Inference in Cutoff Sampling

October 2012

Alan H. Dorfman,

U.S. Bureau of Labor Statistics
2 Massachusetts Ave NE
Washington, DC 20212

Abstract

In cutoff sampling, inference -— for example, interval estimates with associated alpha-levels —- is problematic. Design-based samplers do not find an adequate random design on which to base variance estimates. Model-based samplers worry that gaps in information can lead to biases. We nonetheless describe some schemes for inference in cutoff sampling.

Key Words: confidence intervals; probability proportional to size sampling

1. Introduction

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 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- is the assessment of the accuracy of ones estimate and requires that an interval be placed around the estimate with an associated probability of the interval containing the unknown target.

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 $\left|\hat{Y}-Y\right|\big/Y \le \varepsilon$, where ε is small (in the eyes of the user); in other words, the relative absolute 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 is the practice of taking cutoff samples, where by a cutoff sample we mean a sample from which some portion of the population is deliberately excluded. 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. We will focus on the typical case, where the small fry are excluded. In any case, cutoff sampling 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. In this regard, it is akin to the better known (and more widely accepted) procedures of Small Area Estimation and Non-response Adjustment .

Bridgman et al. (2011) suggested some theoretical results by which interval estimates might be constructed for cutoff samples, but gave no empirical examples. One problem is that suitable populations for studying cutoff sampling tend to lie behind a veil of confidentiality, and not be publishable. In this paper, we construct a set of (we hope realistic) populations to illustrate the effects of various approaches to inference under cutoff sampling. We shall take the target to be the total *Y* over the population of a positive variable of interest *y.*

2. Some Results bearing on Inference in a Cutoff Sample

Consider a population *U* comprising strata U_N , the not sampled or *take none stratum,* and U_S the sampled stratum. Often in practice each unit in U_S will be sampled with certainty, but this need not be the case. Let *Y*, $Y_{N,}$ and Y_S be the sum of values of the (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 $\hat{Y} = \hat{Y}_N + \hat{Y}_S$ an estimate of *Y.* Inevitably $\hat{Y}_N$ and $\hat{Y}_S$ are different in kind: $\hat{Y}_S$ will be a standard design-based or model-based (e.g. 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. $\hat{Y}_N$ will use an explicit or implicit model for the data on U_N, precisely where there is *no data to verify the model* (as is also the case often enough with small area estimation and non-response adjustment.) 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.

We repeat some of the results from Bridgman et al. (2011):

Result 1: If $\left|\hat{Y}_N - Y_N\right| \big/ Y_N \le \varepsilon_N$ and $\left|\hat{Y}_S - Y_S\right| \big/ Y_S \le \varepsilon_S$, then $\left|\hat{Y} - Y\right| \big/ Y \le \eta = \max\left(\varepsilon_N, \varepsilon_S\right)$.
Proof. Recall the 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, $\left|\hat{Y} - Y\right| \big/ Y \le \eta = \max\left(\varepsilon_N, \varepsilon_S\right)$.

Corollary. If U_S is sampled with certainty, and $\left|\hat{Y}_N - Y_N\right| \big/ Y_N \le \varepsilon_N$ then $\left|\hat{Y} - Y\right| \big/ Y \le \varepsilon_N$
There are other results in that paper having to do with "coverage"—the fraction that Y_S is (expected to be) of *Y*—which are germane to cutoff sampling, but which we pass over here for the sake of simplicity of presentation.

Iinference requires a probability *p* that an interval does what it says it does. Iinference requires a probability *p* that an interval does what it says it does. It is useful to remind ourselves of Bonferroni's inequality:

$$P\left(A_1 A_2 \cdots A_n\right) \ge P\left(A_1\right) + P\left(A_2\right) + \cdots + P\left(A_n\right) - \left(n - 1\right),$$

which enables us to go from separate assessments of probabilities on Y_S and Y_N to a probabilistic statement regarding *Y* itself. For example, suppose based on sampling (or well founded modeling) properties we are 95% certain that $\left|\hat{Y}_S - Y_S\right| \big/ Y_S \le \varepsilon_S$ and, based on historical or other considerations, 90% certain that $\left|\hat{Y}_N - Y_N\right| \big/ Y_N \le \varepsilon_N$. Then, for

$\hat{Y} = \hat{Y}_N + \hat{Y}_S$, we can be 85% certain that $\left|\hat{Y} - Y\right| \big/ Y \le \eta = \max\left(\varepsilon_S, \varepsilon_N\right)$. This sort of reasoning enables us to construct interval estimates for *Y*.

We should recognize that we are typically speaking of two probabilities p_S and p_N which are different in nature. The probability p_S will derive from standard properties of well constructed confidence intervals, while p_N will derive from experience of historical data and may require some exercise of judgment.

Result 2. Suppose that we have a $p_S = (1-\alpha)$ confidence interval $\hat{Y}_S \pm z\hat{\sigma}$ for Y_S and that
$\left|\hat{Y}_N - Y_N\right| \big/ Y_N \le \varepsilon_N$ with probability p_N.
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 $p_S + p_N$ -1and we have a $p_S + p_N$ -1 confidence interval

$$\hat{Y} - \left(\frac{\varepsilon_N}{1-\varepsilon_N}\hat{Y}_N + z\hat{\sigma}\right) \le Y \le \hat{Y} + \left(\frac{\varepsilon_N}{1-\varepsilon_N}\hat{Y}_N + z\hat{\sigma}\right).$$

Proof.
Note that $\left|\hat{Y}_N - Y_N\right| \big/ Y_N \le \varepsilon_N$ implies

$$Y_N \le \hat{Y}_N + \varepsilon_N Y_N \le \hat{Y}_N + \varepsilon_N\left(\hat{Y}_N + \varepsilon_N Y_N\right) = \hat{Y}_N\left(1+\varepsilon_N\right) + \varepsilon_N^2 Y_N \le$$

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

Corollary. If U_S is sampled with certainty, then with probability p_N,

$$\left|\hat{Y} - Y\right| = \left|\hat{Y}_N - Y_N\right| \le \varepsilon_N Y_N \le \frac{\varepsilon_N}{1-\varepsilon_N}\hat{Y}_N$$

And corresponding p_N confidence interval

$$\hat{Y} - \frac{\varepsilon_N}{1-\varepsilon_N}\hat{Y}_N \le Y \le \hat{Y} + \frac{\varepsilon_N}{1-\varepsilon_N}\hat{Y}_N$$

The challenge, of course, is getting ε_N and *p**.

3. Simulation Study

See Bridgman et al. (2011) for examples of populations which have been subject in practice to cutoff sampling within U.S. government agencies. Unfortunately, because of confidentiality restrictions, these are not readily available for study. Instead, we shall here make use of artificial populations, which, however, are based firmly on a natural population that possesses some degree of the skewness that cutoff sampling typically requires.

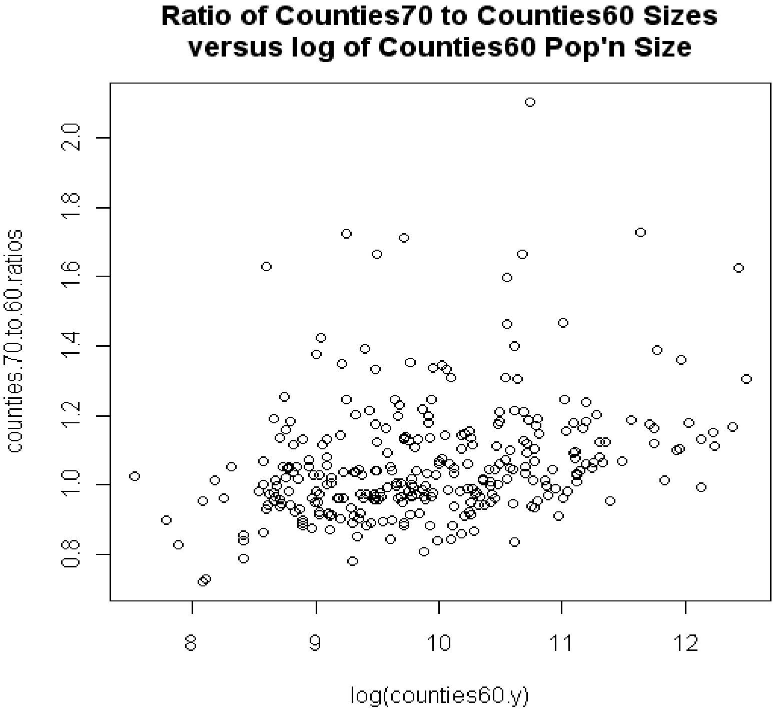
We generate 3 ***sets*** of populations based on the y variables in Counties60 and Counties70 data (Royall and Cumberland 1981). The variable of interest *y* was the number of people living in each of N = 304 southern U.S. counties in 1960 and 1970 respectively.

Each *set* consists of 200 populations, their data running from 1960, 1961, etc. thru 1970. We shall assume that censuses are available in 60 and 65 and the goal will be to get an estimate of the number of people in all the counties in 1970 – call this *Y70* – based on a sample of n = 50 in 1970 and possibly the earlier census data. We shall consider cutoff sampling and *pps* sampling.

Note. This is an unusual context for cutoff sampling. Cutoff samples of n = 50 will have about 55% "coverage" -- ratio of sum *y*'s in the sampled portion to sum of population *y*'s. A rough rule of thumb is that cutoff sampling becomes viable when such coverage reaches 80% (e.g. Knaub 2007, p. 3). Nonetheless we will see what we can learn in this setting.

3.1 Population generation

Background. Over the 10 years, the actual county populations on an annualized basis changed by factors $R_i = y_{70i} / y_{60i}$, the ratio of county *i*'s population in 1970 to its population in 1960. The following figure is a graph of these factors. It will be noted that, as a rough rule, the larger counties increased at a steeper pace than the smaller ones.



For the ***first*** set of 200 populations, for each population we generate succeeding years of data starting with Counties60 data by $y_{ti} = y_{t-1,i} r_i \left(1+\varepsilon_{ti}\right)$, where $r_i = R_i^{1/10}$ and ε_{ti} is a small mean zero normal error (standard deviation = 1/25). Call this the set of ***plain*** populations. Each population contains 11 years of data. Counties60 is kept as base year throughout for all populations. The year 1970 data should bear some resemblance to Counties70 data, but will not be exactly the same.

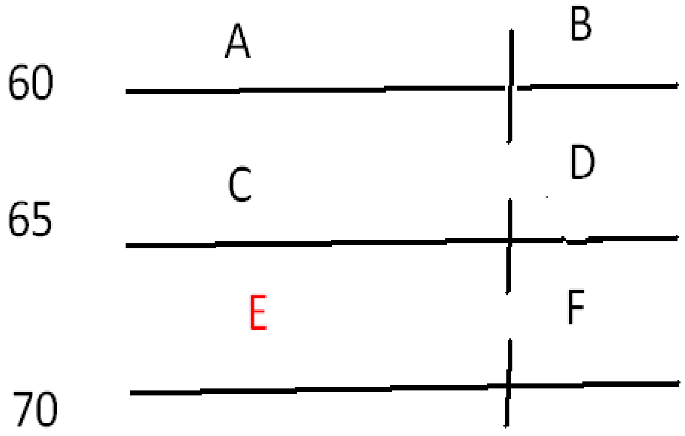
For the ***second*** set of 200 populations, we do the same except that *after* the census year 65, we randomize all the r_i so that units no longer necessarily retain their original growth rate. Those that were originally growing might now be shrinking, etc. We refer to this set as the ***jogged*** populations.

For the ***third*** set of populations there is again a shift after the census year, but now the lower half of the population is growing at a rate 1.05 times that of the upper half. More specifically, for i = 1,…, 152, we set r_i =1.05 $r_{i\ +152}$. This is the set of ***ramped*** up populations.

For each of the 200 populations in each set we shall have use of census data from years 60 and 65. Our target year will be 1970. From this we take a sample of n = 50, either by cutoff sampling or by probability proportional to size sampling (*pps*), using the 65 census data as the size variable.

3.2 S ampling S trategies. We use three sampling strategies: (a) cutoff sampling combined with ratio estimation, (b) cutoff sampling combined with updating of known census values, (c) *pps* sampling combined with mean of ratios estimation. In each of these cases only ***one*** sample will be taken from each population.

The following figure represents the information available with cutoff sampling in '70 and censuses in 60 and 65. B, D, F correspond to the units in the population collected with certainty in 1970; A, C, E represent the units excluded from the sample in 1970. But A, C are later subject to a census. Therefore in the figure only E is unknown.



3.2.1 Estimation and Inference, under *Cutoff sampling*

For cutoff sampling, we use two distinct estimation procedures, namely standard ratio estimation and also an updating procedure.

3.2.1.1 Ratio Estimation

We take $\hat{Y}_{70} = \sum_S y_{70,i} + \hat{\beta}\sum_N y_{65,i}$, where $\hat{\beta} = \dfrac{\sum_S y_{70,i}}{\sum_S y_{65,i}}$ (so $\hat{Y}_{70,N} = \hat{\beta}\sum_N y_{65,i}$)

For inference we in turn use two procedures:

(a) Use jackknife variance estimator v_J and calculate a nominal 95% CI by $\hat{Y}_{70} \pm 2\sqrt{v_J}$

This may be regarded as the "standard procedure".

(b) For the second approach, we assume some data from 1967 is available, enabling us to estimate

$\left|\hat{Y}_{67N} - Y_{67N}\right| \big/ Y_{67N} \le \varepsilon_{67N}$. (In the simulation study we used info from just first five of the 200 runs for this, taking the maximum value this ratio took in the five runs—a crude estimate, "based on historical data".) We set $\varepsilon_{70N} = 5\varepsilon_{67N} / 2$ (this is not a self-evident calculation, but can be given some justification) and use *Result 2* to get an interval estimate $\hat{Y}_{70} \pm \hat{Y}_{70N}\, \varepsilon_{70N} \big/ \left(1 - \varepsilon_{70N}\right)$.

3.2.1.2 Updating procedure

Let $R = \dfrac{\sum_N y_{65,i}}{\sum_N y_{60,i}}$ and $\tilde{Y}_{70} = \sum_S y_{70,i} + R\sum_N y_{65,i}$. Note that no use is made of the newly sampled data (corresponding to F in the above Figure) to estimate $\hat{Y}_{70,N} = R\sum_N y_{65,i}$.

For inference we again use two approaches. (a) Let lower bound of interval be

$\tilde{Y}_{70,L} = \sum_S y_{70,i} + \sum_N y_{65,i}$

Let upper bound of interval be $\tilde{Y}_{70,U} = \sum_S y_{70,i} + R_U \sum_N y_{65,i}$ where R_U is an appropriately selected quantile of $\left\{\dfrac{y_{70,i}}{y_{65,i}}\right\}_{i \in S}$ or $\left\{\dfrac{y_{65,i}}{y_{60,i}}\right\}_{i \in U}$ whichever is greater. Somewhat arbitrarily, we used the 90th quantile.

(b) as above, using Result 2.

3.2.2 Estimation and Inference, *pps* (size variable = counties65)

Here things are straightforward. For point estimation, suitable to the selection probabilities, we used the *mean of ratios estimator*

$$\breve{Y}_{70} = \sum_S y_{70,i} + \hat{\beta}\sum_N y_{65,i}\text{, with } \hat{\beta} = \frac{\sum_S y_{70,i} / y_{65,i}}{n}$$

For inference, we used a jackknife variance estimator v_J and calculated a 95% CI by $\breve{Y}_{70} \pm 2\sqrt{v_J}$

Notes: (i) the *pps* samples had a minimum of 6 certainty units; contrary to usual design

based protocol, these were included in the calculation of $\hat{\beta}$. This was merely a matter of convenience. (ii) About half of each *pps* sample fell below the cutoff point used in the cutoff sampling; thus there was a serious difference between the cutoff and the *pps* samples, the latter filling in the lower half of the population to a large extent.

3.3 Assessment Measures for each set over the *K* = 200 populations

It should be noted that the target differed for each run *k* = 1,…, 200.
We used the following measures to compare the several methods of estimation and inference that were tried:

mean relative error = $100 mean\left(\frac{\hat{Y}_k - Y_k}{Y_k}\right)_{k=1,...,200}$

mean relative absolute error = $100 mean\left(\frac{|\hat{Y}_k - Y_k|}{Y_k}\right)_{k=1,...,200}$

Coverage = $\sum_{k=1}^{200} I\left(Y_k \in I_k\right) / 200$

mean relative interval length = $100 mean(length(I_k)/Y_k)$

3.4 Results

Population Set 1 (Plain Population)
Rate of increase on units same over time, although subject to random error

	relative bias	relative abs error	% coverage	relative int. length
cutoff/ ratio	1.79	1.80	92.5 (v_J) 100.0 (ε)	5.82 8.84
cutoff/ updating	-0.42	0.50	100.0 (*90q*) 88.0 (ε)	9.28 1.97
pps/mean of ratios	0.86	1.16	92.0 (v_J)	4.86

Population Set 2 (Jogged Population)
Rate of increase on units jogged after 65 census

	relative bias	relative abs error	% coverage	relative int. length
cutoff/ ratio	0.00	0.76	99.0 (v_J) 95.5 (ε)	5.35 3.89
cutoff/ updating	0.17	0.39	100.0 (*90q*) 97.0 (ε)	8.98 2.21
pps/mean of ratios	-0.03	1.01	95.5 (v_J)	4.67

Population Set 3 (Ramped Up Population)
Rate of increase on lower units ramped up after 65 census

	relative bias	relative abs error	% coverage	relative int. length
cutoff/ ratio	-2.63	2.63	56.0 (v_J) 80.0 (ε)	5.56 6.59
cutoff/ updating	-4.75	4.75	100.0 (*90q*) 83.0 (ε)	8.88 10.45
pps/mean of ratios	-0.60	1.17	94.5 (v_J)	5.56

3.5 Observations

Here are some observations on these results.

1. *pps* sampling works well: relative absolute error about 1%, coverage about nominal, relative confidence interval length (*rcil*) about 5%, consistently across the two tame and one wild population sets
2. *cutoff/ratio*: its relative absolute error was intermediate between updating and what *pps* gave, for all three population sets. Very mixed bag with respect to inference: coverage as low as 56% using jackknife based intervals (basically because of bias of estimates). Mixed results with respect to the two methods of inference, but method (b) using the '67 data, better on *ramp* population
3. *cutoff/updating*: best with respect to relative absolute error for tame populations, worst for *ramp*. Method (a) inference based on 90th quantiles of ratios of known units, very conservative, giving unduly large 100% intervals. Raises the question whether we can find a more appropriate choice of quantile. Is there a rationale for choosing appropriate quantile? Method (b), using ε method with data supposed known from 67, worked quite well on the *tame* and *jogged* populations, but, mysteriously, had lesser coverage with longer intervals for *ramp* population set.

4. Discussion

We have carried out one simulation study on three sets of populations, none of which would usually be considered ideal for the use of cutoff sampling. We have explored two different ways of capitalizing on fuller data from the past for point estimation, and for each of these two ways of constructing interval estimates. The newer methods are suggestive, but not conclusive, and, as the results on the *ramped* up population shows, there is a vulnerability to things going very badly, if the population behaves in very unexpected ways. The potential increase in efficiency when the population is well behaved, and the promising behavior under those circumstances of the interval estimates, suggests that cutoff sampling should continue to be considered, and that the methods we have suggested for interval estimates deserve further exploration.

References

Bridgman, B.R., Cheng, Y., Dorfman, A.H., Lent, J., Liu, Y.K., Miranda, J., Rumberg, S., Yorgason, D. R. (2011). "Cutoff Sampling in Federal Establishment Surveys: An Inter-Agency Review", Proceedings of Section on Government Statistics, American Statistical Association, July 30 – August 4, 2011, Miami Beach, pp. 76 - 90

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

Knaub, J. (2007) "Cutoff Sampling and Inference", *InterStat*, April, http://interstat.statjournals.net/YEAR/2007/abstracts/0704006

Royall, R.M. and Cumberland, W.G. (1981) "An Empirical Study of the Ratio Estimator and Estimators of its Variance", *Journal of the American Statistical Association*, **73**, 66-77

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