percent of total employees and total payroll. This leads us to construct a modified cut-off sampling method only for sub-counties and special districts.

## 2. Modified cut-off sampling

### 2.1 Two-stage sampling

After exploring possible cut-off sample methods for ASPEP, we proposed an alternative sample method, modified cut-off sampling, based on the current traditional stratified PPS 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 applied the stratified PPS sample based on Total payroll by state and government type. Later, we applied a cumulative square root frequency (CSRF) 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). Primary sampling units (PSUs) for ASPEP are local government units such as counties, cities, towns, special districts, and school districts.

Suppose a sample S is a PPS sample with sample size n drawn from universe U of known size N. Suppose further that the subsample $S_m \subset S$ is to be drawn by simple random sampling taking m out of n. Then, $S_m$ is a PPS sample with size m, and the second-order inclusion probabilities for distinct pairs of elements of the subsample are also proportional to the corresponding joint inclusion probabilities for the sample S. Thus, the modified cut-off sample is also a stratified PPS sample. In the ASPEP case, it is stratified by 50 states and Washington, DC and counties, sub-counties (small, large), special districts (small, large), and school districts.

Now, we attempt to find a way to divide a population into small and large strata and then subsample the strata, drawing even fewer units from the small strata into the sample. In addition, we wish to propose new methods that might be better at defining these cut-off points.

### 2.2 Identify the cut-off point

Choosing cutoff points requires balancing data quality with cost and respondent burden. The cost of obtaining a survey response consists of both the staff resources required to collect, edit, and enter the response and the burden on the respondent. When setting the cutoff point, there is a minimum level of coverage of a cell for a statistical agency to have confidence in the estimate of that cell.

The modified cut-off sample design is specified by two parameters: (1) the cut-off point, which determines how to construct size-based sub-strata; and (2) the small-unit sub-sampling rate, which is the proportion of small government units to be sampled randomly. We know the value of cut-off point must be between the minimum of Total payroll in 2007 and the maximum of Total payroll in 2007, and the reduction rate must be between 0 and 1. In this sub-section, we will discuss several methods to identify the cut-off point.

We notice that eighty percent coverage of attributes (80/20 rule) is the most popular way to identify the cut-off point in federal government agencies, that is, we expend eighty percent of their resources collecting in establishment surveys or sometimes via site visits to small firms that represent about twenty percent of the target population. In such cases, cutoff sampling may offer a reasonable alternative to PPS.

The CSRF method was first introduced in the paper by Dalenius and Hodges (1957). In practice, we apply the CSRF method in ASPEP to determine the cut-off points in every state by sub-counties or special districts. If the population distribution is much skewed, the setting where cutoff sampling is typically used. It is proper to apply the modified geometric method by Gunning and Horgan (2004). A problem arose when we applied the CSRF method in the Taxable Property Values (TPV) feasibility study in the Governments Division of the Census Bureau because it has a much skewed population. But, modified geometric method performed perfectly to construct 3 size-based strata. There are a large number of small entities that contribute little to the aggregate data. In 2010, we explored a numerical solution to determine the turning point by means of a genetic algorithm by Barth and Cheng (2010). We use the turning point as an optimum cut-off point.

### 2.3 Penalized mean square errors

How do we select the parameters of modified cut-off sampling? In the ASPEP redesign, we first applied the CSRF method for each state by government type stratum to determine large and small size-based strata. Then, we identified which stratum to further sub-sample based on the satisfaction of two constrains we mentioned in **sub-section 2.2**. Working with analysts and subject experts, we targeted 800 sample reductions in 2009 ASPEP through cost benefit analysis. The total number of units in all small strata which satisfied the two constraints was approximate 2,000. Thus, we calculated the overall reduction rate to be forty percent. In the second stage, we applied simple random sampling to subsample these qualified small size strata in the state by sub-counties or special districts.

**Figure 1** also demonstrates how the modified cut-off sampling works by using data from sun-counties (cities and townships) in the State of Wisconsin. All points in **Figure 1** are sampled local government units. Applying the CSRF method based on the size variable: Total payroll for each government unit, we locate the cut-off point. The vertical line in **Figure 1** is the cut-off line. The large units' stratum is on the right side, and small units' stratum is on the left side. In the second stage, we eliminated the units with "+" after sub-sampling with a fixed reduction rate.

After the ASPEP redesign in 2009, **Table 1** shows the sample size changes by government types from Fiscal Year (FY) 2008 to FY 2009. The results are that we increased the sample size in cities by 413, reduced the sample in townships by 910, and reduced the number of special districts by 568. The total sample reduction was approximately 1,000 after applying the modified cut-off sampling method in FY2009.

In order to optimize the sample parameters, we introduce the mean square error (MSE) as a loss function. Corcoran and Cheng (2010) showed that the algorithm will almost always choose to keep 100 percent of the samples if we try to minimize the mean squared error with respect to the cut-off point and reduction rate. It is difficult to find a reasonable reduction rate through minimizing the MSE when we apply a model-assisted

method to estimate the total attributes. In most cases, the reduction rate is close to 1. This means that it achieves the minimum without sub-sampling. Our intention is to reduce the sample size and save the costs. Therefore, a penalized loss function is introduced by adding a cost function into MSE.

Let S be the overall sample of size n and $S = S_1 \cup S_2$ where $S_1$ is the small size units sub-stratum with size $n_1$ and $S_2$ is the large size units sub-stratum with size $n_2$. Cut-off point, c, is defined as the breakpoint between size-based strata. This cut-off point distinguishes small and large samples. Also, we define p as the sampling rate with which we subsample $S_1$* from $S_1$. The subsample size $n_1$* is equal to $pn_1$. Thus, $n_1^* \leq n_1$. The penalized loss function is as follows:

$$MSE_{penalized} = MSE_1(c, p) + MSE_2(c) + dn_1^k p^k \tag{1}$$

where d is the cost for each unit in the sample, k=½, 1, or 2, and

$$MSE_1(c, p) = \frac{1}{n_1^* - 2} \sum_{\{x_j \leq c,\, j \in S_1^*\}} (y_{1,j} - \hat{y}_{1,j})^2$$

$$MSE_2(c) = \frac{1}{n_2 - 2} \sum_{\{x_j > c,\, j \in S_2\}} (y_{2,j} - \hat{y}_{2,j})^2$$

After including the penalty term, we are able to minimize the total penalized mean squared error with respect to the cut-off point and proportion of small stratum reduction if given a specified cost. The MSE formulas can be found in many standard sampling text books, where $\hat{y}_j$ is the linear predictor from the sample data corresponding to $y_j$. We are minimizing the sum of MSE over small and large sample sets to obtain the optimum of the cut-off point and subsample rate.

## 3. Decision-based estimation

### 3.1 Decision-based approach

The modified cut-off sampling approach constructed the large size sub-stratum and small size sub-stratum, and then sub-sampled in each small size sub-stratum. At the end, it is still a stratified PPS sampling. Thus, we can apply all standard stratified PPS estimation methods. Since small and large size sub-strata belong to one stratum at the beginning, it is natural to think about combining data together after reducing the small samples. The decision-based idea is to test the equality of the model parameters to determine whether we can combine data in different stratum in order to improve the precision of the estimates. Cheng et al. (2009) analyzed data in the resulting stratified design with a linear regression estimator (using the previous Census value as a predictor) within each stratum.

We explore the formula for estimating the survey total of key variables: full-time employment, full-time payroll, part-time employment, part-time payroll, part-time hours, total employment, total payroll, and full-time equivalent employment. Let Z, the total payroll from the most recent census, be the size variable used in PPS sampling. The design-based Horvitz-Thompson estimator of the unknown total, $Y = \ _{i \in U} y_i$ , is

$$\hat{Y} = \sum_g \sum_h \sum_{i \in S_{gh}} y_i / \pi_i \tag{2}$$

where is $\pi_i$ the first-order inclusion probability of unit i in $S_{gh}$, a known function of $z_i$'s. Z is the size variable of the unit.

From the model approach, we start with a simple linear regression model:

$$y_{ghi} = a_{gh} + b_{gh} x_{ghi} + \varepsilon_{ghi} \tag{3}$$

where $g = 1,\ldots,G; h = 1,2; i = 1,\ldots,N_{gh}$

Following the Särndal et al. (2003) model-assisted approach and applying linear regression equation (3) with each $S_{gh}$, we have the separate ratio estimation as follows:

$$\hat{Y}_{REG1} = \sum_g \sum_h \left[ \frac{N_{gh}}{\hat{N}_{gh}} \hat{Y}_{gh} + \hat{\beta}_{gh} \left( X_{gh} - \frac{N_{gh}}{\hat{N}_{gh}} \hat{X}_{gh} \right) \right] \tag{4}$$

where $X_{gh} = \sum_{i \in U_{gh}} x_i$ , $\hat{Y}_{gh} = \sum_{i \in S_{gh}} y_i / \pi_i$ , $\hat{X}_{gh} = \sum_{i \in S_{gh}} x_i / \pi_i$ , $\hat{N}_{gh} = \sum_{i \in S_{gh}} 1/\pi_i$ , and

$$\hat{\beta}_{gh} = \frac{\sum_{i \in S_{gh}} (x_i - \hat{X}_{gh}/\hat{N}_{gh}) y_i / \pi_i}{\sum_{i \in S_{gh}} (x_i - \hat{X}_{gh}/\hat{N}_{gh})^2 / \pi_i}$$

Alternatively, combining two sub-strata $S_{g1}$ and $S_{g2}$, then the combined regression estimator is:

$$\hat{Y}_{REG2} = \sum_g \left[ \frac{N_g}{\hat{N}_g} \hat{Y}_g + \hat{\beta}_g \left( X_g - \frac{N_g}{\hat{N}_g} \hat{X}_g \right) \right] \tag{5}$$

where $\hat{Y}_g = \sum_h \hat{Y}_{gh}$, $\hat{X}_g = \sum_h \hat{X}_{gh}$ , $\hat{N}_g = \sum_h \hat{N}_{gh}$ , and

$$\hat{\beta}_g = \frac{\sum_h \sum_{i \in S_{gh}} (x_i - \hat{X}_g/\hat{N}_g) y_i / \pi_i}{\sum_h \sum_{i \in S_{gh}} (x_i - \hat{X}_g/\hat{N}_g)^2 / \pi_i}$$

### 3.2 Test of null hypothesis

Cheng et al. (2009) proposed a decision-based method to improve the precision of estimates and reduce the mean square error of weighted survey total estimates. The idea was to test the equality of linear regression lines to determine whether we can combine data in different sub-strata. In Cheng et al. (2009), equality of regression lines is tested in two steps. First, a test is performed of the null hypothesis that the slopes are identical. If the P value is less than 0.05, the null hypothesis is rejected in favor of the conclusion 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, the null hypothesis of equality of slopes cannot be rejected, but intercepts can be compared. If the regression lines for the two sub-strata are not found to be significantly different, then a single line is estimated from the combined sub-strata.

Let

$$t_g = \left(\hat{\beta}_{g1} - \hat{\beta}_{g2}\right)\sqrt{v_g} \tag{6}$$

where $v_g$ is an estimated variance of $\hat{\beta}_{g1} - \hat{\beta}_{g2}$.

If $\left|t_g\right| > t_{0.975,n_g-4}$, where $t_{0.975,d}$ is the 0.975 quantile of the t-distribution with d degrees of freedom, then we reject the null hypothesis and conclude that the regression slopes are significantly different.

In general, we define the null hypothesis: $H_0: \beta_S = \beta_L$ for model parameters in small and large sub-strata. Then, the decision-based estimator is

$$\hat{Y}_{DEC} = \begin{cases} \hat{Y}_S + \hat{Y}_L & \text{if we reject } H_0 \\ \hat{Y}_{S\&L} & \text{if we cannot reject } H_0 \end{cases} \tag{7}$$

We can apply either a model-assisted method, generalized weighted regression estimation (GREG), or calibration estimation in the decision-based estimation formula in equation (7). Equivalently, we use the test-based indicator function as follows:

$$\zeta_g = \begin{cases} 0 & \text{If } \left|t_g\right| \le t_{0.975,n_g-4} \\ 1 & \text{Otherwise} \end{cases} \tag{8}$$

In the model-assisted format, the decision-based estimator of Y is

$$\hat{Y}_{DEC} = \sum_g \sum_h \zeta_g \left[ \frac{N_{gh}}{\hat{N}_{gh}} \hat{Y}_{gh} + \hat{\beta}_{gh}\left( X_{gh} - \frac{N_{gh}}{\hat{N}_{gh}} \hat{X}_{gh} \right) \right] + \sum_g \left(1 - \zeta_g\right) \left[ \frac{N_g}{\hat{N}_g} \hat{Y}_g + \hat{\beta}_g \left( X_g - \frac{N_g}{\hat{N}_g} \hat{X}_g \right) \right] \tag{9}$$

### 3.3 Asymptotic property

We developed some asymptotic properties. Shao et al. (2011) shows that under some conditions and linear regression assumptions, we have the following conclusions:

- If $a_1 = a_2$ and $b_1 = b_2$, then $\hat{Y}_{reg1}, \hat{Y}_{reg2}$, and $\hat{Y}_{dec}$ have the same asymptotic normal distribution.
- If $b_1 \neq b_2$, then $\hat{Y}_{reg1}$ and $\hat{Y}_{dec}$ have the same asymptotic normal distribution and $\hat{Y}_{reg2}$ has a different asymptotic normal distribution with a larger variance.

where $a_1$ and $a_2$ are regression intercepts for small and large sub-stratum, and $b_1$ and $b_2$ are regression slopes from linear model (3).

## 4. Variance estimators

In order to compute the variance estimator for survey estimates based on unequal probability sampling when we apply Horvitz-Thompson variance estimators such as the Yates-Grundy estimator. This estimator relies on the joint inclusion probabilities, which can be laborious to specify for PPS selection methods without replacement. For at least one commonly used PPS without-replacement design, Vijayan's (1968) extension of a method of Hanurav (1967), this is now easy in SAS: the Vijayan-Hanurav design is the default PPS method in the new SAS sample selection procedure, PROC SURVEYSELECT, which is now used in selecting PPS samples within state by government type stratum. The formula-based joint inclusion probabilities can be computed directly and stored using options in PROC SURVEYSELECT.

However, a convenient and generally accurate approximate variance formula which avoids the need for joint inclusion probabilities is the PPS *with replacement* (PPSWR) variance estimator

$$\hat{V}_2(\hat{Y}_{DEC}) = \frac{n}{(n-1)} \sum_{i \in S} (z_i - \bar{z})^2 \tag{10}$$

where $z_i = \frac{y_i}{\pi_i}$ and $\bar{z} = \frac{1}{n} \sum_{i \in S} z_i$ . The PPSWR variance estimator usually overestimates the true variance slightly.

We adapt the PPSWR variance formula to cover the model-assisted decision-based estimator (9) based on a PPS sample. The approximate naive theoretical variance obtained by combining variances for stratum-wise regression estimators and using PPSWR variance formula within each stratum is

$$V(\hat{Y}_{DEC}) = \sum_{i=1}^{N} \frac{e_i^2}{p_i} \tag{11}$$

where $p_i$ is the single-draw probability of selecting a sample unit i and $\pi_i = np_i$ for a PPS sample of size n, and $e_i$ is the population regression residual. The variance is estimated by the quantity

$$\hat{V}(\hat{Y}_{DEC}) = \frac{n}{n-1} \sum_{i \in S} \left( \frac{y_i - \hat{y}_i}{\pi_i} \right)^2 \tag{12}$$

Based upon the estimator (9), we produce a naive variance estimator, $\hat{V}_N$ , for the decision-based model-assisted estimator. Using sample data from two sub-strata, the decision-based estimator is either the model-assisted estimator on the stratum formed as the union of two sub-strata when $H_0 : b_1 = b_2$ is accepted, or is the sum of separate sub-stratum model-assisted estimators when this hypothesis is rejected. In the case of accepting the null hypothesis, $\hat{V}_N$ is the whole-stratum variance estimator (10). When

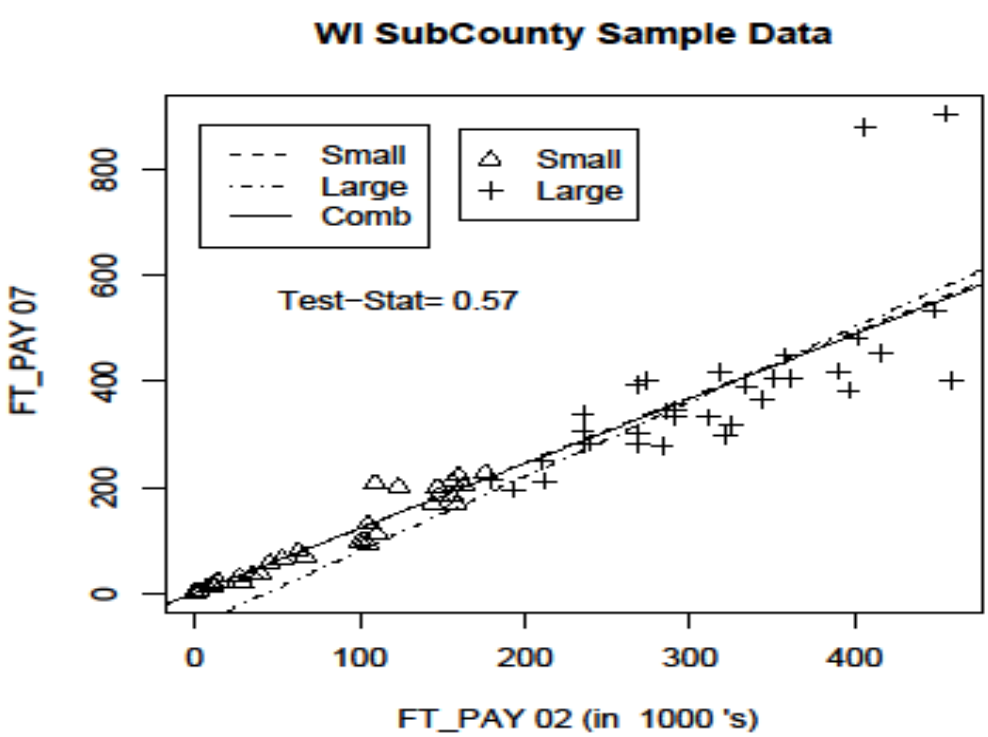
the hypothesis is rejected, $\widehat{V}_N$ is defined as the sum of the separate sub-stratum estimators (10). This naïve variance estimator ignores the randomness inherent in the hypothesis-test-based decision, and treats that decision as though it were known in advance.

## 5. Application

We apply the modified cut-off sampling method and its corresponding decision-based estimation with the real data. We used the 2002 Census of Governments: Employment component as the universe, and drew samples based on the modified cut-off sample design. We obtained data information from the 2007 Census of Governments: Employment component. Since we know all values in both years, we are able to compare the estimation from decision-based model-assisted estimators with the true population in FY 2007.

**Figure 4** shows a real example with employment information from the 2007 Census of Governments. There are 3 linear regressions to fit data from the small, large, and combined stratum for sub-counties in the State of Wisconsin with full-time payroll as the response variable. The test statistic is 0.57, which is less than 1.96. So, the decision-based estimator is the model-assisted combined ratio estimator (5) for the combination of large and small government units together.

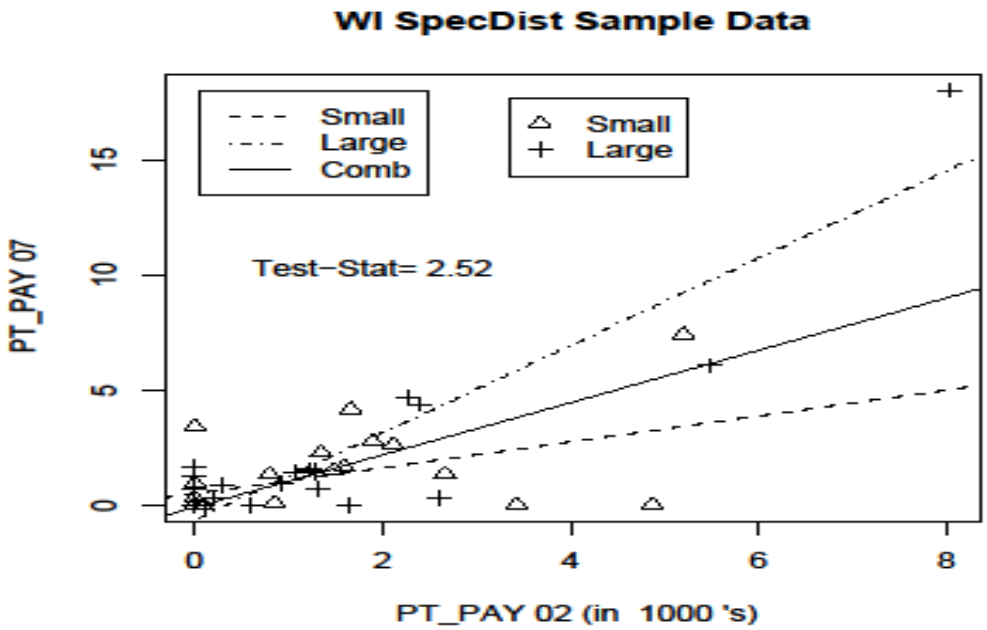
**Figure 4: Decision-based estimate for Wisconsin sub-counties**



U.S. Census Bureau, 2002 and 2007 Census of Governments: Employment

**Figure 5** similarly shows 3 linear regressions, which fit large and small sub-strata as well as the combined-stratum for special districts in the State of Wisconsin with part-time payroll as the y-variable. But in this example, the test statistic is 2.52, which is larger than 1.96. Thus, we reject the null hypothesis and conclude that the decision-based estimator is the separate ratio estimator (4) for the small and large sub-strata.

**Figure 5: Decision-based estimate for Wisconsin special districts data**




U.S. Census Bureau, 2002 and 2007 Census of Governments: Employment

**Table 2: Test results for decision-based model-assisted estimation**

| | FT_Pay | | FT_Emp | | PT_Pay | |
|---|---|---|---|---|---|---|
| (State,Type) | Test-Stat | Decision | Test-Stat | Decision | Test-Stat | Decision |
| (AL,SubCounty) | 2.06 | Reject | 2.04 | Reject | 3.62 | Reject |
| (CA,Specdist) | 0.98 | Accept | 1.02 | Accept | 0.29 | Accept |
| (PA,SubCounty) | 0.54 | Accept | 0.62 | Accept | 0.08 | Accept |
| (PA,SpecDist) | 0.24 | Accept | 0.65 | Accept | 1.09 | Accept |
| (WI,SubCounty) | 0.57 | Accept | 0.85 | Accept | 2.11 | Reject |

| | | | | | | |
|---|---|---|---|---|---|---|
| (WI,SpecDist) | 1.33 | Accept | 0.85 | Accept | 2.52 | Reject |

U.S. Census Bureau, 2002 and 2007 Census of Governments: Employment

**Table 2** shows real data examples of the hypothesis test statistics and decisions in the hypothesis test of $H_0: \beta_S = \beta_L$ . **Table 2** contains 18 tests for full-time payroll, full-time employment, and part-time payroll in the States of Alabama (AL), California (CA), Pennsylvania (PA), and Wisconsin (WI) and sub-counties and special districts.

**Acknowledgements**

I would like to thank Lisa Blumerman and Carma Hogue of the Governments Division at the U.S. Census Bureau for serving as reviewers as well as Dr. Eric Slud from the Center of Statistical Research and Methodology of the U.S. Census Bureau and the University of Maryland. Also, thanks to Dr. Jun Shao from the University of Wisconsin-Madison for useful suggestions and comments.

**References**

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

Barth, J, Cheng, Y. (2010), Stratification of a Sampling Frame with Auxiliary Data into Piecewise Linear Segments by Means of a Genetic Algorithm, JSM Proceedings.

Cheng, Y., Corcoran, C., Barth, J., Hogue, C. (2009). An Estimation Procedure for the New Public Employment Survey, JSM Proceedings

Cheng, Y., Slud, E., Hogue, C. (2010). Variance Estimation for Decision-Based Estimators with Application to the Annual Survey of Public Employment and, JSM Proceedings.

Corcoran, C., Cheng, Y. (2010). Alternative Sample Approach for the Annual Survey of Public Employment and Payroll, JSM Proceedings.

Dalenius, T., Hodges, J. (1957). The Choice of Stratification Points. Skandinavisk Aktuarietidskrift.

Gunning, P., Horgan, J. (2004). A New Algorithm for the Construction of Stratum Boundaries in Skewed Populations, Survey Methodology, 30(2), 159-166.

Knaub, J. R. (2007). Cutoff Sampling and Inference, InterStat.

Sarndal, C., Swensson, B., Wretman, J. (2003). Model Assisted Survey Sampling. Springer.

Shao, J., Slud, E., Cheng, Y., Wang, E.S., Hogue, C. (2011). Model-Assisted Adaptive Inference to the Annual Survey of Public Employment and Payroll, preprint.