Optimal Cutoff Sampling for the Annual Survey of Public Employment and Payroll

Brian Dumbacher[1], Carma Hogue[1]

[1]U.S. Census Bureau
4600 Silver Hill Road, Washington, DC 20233
Brian.Dumbacher@census.gov, Carma.Ray.Hogue@census.gov

Proceedings of the 2013 Federal Committee on Statistical Methodology (FCSM) Research Conference

Abstract

The goal of cutoff sampling is to save cost, reduce respondent burden, and maintain accuracy of estimates by reducing the number of small units in sample. For the Annual Survey of Public Employment and Payroll, the Governments Division of the U.S. Census Bureau uses a modified version of cutoff sampling in which a subsample of units below the cutoff is selected. In this paper, we examine a numerical method based on minimizing the average of mean squared errors from linear regression models to find an optimal combination of cutoff and subsampling rate given a specified cost. Data from the 2002 and 2007 Censuses of Governments: Employment are used for this study.

Key Words: Cutoff sampling; Decision-based estimation; Linear regression; Mean squared error

1. Introduction

1.1 Survey Overview

The Annual Survey of Public Employment and Payroll (ASPEP) is conducted by the Governments Division of the U.S. Census Bureau to collect data on federal, state, and local government civilian employees and their gross payrolls. Key study variables for ASPEP include the total number of employees, total pay, and the number of full-time equivalent employees. Small area composite methodology is used to estimate local government totals for each combination of state and government function. Government functions are identified by item code, and a complete list of item codes is provided in Appendix A.

1.2 Sample Design

The sampling frame for ASPEP is a list of the 89,476 local governments identified during the 2007 Census of Governments and is updated annually with births (newly discovered governments), deaths (disincorporated governments), and mergers. Initial certainties are determined based on population size, school enrollment, and government function, and then a first-stage, stratified, probability-proportional-to-size sample (Särndal, Swensson, & Wretman, 1992, p. 90) is selected from the remaining local governments, where the strata are determined by the cross-classification of state and government type and the measure of size is total pay in 2007.

The different types of local government are counties, cities, townships, special districts, and independent school districts, but cities and townships are deemed similar enough to group together and are known collectively as subcounty governments. Counties and subcounties are known as general-purpose governments, and they tend to perform several of the functions listed in Appendix A. Special districts and school districts, on the other hand, are known as single-purpose governments and tend to perform one or a very limited number of functions. The purpose of school districts is education, while special districts may be cemetery districts, public utilities, transit authorities, etc. As such, their contribution to a single function like air transportation may be great, but they would have no contribution to other functions. Table 1 gives a breakdown of the local governments in 2007 by type. As you can see, there are many subcounties and special districts, but these units' shares of total employees and total pay are disproportionately small.

Table 1: Local governments in 2007

Government type	Number	%	Total employees	%	Total pay ($)	%
County	3,033	3.39	2,928,244	20.64	10,093,125,772	21.77
Subcounty	36,011	40.25	3,510,995	24.75	12,717,946,464	27.43
Special district	37,381	41.78	821,369	5.79	2,651,730,327	5.72
Independent school district	13,051	14.59	6,925,014	48.82	20,904,942,336	45.09
Total	89,476	100.00	14,185,622	100.00	46,367,744,899	100.00

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

To reduce the number of non-contributory units in sample, the Governments Division uses cutoff sampling in the subcounty and special district strata to divide each first-stage sample into a small cutoff stratum and a large cutoff stratum (Barth, Cheng, & Hogue, 2009). For the 2009 sample design, the cutoffs were determined by the cumulative square root of the frequency method (Dalenius & Hodges, 1959). [See Cochran (1977, p. 130) for an overview and Appendix B for an illustration of this method.] Instead of completely ignoring the sample units in the small cutoff strata as is done in standard cutoff sampling, the Governments Division selects a subsample of them. This helps protect against bias that could result from small units becoming large units during intercensal years. A subsampling rate equal to 0.56 was used in all small cutoff strata, which resulted in a target reduction of 800 subcounty and special district units. This target was based on a rough cost-benefit analysis, and we have planned more optimal methods for the future when a new ASPEP sample based on the 2012 Census of Governments is selected.

1.3 Small Area Estimation Methodology

Estimates of local government totals are calculated for each combination of state and item code using small area composite methodology (Tran & Cheng, 2012). Each composite estimate is a weighted average of the direct Horvitz-Thompson estimate and a synthetic estimate. This synthetic estimate equals the product of a decision-based regression estimate of the state total and a proportion for the item code within the state. The term "decision-based" refers to statistical hypothesis tests that are carried out to determine whether the regression relationships in the small and large cutoff strata are similar enough that the strata can be combined for estimation purposes.

Figure 2 is a scatterplot of total pay in a survey year versus total pay in the most recent Census year for a hypothetical sample after the cutoff is determined and subsampling is performed. Separate linear regressions are fitted in the small and large cutoff strata using sample data, and then a statistical hypothesis test of the equality of the regression slopes is carried out. If the null hypothesis is rejected, then the cutoff strata are kept separate. If the null hypothesis is not rejected, then the cutoff strata are combined and a new regression is fitted to all the units. Whichever regression is decided on is then applied to Census data to estimate the state total. This is a simplified description of the process as robust regression and auxiliary data are used to handle outliers and to strengthen poor fitting models. Also, the variable total pay is used in this example, but in production, the decision-based methodology would be applied to full-time pay and part-time pay separately. For a much more detailed description of decision-based estimation, see Cheng, Corcoran, Barth, and Hogue (2009).