Cutoff Sampling in Federal Surveys: An Inter-Agency Review

Daniel Yorgason,[1] Benjamin Bridgman,[1] Yang Cheng,[2] Alan H. Dorfman,[3] Janice Lent,[4] Yan K. Liu,[5] Javier Miranda,[2] Scot Rumburg,[6]

[1]Bureau of Economic Analysis, 1441 L St. NW, Washington, DC 20230
[2]Census Bureau, Washington, DC 20233-0001
[3]Bureau of Labor Statistics, 2 Massachusetts Ave. NE, Washington DC 20212
[4]Energy Information Administration, 1000 Independence Ave. SW, Washington, DC 20585
[5]IRS Statistics of Income Division, 77 K St. NE, Washington, DC 20002
[6]USDA/NASS 1400 Independence Ave., SW, Washington, DC 20250

1. Introduction

Government policy makers, economic analysts, and the general public rely heavily on data gathered from federal establishment surveys for information on the U.S. economy. All data collections conducted by federal statistical agencies are subject to the Office of Management and Budget's (OMB's) Standards and Guidelines for Statistical Surveys. These requirements ensure that, for each survey, the sponsoring agency develops and documents a detailed survey design encompassing all aspects of the survey, e.g., the definition of the target population, the sampling plan, the data collection instrument and methods.

Under the Paperwork Reduction Act of 1995, agencies must submit survey plans to OMB prior to data collection. The plans encompass data collection timetables and estimated costs, including the time cost that the respondents are expected to incur—referred to as "respondent burden"—which the Paperwork Reduction Act was designed to reduce.

Faced with increasing data collection costs and concerns about heavy respondent burden, some agencies are turning to model-based estimation techniques to maximize the information they extract from their survey data, e.g., for small area estimation. One highly cost-effective solution for many establishment surveys is "cutoff sampling." 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.

1.1. Establishment Surveys in the Federal Statistical System

The sampling method traditionally used for surveys of business establishments is stratified *probability proportional to size* (PPS) sampling. In this method, each business establishment in the target population is assigned a *measure of size* (MOS), usually associated with the establishments revenues, output, or number of employees. The population is usually grouped into strata defined by variables that may include size, industry, geographic location, and other characteristics. The largest establishments in the population—or within each subpopulation, defined by non-size variables—are usually placed in a "certainty stratum" and selected for the sample with probability one.

The remaining units form non-certainty strata. In general, these units are selected with probability proportional to their MOS. In order to satisfy reliability requirements for subpopulations, however, sample designers may vary the selection probabilities for subpopulations, even for units

with equal MOS. For example, if industry-specific estimates are desired, units in industries with fewer establishments may receive higher selection probabilities.

In some federal surveys (particularly those targeting specific industries), large establishments cover the bulk of the target population for federal surveys, and these establishments are more likely than smaller ones to be set up for high tech, low cost data collection methods, e.g., online collection, automated electronic data transfer. As a result, data collectors sometimes find the "80/20 rule" at work in establishment surveys: they expend 80% of their resources collecting data—sometimes via site visits—from small firms that represent about 20% of the target population. In such cases, cutoff sampling may offer a reasonable alternative to PPS.

Cutoff sampling may be performed using a variety of size measures gleaned from establishment census data. Information on large firms often drives PPS sample-based estimates, and, in this practical sense, cutoff sampling may not be a major departure from tradition. It is, however, a theoretical departure from probability-based sampling, and statistical agencies are sometimes reluctant to embrace cutoff sampling because of the risk of model failure.

In recent decades, the Bureau of Economic Analysis (BEA) and the Energy Information Administration (EIA) have pioneered the use of cutoff sampling in federal establishment surveys as a means of saving taxpayer dollars and reducing respondent burden. The experiences of these agencies have revealed many benefits and some pitfalls of the technique. It is clear that building robust models for establishment surveys requires a deep understanding of the industries in which the establishments are engaged. Models must account for the establishments' reactions to changing economic conditions (e.g., turning points in the business cycle, rapid price fluctuations), effects of legislative changes, technological advances, industry restructuring, etc.

Firm size—defined in terms of outputs, revenues, and/or employment—is one of the key variables needed for both cutoff sampling and model-based estimation. Modeling changes in firm-size dynamics (i.e., the extent to which larger or smaller firms gain or lose market share over time) is therefore one of the major challenges faced by agencies who desire to save funds and reduce respondent burden through the use of cutoff samples. Without accurate models of firm size changes, the target population coverage provided by cutoff samples is impossible to track over time.

1.2. An Example from the Energy Information Administration

The case of EIA's Monthly Natural Gas Report (form EIA-914) illustrates the importance of industry-economic considerations in the selection of estimation methods for use with cutoff samples. Since 1995, a cutoff sample of about 220 companies producing large volumes of natural gas has reported monthly natural gas production totals to EIA through form EIA-914. The sample is reselected annually, and the natural gas produced by the companies in the sample accounts for 85% to 90% of all the natural gas produced in the lower 48 states during the reference period used for sample selection.

To estimate total natural gas production from the cutoff sample data, EIA uses universe production data obtained from a private vendor. These data, however, are incomplete for the first 6 to 24 months after the reference month. (The lag period varies by state.) For each state that produces large volumes of natural gas, EIA estimates the sample coverage rate—the proportion of total natural gas production that is accounted for by the sample operators—from prior data and projects this rate forward, implicitly assuming that it changes only very slowly over time. The state-level sample production totals are multiplied by the inverses of the estimated coverage rates to compute state-level estimates of total natural gas production.

When petroleum and natural gas prices began to rise rapidly in 2008, the large (in-sample) natural gas well operators increased their production rates faster than did the smaller (non-sample) operators. In addition to the production incentive created by rising prices, technological advances allowed some large companies to increase their shale gas extraction rates. The actual EIA-914 sample coverage rates increased, and the estimated coverage rates, based on prior data, failed to reflect the changes quickly enough. As a result, EIA overestimated natural gas production for some states.

EIA published a disclaimer in early 2009, warning analysts of the possibility of overestimates due to the use of out-dated coverage rates in the EIA-914 estimation process. Analysts within EIA began researching alternative sampling and estimation schemes. In 2009 and 2010, non-government energy industry analysts published articles claiming that the overstated EIA production estimates had artificially deflated market prices for natural gas. EIA has revised its estimation method to make use of the most up-to-date administrative data available, even though the information may be somewhat incomplete for some states.

The possibility of rapid changes (economic, technological, etc.) disproportionately affecting the sample vs. the non-sample units in a population poses a challenge for agencies employing cutoff samples in important surveys. The benefits of cutoff sampling may nevertheless render it an attractive option for many applications, and it is to these that we now turn.

2. Reasons for Using Cutoff Samples

Efficiencies are the primary reason for choosing cutoff sampling over a probability-based sampling scheme in which every unit in the target population has a positive probability of selection. Several types of efficiency gains can be achieved by using cutoff sampling under various conditions.

2.1. Physical Efficiencies

Units in the target population may be inaccessible, for various reasons, in some survey applications. In a household telephone survey, for example, households without telephones are often excluded from the survey although they may be in the target population. This is a type of cutoff sample. In order to compute estimates for the target population based on the cutoff sample data, one may assume that (a) the households without telephones are similar to the sample households with respect to the variables of interest or (b) non-telephone households are so rare in the target population that their contribution to the estimates would be negligible. Alternatively, one may use a model, incorporating other data and/or assumptions, to estimate the variables of interest for the non-telephone households. Similar examples include internet-only surveys or even household surveys conducted by personal visit when the homeless and institutionalized populations are not surveyed but are included in the target population.

2.2. Costs, Respondent Burden, and Data Quality

Reducing costs and respondent burden is perhaps the primary motivation behind cutoff sampling. 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 so poor that the potential error caused by including this subset in the sample outweighs the statistical advantage gained through use of a more inclusive sample. Cutoff sampling may therefore produce more accurate estimates, in terms of total survey error. This phenomenon might be viewed as another

example of the "80/20 rule": inaccuracies in data from a small subset of the sample accounts for a large portion of the error in the sample-based estimates.

The data quality issue may also be related to the cost issue. Although data quality may be improved by meeting with individual respondents and instructing them in providing accurate data, the cost of doing this for a large number of respondents may be inordinately high.

2.3. Statistics Estimated

Cutoff sampling may be more cost effective than probability sampling, while still producing accurate estimates, if one is only interested in the ratio of two variables of interest as opposed to a population total. Average yield per acre is one example. Other examples include cases where only change over time is to be estimated, rather than population characteristics at a particular point in time. Cutoff sampling may be efficiently used to estimate "links" (ratios indicating change from a prior time period) for use in a link-relative estimator or for estimating a price index (see, for example, Dorfman et al. 2006). The technique could also be quite useful when one is only attempting to find anomalies or proof of the existence of a particular trait in nonhomogeneous members of the target population.

In general, cutoff sampling is an effective sampling methodology when one has a good working knowledge of the characteristics of the total population and subpopulations of interest and is willing to take some calibrated 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.

3. Federal Agencies Experiences with Cutoff Sampling

3.1. Selecting the Sample

The first step in implementing cutoff sampling is to settle on a measure of size (MOS) that will be used to stratify the sample. A variable or variables of primary importance for the data collection is usually used.

For more focused surveys (those with few data items), the choice of MOS is often obvious. For example, in BEA's surveys of international transactions in services, cutoffs are established using the size of service imports and exports. In EIA's surveys of energy producers, the MOS is usually an energy output measure—usually quantities of energy products either produced or sold (e.g., cubic feet of natural gas produced, kilowatt hours of electricity sold).

More than one MOS may be used to make sure that all cells of interest are sufficiently populated. For example, EIA's monthly Power Plant Operations survey uses power plant nameplate capacity (how much electricity a plant would generate at full capacity, assuming optimal fuel types) and annual electricity generation reported on the most recent Annual Electric Generator report. After an initial cutoff sample is selected based on these two MOS, the sample count for each publication cell is reviewed. As needed, sample is added within cells to ensure minimal within-cell sample counts.

Multiple MOS may be used for surveys with many data items. For example, BEA's surveys of direct investment have a large number of data items covering a number of areas of economic activity of multinational corporations. These surveys stratify firms by the maximum of the absolute value of three items: assets, sales, and net income. Companies that are small according to this composite size measure will have relatively little activity in most other activities of interest.

3.1.1. Stratification (e.g., by age, other variables of interest)

The next step in implementing cutoff sampling is to stratify the sample according to the MOS selected. The simplest form of stratification is to ask units above a MOS cutoff to report and exclude units below the cutoff. In BEA's surveys of international transactions in services, size cutoffs are established separately for transaction direction (exports or imports) and apply separately to the combination of transaction direction and service type (e.g., rights related to industrial processes and products, computer and data processing services, legal services, financial services). So, for instance, if a company has exports in financial services exceeding the export cutoff value, it must respond to questions regarding those transactions, but not necessarily to questions about its imports of legal services unless these exceed the import cutoff.

For more complex surveys, there may be additional stratification. One method of stratification is to ask larger firms to report on more data items. A survey might consist of increasing cutoffs *A*, *B*, and *C*. The sub-survey with the lowest cutoff value (*A*) will consist of some core survey items; the sub-survey with the next lowest cutoff value (*B*) will contain additional items and more detail on the core items. The sub-survey with the highest cutoff (*C*) will consist of the largest number of items and the most detail on core items. Therefore, the largest companies file all three sub-surveys, a group of intermediate-size companies files *A* and *B*, and the smallest file only *A*. Totals are calculated for all items, with items in *A* used (along with data from the benchmark, when applicable) to impute for items in *B* and *C*. This method is used by BEA in its surveys of foreign direct investment.

An alternative stratification is to ask for the same number of items, but varying the probability of selection. The Census Bureau's economic surveys are required to provide reliable estimates for detailed industries, geographies and sometimes product classes. These sub-populations vary considerably so the Census may sample each subpopulation (stratum) independently and in a way that is proportional to that of the total population. For most economic surveys, a number of units are also included in the sample with certainty; i.e. their probability of selection for the survey is 1.00.

In Census's Annual Survey of Manufactures (ASM), establishments are sampled with probabilities ranging from .05 to 1.00. Each industry and product class are considered to be a separate population. Using variable reliability constraints, each establishment within a given population is assigned an initial probability of selection that reflected its relative importance within the population. Establishments producing products in multiple product classes receive multiple initial probabilities. The final probability of selection for a specific establishment is defined as the largest of their initial probabilities.

This method of assigning probabilities is motivated by a desire to produce reliable estimates of both product class and industry shipments. The high correlation between shipments and employment, value-added, and other general statistics assures that these variables will also be well represented. For sample selection purposes, each establishment is assigned to an industry stratum. Within each of the 473 industry strata, an independent sample is selected using the final probability of selection associated with the establishments classified within the stratum. A fixed-sample size methodology is used to assure that the desired sample size is realized.

Non-size information may be used in conjunction with a MOS in stratification. For example, EIA often uses geographic stratification. The sampling strata are usually states or groups of states (e.g., Census divisions).

3.1.2. Optimizing Cutoff Points

Choosing cutoff points requires balancing data quality with cost and respondent burden. The costs of obtaining a survey response consist of both the staff resources required to collect, edit, and enter the response and the burden on the respondent. In the skewed population distributions where cutoff sampling is typically used, there are a large number of small entities that contribute little to aggregate data. For both types of survey costs, a small entity response is nearly equivalent to a large entity response, at least in comparison to the large difference between the two in the share of data accounted for by each. Obtaining reliable data from the smallest units will pull resources from obtaining and processing reports from large units. In general, total survey error is reduced when a large entity, rather than a small entity, is in the sample.

When setting cutoffs, there is a minimum level of coverage of a cell for a statistical agency to have confidence in the estimate of that cell. EIA's general practice is to select sample cutoff points to ensure minimal sample coverage rates for important publication cells. The target coverage rates are usually determined through past experience. In some cases, however, model-based error measures are used to indirectly determine cutoff points by providing minimal sample counts for publication cells (e.g., in the monthly Power Plant Operations Report discussed above).

In the case of the ASM establishments in the 2002 Economic Census - Manufacturing that satisfied any of the following criteria are included in the sample with certainty:

(1) Total employment for the establishment for 2002 is greater than or equal to 1,000 or their receipts are greater than $50 million;
(2) The establishment is identified as one of the ten largest establishments within the industry (employment);
(3) There are a few industries that have very few establishments (e.g. flat glass). For these industries, sampling is not efficient; therefore, every establishment is included in the sample. There are 6 such industries that account for about 200 establishments.
(4) The establishment is located within a State where there are less than 20 additional establishments in the same North American Industry Classification System (NAICS) group (NAICS group is defined as the set of NAICS industries that have the same first four digits.).

Collectively, the ASM includes approximately 9,700 establishments that are selected with certainty. These establishments accounted for approximately 44 percent of the total value of shipments in the 2002 Economic Census, Manufacturing.

The availability of alternative data sources may allow cutoffs to be increased without serious harm to data quality. For example, the Census Bureau decided to increase cutoffs substantially for the 2004 ASM panel since (less detailed) administrative information was available to impute data for the smallest units.

3.1.3. Adding New Units (e.g., Births)

Populations are continually changing. Despite tightly controlled procedures for maintaining the sample, some deterioration in coverage and representation might occur-- "drift". In order to avoid drift and retain the representativeness of the sample, surveys may need to include births (newly operating units). Existing entities that had previously been below the cutoff may grow above the cutoff.

It is not always necessary to add births of small entities. The impact of the smallest units on aggregate data may be so small as to not justify the cost of surveying them. In EIA cutoff sample

surveys, no effort is made to add units whose measures of size fall below the cutoff points, unless these units fall into special categories (e.g., nuclear power plants). Small firms are not added simply because they are new.

Births can be discovered in different ways. New units covered by the survey may be required to identify themselves. For example, in BEA's surveys on transactions in international services, a firm with a known history in exporting R&D services but no history of importing any service is required to report imports of, say, database services as soon as they exceed the cutoff value, whether the existence of those imports is known to BEA or not. In addition, in BEA's surveys on direct investment abroad, a known U.S. multinational company is required to provide a survey response for each new (and, likely, unknown to BEA) affiliate it acquires or establishes. Self selection in this manner is not complete in practice, but it does generate a more complete frame than if BEA were to have to identify each new unit on its own. Census adds all births that are part of a multiunit enterprise in the year of birth. For single unit births, data are imputed in the first year and a number of births are added to the cut-off sample in subsequent years (distributed by industry). Remaining single unit births do not file a report and data are imputed based on administrative records data.

Statistical agencies typically must actively search out births, since new births may not be aware of reporting requirements or choose not to comply. BEA uses a variety of means of identifying new units in the population (news accounts, privately produced business databases, etc.). BEA does not have access to administrative records for these surveys, but on occasion links records with the Bureau of Census (or other agencies) as part of data sharing agreements. These links can identify companies in the relevant population that are not currently on BEA's frame.

Existing entities that grow into the cutoff sample are also a source of drift. To maintain the cutoff sample, BEA sends reports to all known units in the frame with the requirement that responses be provided if the units exceed the cutoff size in the reference period. Census uses administrative data to track whether small units have grown into the cutoff sample.

3.2. Estimation Methods for Cutoff Samples x

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".[1] The Census Bureau makes use of administrative data, census information, and prior year data to minimize the potential of sampling error.[2]

3.2.1. Standard approach: use sample data to estimate change over time.

Let the population U be divided into a take-none stratum U_N and a sampled stratum U_S. Let Y, Y_N, and Y_S be the sum of values of a positive variable of interest y attaching to each of the units in U, U_N and U_S, respectively. Let $S_s \subseteq U_s$ be a sample drawn from U_s.

- Negligible cutoff portion: Assume $Y_N = \sum_{k \in U_N} y_k$ is very small; then

$$\hat{Y}_S \approx \sum_{k \in S_s} \frac{y_k}{\pi_k},$$

[1] Importantly, model failure will not be known until a later date.

[2] The estimates are also subject to various types of nonsampling error. Examples of nonsampling error include measurement error and processing error. Nonsampling error is inherent to all surveys, including complete censuses.

where π_k is the inclusion selection probability for unit $k \in U_s$.

- Ratio adjustment method (Sarndal, 2003): Let x_k be an auxiliary variable for unit $k \in U$. We use the information from the sample data to estimate the relationship between x_k and the variable of interest y_k. We then estimate the population value Y by assuming that the estimated relationship holds within the non-sampled population:

$$\hat{Y} = \sum_U x_k \frac{\sum_{S_s} y_k/\pi_k}{\sum_{S_s} x_k/\pi_k}.$$

The most common way of estimating data for units that fall below a cut-off threshold is to extrapolate from reported data using ratio estimators. The data that a unit reported in an earlier period is adjusted for the change between the earlier reference date and the current period using growth rates calculated from similar units above the cutoff. Here are a few examples:

For BEA surveys, in most cases, a benchmark survey covering the full universe of reporters is performed periodically (usually every five years) to provide a basis for estimating data from reporters that are excluded in non-benchmark years. In non-benchmark years, data are estimated for those universe members that are excluded from the sample by cut-off sampling by extrapolating reported data using ratio estimators. The estimation methodology uses data that a reporter filed in an earlier year and adjusts it for the current year using growth rates estimated from available data.

BEA's foreign direct investment program conducts a census (called a benchmark) of all firms in the sample frame every five years. Firms that enter the sample frame are required to file a report which provides a basis for extrapolation. The growth rates are computed from reports filed by units above the cut-off. The ratio may be estimated from a subsample of units that most closely resemble the units below the cut-off (e.g., firms in the same industry group and geographical area) to estimate the ratios. This approach, similar to a post-stratification, is called a *matched sample* estimation approach.

The ratio estimation methodology can be used to fill in components of aggregate data. For example, BEA's foreign direct investment program asks medium-sized firms for a limited number of items on a short form to reduce their reporting burden. Ratio estimation is used to fill in data items that are not on the short form but are on the long form submitted by the largest firms. Using a sample of reporting affiliates, ratios linking reported data to unreported data are computed. For example, current liabilities are computed as reported total liabilities multiplied by the industry ratio of current liabilities to total liabilities.

In some EIA surveys, data for the small firms is estimated by assuming that the percentage change over time is the same for the sample and non-sample companies. The assumption can be applied at an aggregate level, through a ratio adjustment within a geographic area, or at the company level, by imputing data for the small firms as if they were non-responding sample firms. Nonresponse adjustment cells may be defined using annual data on geography and other company characteristics. EIA collects annual data from all firms within the target population and monthly data from the large firms only. The change over time, estimated from the large firms' data, is then used to adjust the most recent annual data for the small firms.

3.2.2. Using Administrative Data for Nonsampled Units

Statistical uses of administrative records are more popular and important for federal agencies now. According a to subcommittee chartered by the Federal Committee on Statistical Methodology to examine statistical uses of administrative records, the Internal Revenue Service (IRS), the Social Security Administration (SSA), and U.S. Census Bureau have all devoted considerable resources to create and maintain a wealth of administrative records. Statistical research on administrative records use and linkages is one of the top 5 research priorities of the US Census Bureau. For example, for the take-none stratum (also called the "non-mail stratum") from the Annual Survey of Manufactures (ASM), the Census Bureau relies on a mix of current and prior year establishment-specific information for imputation. Payroll and employment information are directly tabulated from the administrative-record data provided by the Internal Revenue Service (IRS) and the Social Security Administration (SSA). Payroll information then drives the imputation for all the remaining variables. The imputation makes use of prior year data from the economic census or the prior year's survey. If prior year data are not available for a particular establishment, then industry averages are used (electricity is an exception; geographic averages are used for the electricity industry.)[3]

The Annual Retail Trade Survey (ARTS) and Service Annual Survey (SAS) provide further examples. The US Census Bureau has applied cut-off sampling methods for the SAS since 1999 and for the ARTS since 2000. The cut-off sampling method is only applied to single-establishment businesses. Census then estimates the excluded small single-establishment businesses for ARTS and SAS by imputing available administrative information from the Economic Census, the Business Register, and the Small Business Administration (SBA).

3.2.3. Econometric Modeling Approach

Most of the estimates derived for the mail stratum from the Annual Survey of Manufactures (ASM) are computed using a difference estimator. At the establishment level, there is a strong correlation between the current-year data values and the corresponding historical data values. At each level of aggregation, an estimate of the difference between the current reference period value and the base reference period value is computed for each item from the sample and then added to the corresponding base-year value. For example, the 2002 Census of Manufactures values are used as base year values for the 2002-2006 ASM. Because of the positive year-to-year correlation, the estimated differences between the current year and base year (i.e., census year) values are more reliable than comparable estimates developed from the current sample data alone.

Note that, in the ASM, some variables lack positive year-to-year correlation. (Examples include the capital expenditures variable.) A standard linear estimator is used for these variables.

4. Error Measures for Cutoff Sample Estimates

In general, a sampling strategy consists of a sample plan (design) and a method of estimation, together aimed at estimating one or more target quantities, subject to agency resource constraints and limits on allowable respondent burden. The twin goals of a sampling strategy are accurate estimation and sound inference. We can aim at optimal accuracy or accuracy that meets some standard. Inference-the assessing of error- gives us an idea of the degree of accuracy we have achieved.

[3] For example, the ASM makes use of establishment-specific information from the 2002 Census of Manufactures to impute 2004-2008 take none cases.

To make this a bit more specific, consider the following scenario: if Y is a population target, then estimation is said to be accurate if there is good reason to think that the estimate $\hat{Y}$ satisfies $\left|\hat{Y} - Y\right| / Y \leq \varepsilon$, where ε is small; in other words, the relative error is small. For inference we typically rely on confidence intervals, that is, an interval I (based on $\hat{Y}$ and also usually a variance estimate) such that $Y \in I$ a certain specified percent of the time.

Cutoff sampling typically entails omitting units that are in some respect much smaller than the units that are sampled, but this is not always the case. For example, in Haziza et al (2010), the portion cutoff is the part of the population of establishments inaccessible to electronic sampling, which might include some large as well as small firms. Cutoff sampling therefore poses special difficulties with regard to *assessing* the accuracy of the estimates and performing inference through confidence intervals. This is essentially due to the fact that an estimate based on a cutoff sample may have a bias, the magnitude and direction of which may be hard to assess. This is an especially important concern, mentioned in the *OMB Standards and Guidelines for Statistical Surveys* (2006) to which the literature on cutoff sampling has so far paid little attention. The OMB guidelines (Section 1.2) state, "Agencies must develop a survey design, including defining the target population . . . and selecting samples using generally accepted statistical methods (e.g., probabilistic methods that can provide estimates of sampling error). Any use of nonprobability sampling methods (e.g., cut-off or model-based samples) must be justified statistically and be able to measure estimation error."

Despite these concerns, cutoff sampling is used for many federal establishment surveys. Its use rests on the supposition, borne out by experience, that in some circumstances it can be the most accurate way to sample within the constraints of a given budget.

4.1. Examining the Accuracy of Estimates Based on a Cutoff Sample

We use the notation of Section 3 above, i.e., we consider a population U comprising strata U_N (not sampled) and U_S. (sampled). Let Y, Y_N, and Y_S be the sum of values of a positive variable of interest y attaching to each of the units in U, U_N and U_S respectively. Let $\hat{Y}_N$ and $\hat{Y}_S$ be estimates of Y_N and Y_S and $\hat{Y} = \hat{Y}_S + \hat{Y}_N$ an estimate of Y. The important point is that $\hat{Y}_N$ and $\hat{Y}_S$ are different in *kind*: $\hat{Y}_S$ will be a standard design-based or model-based (Valliant et al. 2000) estimator; $\hat{Y}_N$ will be based on some sort of extrapolation from the data at hand, possibly guided by auxiliary or historic data. Thus getting bounds on the relative error of $\hat{Y}_S$ will be straightforward, whereas bounds on the relative error for $\hat{Y}_N$ will tend to be more conjectural and tenuous.

Here are two simple results worth noting:

Result 4.1: If $\left|\hat{Y}_N - Y_N\right| / Y_N \leq \varepsilon_N$ and $\left|\hat{Y}_S - Y_S\right| / Y_S \leq \varepsilon_S$, then $\left|\hat{Y} - Y\right| / Y \leq \eta = \max\left(\varepsilon_N, \varepsilon_S\right)$.

Result 4.2. If (a) $Y_N / Y \leq \varepsilon$, (or equivalently the "coverage" Y_S / Y exceeds 1-ε), (b) $0 \leq \hat{Y}_N / Y_N \leq B$, and (c) $\left|\hat{Y}_S - Y_S\right| / Y_S \leq \varepsilon_S$, then $\left|\hat{Y} - Y\right| / Y \leq \eta^* = (B+1)\varepsilon + \varepsilon_S$.

In particular, if $\hat{Y}_N = 0$ (that is, $\hat{Y} = \hat{Y}_S$) and conditions (a), (c) are met, then $\left|\hat{Y} - Y\right| / Y \leq \eta^* = \varepsilon + \varepsilon_S$.

In practice, the conditions of Result 4.2 are met more often than those of Result 4.1. Proofs of these results are given in the Appendix.

4.2. Inference with Estimates from a Cutoff Sample

We now turn to the question of inference. Strict confidence intervals are doubtless impossible, but something very much like them may be feasible.

We recall Bonferroni's useful inequality:

$$P(A_1A_2\cdots A_n) \geq P(A_1)+P(A_2)+\cdots+P(A_n)-(n-1).$$

This can enable us to go from assessments of probabilities concerning the separate components of *Y* to a probabilistic statement regarding *Y* itself. For example, suppose based on sampling (or well founded modeling) properties we are 95% certain that $\left|\hat{Y}_S - Y_S\right| / Y_S \leq \varepsilon_S$ and, based on historical or other considerations, 90% certain that $Y_N / Y \leq \varepsilon$. Then, if we take $\hat{Y} = \hat{Y}_S$, we can be 85% certain that $\left|\hat{Y} - Y\right| / Y \leq \eta^* = \varepsilon + \varepsilon_S$.

This sort of reasoning enables us to construct interval estimates for *Y*.

4.2.1. Quasi-Confidence Intervals (for totals)

In the discussion to follow we speak of two probabilities *p* and *p** which are slightly different in nature. The probability *p* will derive from standard properties of well constructed confidence intervals, while *p** will derive from experience of historical data and may require some exercise of judgment.

Case A: $\hat{Y} = \hat{Y}_S$. Suppose that we have a $p = (1-\alpha)$ confidence interval $\hat{Y}_S \pm z\hat{\sigma}$ for Y_S and that $Y_N / Y \leq \varepsilon$ with probability *p**. As noted above, the latter condition implies $Y_S \geq (1-\varepsilon)Y$. Then, with probability at least $p + p^* - 1$,

$$\left|\hat{Y} - Y\right| = \left|\hat{Y}_S - Y_S - Y_N\right| \leq \left|\hat{Y}_S - Y_S\right| + \left|Y_N\right| \leq z\hat{\sigma} + \varepsilon Y \leq z\hat{\sigma} + \frac{\varepsilon}{1-\varepsilon} Y_S \leq z\hat{\sigma} + \frac{\varepsilon}{1-\varepsilon}\left(\hat{Y}_S + z\hat{\sigma}\right)$$
$$= \frac{\varepsilon}{1-\varepsilon}\hat{Y}_S + \frac{z\hat{\sigma}}{1-\varepsilon}.$$

Special Case. If the sample *s* consists solely of certainties, i.e. $U_S = S_S$, and there is no non-response, then $\hat{Y} = \hat{Y}_S = Y_S$, $\hat{\sigma} = 0$, and *p* = 1. In this case we get a *p** interval $\left|\hat{Y} - Y\right| \leq \left(\frac{\varepsilon}{1-\varepsilon}\right) Y_S$.

Case B. We allow $\hat{Y}_N \neq 0$ and take $\hat{Y} = \hat{Y}_N + \hat{Y}_S$. Suppose (a) $Y_N / Y \leq \varepsilon$ and $\hat{Y}_N / Y_N \leq B$ with probability *p** and (b) we have a $p = (1-\alpha)$ confidence interval $\hat{Y}_S \pm z\hat{\sigma}$ for Y_S. Then (b) implies $\left|\hat{Y}_S - Y_S\right| / Y_S \leq \varepsilon_S = z\hat{\sigma} / Y_S$ and applying Result 4.2 we have, with probability *p* + *p** - 1,

$$\left|\hat{Y} - Y\right| / Y \leq \eta^* = (B+1)\varepsilon + \varepsilon_S$$

so that

$$\left|\hat{Y} - Y\right| \leq \eta^* Y \leq \eta^* \frac{Y_S}{1-\varepsilon}$$

or

$$\left|\hat{Y}-Y\right| \le \left[\left(B+1\right)\varepsilon + \frac{z\hat{\sigma}}{Y_S}\right]\frac{Y_S}{1-\varepsilon} = \frac{1}{1-\varepsilon}\left[\left(B+1\right)\varepsilon Y_S + z\hat{\sigma}\right]$$
$$\le \frac{1}{1-\varepsilon}\left[\left(B+1\right)\varepsilon\left(\hat{Y}_S + z\hat{\sigma}\right) + z\hat{\sigma}\right]$$

Provided that *B*, *p** and ε are well chosen, we would expect this to be a conservative interval. In the case where *B* = 0, *Case B* reduces to *Case A*.

Case C. Condition (a) of *Case B* is replaced by (a') $\left|\hat{Y}_N - Y_N\right| \Big/ Y_N \le \varepsilon_N$ with probability *p**; (b) as above. Note that (a') implies

$$Y_N \le \hat{Y}_N + \varepsilon_N Y_N \le \hat{Y}_N + \varepsilon_N\left(\hat{Y}_N + \varepsilon_N Y_N\right) = \hat{Y}_N\left(1+\varepsilon_N\right) + \varepsilon_N^2 Y_N \le$$
$$\cdots \le \hat{Y}_N\left(1+\varepsilon_N + ... + \varepsilon_N^k\right) + \varepsilon_N^{k+1} Y_N \approx \frac{\hat{Y}_N}{1-\varepsilon_N}, \text{ for } \varepsilon_N \text{ small.}$$

Then

$$\left|\hat{Y}-Y\right| \le \left|\hat{Y}_N - Y_N\right| + \left|\hat{Y}_S - Y_S\right| \le \varepsilon_N Y_N + z\hat{\sigma} \le \frac{\varepsilon_N}{1-\varepsilon_N}\hat{Y}_N + z\hat{\sigma},$$

with probability *p* + *p** -1.

These results on forming quasi-confidence intervals deserve further exploration, especially on actual population data, but they do suggest that inference based on cutoff sampling is achievable.

4.2.2. Quasi-Confidence Intervals (for ratios)

So far, we have considered inference for totals of a variate of interest. Cutoff sampling is often thought of as being especially suitable for the estimation of ratios. We here give some brief consideration as to how we might carry out inference, i.e. construct quasi-confidence intervals, in this important case.

Suppose two *positive* variables of interest *x* and *y*. Let $X = \sum_{i\in U} x_i$ and $Y = \sum_{i\in U} y_i$ and suppose the target of interest is the ratio *R* = *Y*/*X*. Let $X_k = \sum_{i\in U_k} x_i$, $Y_k = \sum_{i\in U_k} y_i$ and $R_k = Y_k/X_k$ for $k \in \{N, S\}$. We note the following relationship:

$$R = R_S + r, \text{ where } r = \frac{Y_N - R_S X_N}{X} \quad . \qquad (4.2.1)$$

Suppose (a) R_S is estimated by $\hat{R}_S$ yielding a confidence interval $\left|\hat{R}_S - R_S\right| \le \varepsilon_S$ with probability p_2. Suppose also (b) with probability p_1 we have conditions $Y_N \le \varepsilon_y Y$ and $\varepsilon_{xL} X \le X_N \le \varepsilon_{xU} X$. Note: one default possibility for ε_{xL} is $\varepsilon_{xL} = 0$, which holds trivially. Often X_N will be known and we can take $\varepsilon_{xL} = \varepsilon_{xU} = \varepsilon_x$. Note that (b) implies coverages

$$Y_S \ge \left(1-\varepsilon_y\right)Y \text{ and } \left(1-\varepsilon_{xU}\right)X \le X_S \le \left(1-\varepsilon_{xL}\right)X \ . \qquad (4.2.2)$$

If the epsilons are small enough we can take $\hat{R}_S$ as an estimate of *R* itself, and the following result gives a (non-symmetric) quasi-confidence interval for *R*.

Result 4.3. Under conditions (a) and (b), $(1-\varepsilon_{xU})\left(\hat{R}_S-\varepsilon_S\right) \le R \le \frac{1-\varepsilon_{xL}}{1-\varepsilon_y}\left(\hat{R}_S+\varepsilon_S\right)$, with probability $p = p_1 + p_2 - 1$.

Proof. By eqtn (4.2.2), $\frac{\left(Y_S/Y\right)}{\left(X_S/X\right)} \ge \frac{1-\varepsilon_y}{1-\varepsilon_{xL}}$, that is, $\frac{R_S}{R} \ge \frac{1-\varepsilon_y}{1-\varepsilon_{xL}}$ and then, invoking (a),

$$R \le \frac{1-\varepsilon_{xL}}{1-\varepsilon_y} R_S \le \frac{1-\varepsilon_{xL}}{1-\varepsilon_y}\left(\hat{R}_S + \varepsilon_S\right).$$

We also have

$$R = \frac{Y}{X} \ge \frac{Y_S}{X} \ge \frac{Y_S}{X_S}(1-\varepsilon_{xU}) \ge \left(\hat{R}_S - \varepsilon_S\right)(1-\varepsilon_{xU}),$$

giving the lower bound.

*Special Case***.** When $U_S = U_A$, i.e., U_S is a certainty stratum, (and there is no non-response) we have $\hat{R}_S = R_S$, $\varepsilon_S = 0$, and $p_2 = 1$. Then under conditions (b), $(1-\varepsilon_{xU})R_S \le R \le \frac{1-\varepsilon_{xL}}{1-\varepsilon_y} R_S$, with probability p_1.

Remark. The estimator $\hat{R} = \hat{R}_S$, which is probably the most commonly used estimator of ratios in cutoff sampling, is equivalent to assuming either that $Y_N/X_N = Y_S/X_S$ *or* that Y_N and X_N are at their lowest possible bounds under conditions (b). But if (b) holds we can ask whether a somewhat improved estimator is possible. One possibility that suggests itself is to take $\hat{Y}_N = \varepsilon_y' Y$ and $\hat{X}_N = \varepsilon_x' X$, where $0 \le \varepsilon_y' \le \varepsilon_y$ and $\varepsilon_{xL} \le \varepsilon_x' \le \varepsilon_{xU}$. Methods for selecting $\varepsilon_x', \varepsilon_y'$ require investigation. One might also consider more explicit model-based estimators. In either case, it may be regarded as open how best to construct intervals around these estimates.

5. Other Issues

We here touch briefly on some other issues that pertain to cutoff sampling.

5.1. The role of a Take Portion Stratum

In general, cutoff sampling divides the population into two broad strata, the take none and the take sample: $U = U_N \cup U_S$, and U_S can, in general, be further sub-divided: $U_S = U_P \cup U_A$, where U_P is a take portion stratum, in which only some units are selected, usually by some probabilistic mechanism, and U_A is a take all stratum, in which every unit is selected. Probably most commonly used is a design where $U_S = U_A$ and the take portion stratum is empty. Baillargeon and Rivest (2009) and Benedetti *et al.* (2010) suggest the utility of having U_P non-

empty, from the viewpoint of efficiency. Its presence also brings cutoff sampling closer to the standard design-based canon, for in that approach, certainty units can represent only themselves. From a model-based view, a model based on data from U_P may well reflect better than the certainties what the data are like in U_N. Typically, U_A would be the largest units, U_P would be midsize units, and U_N would be the smallest units.

Suppose we are interested in the total *Y* and an auxiliary *x* is available for all units in *U*. One estimator we can consider takes the form $\hat{Y} = \sum_s y_i + \hat{\beta}\sum_{\bar{s}\cap P} x_j + \hat{\beta}'\sum_N x_j$, where *s* represents the sample, $\bar{s}$ the non-sample and $\hat{\beta}$, $\hat{\beta}'$ are ratio estimates calculated from the sample data. Usually we are inclined to take $\hat{\beta} = \hat{\beta}'$, but, depending on the population structure, we might want to allow that the coefficient for the take none units differs from that for the unsampled take portion units. For example, we might want to use some of the certainty units in calculating $\hat{\beta}$ and restrict ourselves to some smaller units in the available take portion stratum for $\hat{\beta}'$. There are many as yet unexplored possibilities.

5.2. The Possibility of Diagnostics

Consider the example of the last sub-section. We may be able to get some handle on how well extrapolation to U_N will work, by calculating several values of $\hat{\beta}'$ using shrinking amounts of the sample data available in U_P. If these values remain close to each other or show some trend as the *x*'s used get restricted and closer in size to those in U_N, then we have some reassurance that the extrapolation is not wild. On the other hand, if the $\hat{\beta}'$'s jump around, then we may wish to state our conclusions more cautiously.

5.3. The Issue of Outliers

One of difficulties that cutoff sampling must reckon with is the increased chance of not detecting outliers. If a unit with small *x*, where *x* is say *y* at a previous time period, undergoes sudden expansion, then there will be no way to ascertain this if *x* belongs to U_N. Perhaps no other factor can as much undermine inferences based on cutoff sampling as outliers. Estimates and statements of conclusions drawn from cutoff sampling should perhaps always incorporate an acknowledgment of the possibility of "poppers" and what effect they might have. Recalling historical examples may be salutary.

On the other hand, it should be recognized that cutoff sampling is not unique in this regard. From a practical standpoint, "poppers" are almost as unlikely to be picked up by standard sampling methods, especially probability proportional to size algorithms. In fact, there may be a greater risk in this case because, reassured by our sound methodology, we may be less likely to issue precautions.

6. Conclusions

An interagency research group formed in 2009 to examine the issue of cutoff sampling in federal establishment surveys found that the technique was employed in many surveys within the federal statistical system. Although traditional design-based error measures are impossible to compute for cutoff sample based estimates, meaningful error measures can often be derived and used in statistical inference. Given the cost savings that cutoff sampling can achieve in many surveys, general acceptance of the technique is warranted. Agencies wishing to employ cutoff sampling

must, however, give considerable forethought to the issues of error estimation and the verification of any statistical models employed in the estimation process.

References

Baillargeon S. and Rivest, L.P. (2009) "A General Algorithm for Univariate Sampling," *International Statistical Review*, Volume 77, Issue 3, pp. 331-344, December, 2009.

Benedetti, R., Bee, M., and Espa, G. (2010) "A Framework for Cut-off Sampling in Business Survey Design," *Journal of Official Statistics, Vol. 26*, No. 4, pp. 651–67.

Dorfman, A. H., Lent, J., Leaver, S., and Wegman, E. (2006). "On Sample Survey Designs for Consumer Price Indexes," *Survey Methodology, Vol. 32*, No. 2, December, 2006. Statistics Canada, Catalogue No. 12-001-XPB.

Haziza, D., Chauvet, G. & DeVille, J-C (2010) "Sampling and Estimation in the Presence of Cutoff Sampling", *Australian and New Zealand Journal of Statistics*, **52**, 303-319

Valliant, R., Dorfman, A.H., and Royall, R.M. (2000) *Finite Population Sampling and Inference*: *A Prediction Approach*, Wiley, New York

U.S. Office of Management and Budget (2006) *Standards and Guidelines for Statistical Surveys*, www.whitehouse.gov/sites/default/files/omb/inforeg/statpolicy/standards_stat_surveys.pdf.

Appendix

Proofs of Results 4.1 and 4.2

Proof of Result 4.1. Recall assumption that components of Y are positive. We have

$$\left|\hat{Y}-Y\right| = \left|\hat{Y}_N - Y_N + \hat{Y}_S - Y_S\right| \le \left|\hat{Y}_N - Y_N\right| + \left|\hat{Y}_S - Y_S\right| \le \varepsilon_N Y_N + \varepsilon_S Y_S$$

$$\le \max\left(\varepsilon_N, \varepsilon_S\right)\left(Y_N + Y_S\right) = \max\left(\varepsilon_N, \varepsilon_S\right) Y .$$

Hence, $\left|\hat{Y}-Y\right| \Big/ Y \le \eta = \max\left(\varepsilon_N, \varepsilon_S\right)$.

Proof of Result4.2. $\dfrac{\left|\hat{Y}-Y\right|}{Y} \le \dfrac{\left|\hat{Y}_N - Y_N\right|}{Y} + \dfrac{\left|\hat{Y}_S - Y_S\right|}{Y} \le \dfrac{\left|\hat{Y}_N\right| + \left|Y_N\right|}{Y} + \dfrac{\left|\hat{Y}_S - Y_S\right|}{Y_S}$, since $Y_S \le Y$.

Hence

$$\frac{\left|\hat{Y}-Y\right|}{Y} \le \frac{\left(B+1\right)Y_N}{Y} + \frac{\left|\hat{Y}_S - Y_S\right|}{Y_S} \le \left(B+1\right)\varepsilon + \varepsilon_S .$$