Assessing the Relationship between Balanced Sample and Sample Representativity†

Yonil Park[1], John Chesnut[1]
[1]Demographic Statistical Methods Division, U.S. Census Bureau, Washington DC 20233

Abstract
The Representativeness Index (RI) is a measure of how much the observed sample represents its population. It is a min-max normalized index and a distribution-free measure. There has been little empirical validation that a balanced sample is associated with a representative sample. Here, we investigate the relationship between the RI and the balanced sample by simulating sampling from uniform distribution and normal distribution. We quantify the quality of the balanced sample based on its mathematical definition, comparing the sample mean with the population mean for a variable of interest. Then, we propose a length-biased correction to a simple random sampling distribution to improve the RI. The length-biased distribution, the limiting distribution of spread in a renewal process, weights the density function at each x by its length x. Lastly, we demonstrate how much the proposed length-biased correction to a simple random sample improves the RI using the 2010 Sample Redesign Primary Sampling Units for the American Housing Survey, Current Population Survey, and Survey of Income and Program Participation.

Key Words: Representativeness Index, Balanced Sample, Length-biased Distribution, Sample Redesign Primary Sampling Units

1. Introduction

Sample representativity is important for statistical inference in terms of precision and statistical power (Martínez-Mesa, González-Chica, Duquia, Bonamigo, & Bastos, 2016). Sampling bias or non-response bias can cause the lack of sample representativity. Bertino (2006) proposed a univariate measure of sample representativity for inferential purposes. This representativeness index is a distribution-free measure based on the Smirnov – Cramér – Von Mises statistic. It has been utilized to select the best parametric distribution for inequality measures in economic studies (Kpanzou, Tertius, & Lo, 2017).

Hájek (1981) connected sample representativity with the estimation of population parameters by pairing a sampling design and an estimator. If the population parameter can be estimated from a sample without bias and with a null variance, this sample is representative. This definition is similar to the definition of a balanced sample, where the estimated total from a sample is equal to the population total (Yates, 1953). Although we can conjecture that sample representativity is associated with a balanced sample, there is little empirical validation. Both sample representativity and balanced sample has been interchangeably used with ambiguity.

The main goal of this study is to assess the relationship between sample representativity and balanced sampling by Monte Carlo simulation. We utilize Bertino's measure to

† Disclaimer: This article is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau. Disclosure Avoidance Officer Approval number: CBDRB-FY19-245

evaluate sample representativity against the balanced sample equation to study their relationship. We also use these measures to evaluate the 2010 Sample Redesign Primary Sampling Units (PSUs) to see whether the US Census Bureau Demographic Surveys' PSU samples are representative and balanced for selected demographic characteristics over time.

The organization of this article is as follows. Section 2 describes our methodology to evaluate the relationship between sample representativity and balanced sampling. Section 2.1 and Section 2.2 present Bertino's measure for sample representativity and the definition of a balanced sample, respectively. Section 2.3 and Section 2.4 review the 2010 Sample Redesign demographic surveys and their sampling methods. In Section 2.5, we propose a length-bias correction distribution to the simple random sampling (SRS) method to improve the sample representativity. Section 3 provides our simulation results. Section 3.1 shows Monte Carlo simulation results for the relationship between balanced sample and sample representativity under the standard uniform and the standard normal distributions. Section 3.2 compares the cumulative probability distribution from probability proportional to size (PPS) to the cumulative probability distribution from the length-bias correction distribution to SRS. Section 3.3 computes Bertino's measures for the 2010 Sample Redesign PSUs samples. Finally, Section 4 reflects on our results.

2. Methodology

2.1 Representativeness Index

We measure sample representativity with the Representativeness Index (RI) proposed by Bertino (2006). The RI is defined by

$$R(\boldsymbol{x}, F) = 1 - \frac{12n}{4n^2 - 1} \sum_{r=1}^{n} \left(F\left(x_{(r)}\right) - \frac{2r-1}{2n} \right)^2,$$

where $\boldsymbol{x} = (x_1,\ x_2, \dots, x_n)$ denotes a random sample of size n of random variable X with a cumulative distribution function $F(\cdot)$ and r a rank for a given sample value x such that $x_{(1)} < x_{(2)} \dots < x_{(n)}$. This index is based on the Smirnov – Cramér – Von Mises test statistic, a goodness of fit test statistic comparing a sample distribution and a population distribution. With this index, we can interpret how much of the observed sample represents its population distribution. It is also min-max normalized such that $0 \leq R(\boldsymbol{x}, F) \leq 1$. If the sample distribution is close to the population distribution and exhibits the population's characteristics fully, its representativeness index is close to one. If not, it is close to zero.

2.2 Balanced Sampling

A balanced sample is mathematically defined (Valliant, Dorfman, & Royall, 2000; Yates, 1953). Consider a sample s that is a subset of a finite population U. In our study, we consider the totals for the key variables of interest for each survey and our PSUs have inclusion probabilities π_i of being selected. Thus, we used π_i-balanced sample definition as the following:

$$\sum_{i \in s} \frac{x_i}{\pi_i} = \sum_{i \in U} x_i \,.$$

The sample s satisfying the above equation is called a π_i-balanced sample. We call the above equation a balancing equation. The left side is an estimate of the total for the characteristic of interest derived from a sample and the right side is the population total. The balancing equation can be generalized to the *k*-th order by the following:

$$\sum_{i \in s} \frac{x_i^k}{\pi_i} = \sum_{i \in U} x_i^k .$$

Using the above balancing equation, we compare the total sample estimates for the key survey variables with the corresponding population totals. Then we compute the relative errors by dividing the difference of the two totals by the population total.

2.3 The 2010 Sample Redesign

The 2010 Sample Redesign is the U.S. Census Bureau program that selected and disseminated updated samples for a number of demographic surveys based on the 2010 Census, American Community Survey (ACS) and administrative records (Nguyen & Gerstein, 2011). We are interested in evaluating whether a sample of Primary Sampling Units (PSUs) from the 2010 Sample Redesign are representative samples compared to the current demographic characteristics and projecting forward a decade. In our study, we consider the American Housing Survey (AHS), the Current Population Survey (CPS) and the Survey of Income and Program Participation (SIPP). We select the following key variables for each survey: the number of tenured housing units for AHS; total unemployment level for CPS; and total poverty level for SIPP. Since we do not have population totals for the key variables, we used the ACS 5-year estimates as proxies of the population totals for all surveys' key variables.

2.4 Probability Proportional to Size Sampling (PPS)

The 2010 Sample Redesign uses a two-stage sample design requiring selection of a sample of PSUs using the PPS method in the first stage. Let m_{hi} denote the measure of size for PSU i in the cluster h and M_h the measure of size of the cluster h. Then the selection probability of PSU *i* that belongs to the cluster h, denoted by π_{hi}, is computed by m_{hi}/M_h. Since $\sum_h \sum_i \pi_{hi} = n$, we need to normalize the probability mass function $f(x_{hi}) = \pi_{hi}/n$ for the RI computation under the PPS, where $F(x) = \sum_{(i,h) \in s,\, x_{hi} \le x} f(x_{hi})$.

2.5 Length-bias Correction to Simple Random Sampling (SRS)

Suppose that we randomly select one student from a certain school. Then the family size of the randomly selected student tends to be stochastically larger than that of a regular family. This tendency is known as the *inspection paradox* or *length-bias correction* or *limiting distribution of spread* in a renewal process (Ross, 2003). The intuitive idea is that a random selection is more likely to occur to a student with a large family size than a small one. This is reflected by biasing the probability mass function at each x by its size x. We can improve the RI by applying the length-bias correction to SRS. Under the SRS, probability mass function $f(x)$ becomes $1/n$. The length-bias correction to SRS updates the $f(x)$ with $xf(x)/E(X)$, where $E(X)$ is the expected value of a random variable X.

3. Results

3.1 Relationship between Representativeness Index and Balanced Sample

Figure 1 displays the relationship between sample representativity and balanced sampling with 10,000 simulation runs under the Uniform (0, 1) and standard normal distributions. It shows that there is an explicit relationship between the representativeness index and sampling error. If a sample distribution resembles the population distribution, its representativeness index is very close to 1 and yields a very small bias in estimation. If a sample is not representing a population fully, then its representativeness index moves away from 1 and yields a large bias in estimation. As the sample size increases, the RI improves, the bias decreases, and their relationship is clearer.

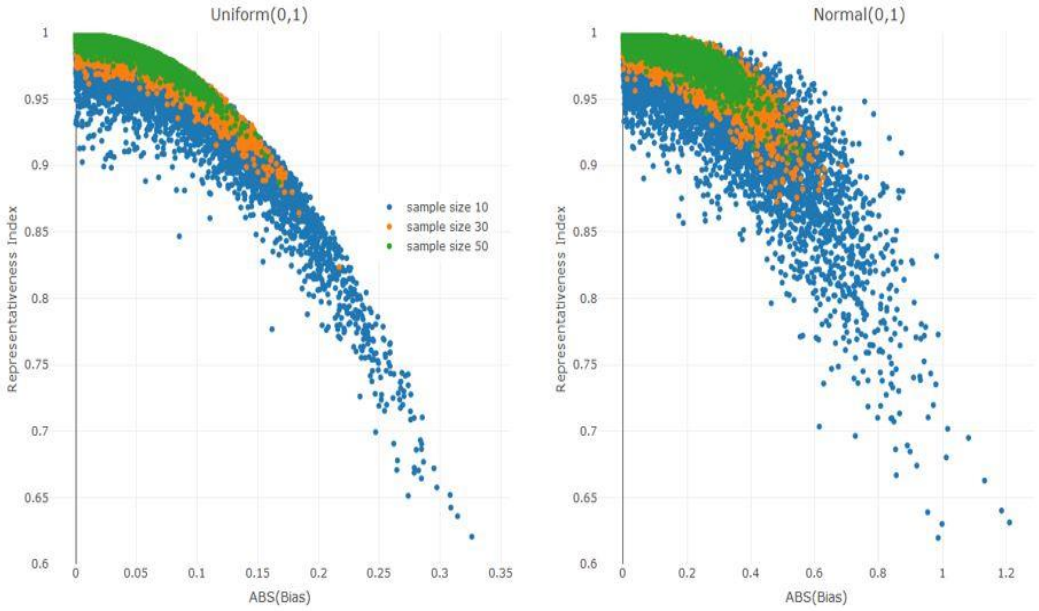

Figure 1: Plots of representativeness index (RI) against absolute value of bias for Uniform (0, 1) and standard normal distributions. Each plot shows 10,000 simulation runs for the sample size of 10, 30 and 50, respectively. X-axis shows the absolute value of difference between population mean and sample mean, and Y-axis shows the representativeness index for each selected samples. Green, orange, and blue points are for sample size 10, 30, and 50, respectively.

3.2 Comparison of Length-bias Correction to SRS and PPS

In two stage PPS sampling, the larger clusters have a higher chance of selection in the first stage. It is conceptually similar to the length-bias correction to SRS as the probability of randomly being selected sample depends on the value of a variable.

Figure 2 shows the cumulative probability distributions of three different sampling methods - PPS, SRS and length-bias correction to SRS – for the number of tenured housing units for the AHS survey. It confirms that PPS and length-bias correction to SRS methods yield similar probability distributions while SRS method overestimates. PPS sampling requires us to know the measure of sizes for each cluster in advance so that we can compute the corresponding sampling weights. The length-bias correction to SRS can be a quick but useful heuristic correction when the measure of sizes for each cluster is not available.

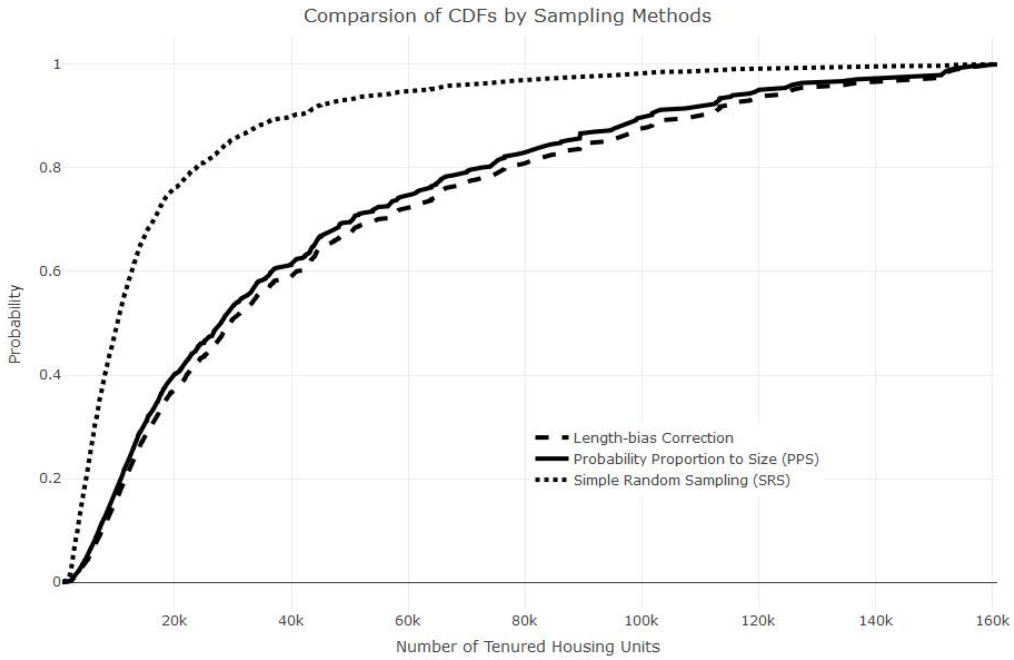

Figure 2: Comparison of cumulative distribution function from different sampling method. It displays the distribution of the number of tenured housing units from 2012-2016 ACS 5-year estimates. The solid, dotted, and dashed lines show the cumulative probability distributions from PPS, SRS, and length-bias correction to SRS, respectively.

3.3 Evaluation for the 2010 Sample Redesign PSUs

Figure 3 displays the national-level representativeness index results with three different sampling methods for the AHS, CPS and SIPP surveys from 2011 to 2016. With the PPS sampling method, all survey's representativeness indexes were close to 1. Based on the simulation study results in Section 3.1, we can interpret that the 2010 Sample Redesign PSUs are balanced samples even though there has been some demographic changes in the population since 2010. We also performed the sub-national level analysis and the overall median representativeness indexes were near 0.95 for all surveys (data not shown in the article). As we showed in Section 3.2, the length-bias correction to SRS improved the RI value significantly in Figure 3.

Finally, we computed relative errors by dividing the absolute differences of the survey sample estimates from population estimates by the population estimates to evaluate our sample PSUs. Note that the population estimates are the ACS 5-year estimates at the national level and the sample estimates are derived from the sampled PSUs. For the national-level analysis, relative errors were within 3.5 percent for the three surveys. The overall median for relative errors at the sub-national-level was 1.7 percent (data not shown in this article).

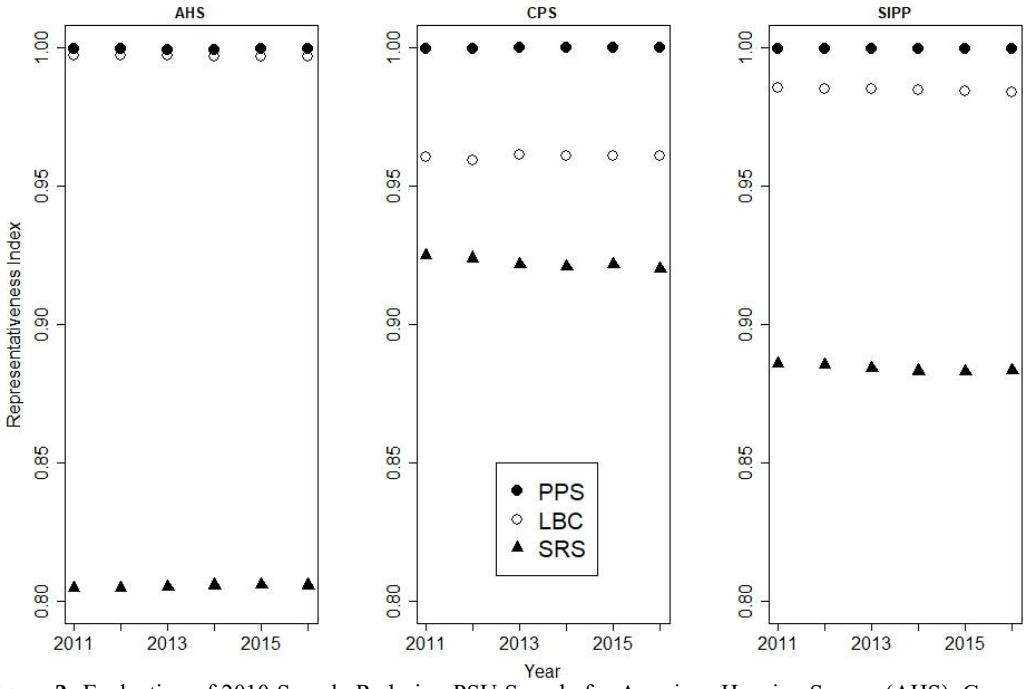

Figure 3: Evaluation of 2010 Sample Redesign PSU Sample for American Housing Survey (AHS), Current Population Survey (CPS), and Survey of Income and Program Participation (SIPP). We consider the number of tenured housing units for AHS; unemployment level for CPS; poverty level for SIPP from the ACS 5-year estimates. Each plot displays the RI over time from 2011 to 2016 for the three different sampling methods. ● shows the RI for 2010 Sample Redesign Sampling method - PPS; ○ length-bias correction (LBC) to SRS; ▲ SRS.

4. Concluding Remarks

We presented that sample representativity is highly associated with balanced sample from the simulation study. We observed the similar relationship for various distributions including exponential and chi-square distributions (data not shown in the article). Accordingly, if a sample gives a high representativeness index value, we can interpret that this sample is a highly balanced sample. The 2010 Sample Redesign PSUs demonstrated a very high representativeness index value for the AHS, CPS and SIPP surveys even in the presence of changing demographic characteristics in the population over time. We also proposed the length-bias correction to SRS method to improve the utility of the representativeness index for SRS. Note that Bertino's (2006) representativeness index played an important role in our study. However, this index is a univariate measure. As a further study, we can explore the extension of this index to a multivariate measure and investigate if the relationship between the sample representativity and balanced sample is valid for multivariate measure.

Acknowledgements

This research was supported by the Sample Redesign Research Program of Demographic Statistical Method Division of the U.S. Census Bureau. We thank Darcy Steeg Morris of Center for Statistical Research and Methodology of the U.S. Census Bureau for her statistical review and helpful comments that led to an improved paper.

References

Bertino, S. (2006). A Measure of Representativeness of a Sample for Inferential Purposes. *International Statistical Review, 74*(2), 149-159.

Hájek, J. (1981). *Sampling from a Finite Population* New York: Marcel Dekker, USA.

Kpanzou, T. A., Tertius, D., & Lo, G. S. (2017). Measuring inequality: application of semi-parametric methods to real life data. *African Journal of Applied Statistics, 4*(1), 157-164.

Martínez-Mesa, J., González-Chica, D. A., Duquia, R. P., Bonamigo, R. R., & Bastos, J. L. (2016). Sampling: how to select participants in my research study? *Anais brasileiros de dermatologia, 91*(3), 326-330. doi:10.1590/abd1806-4841.20165254

Nguyen, T., & Gerstein, A. (2011). *Sample Design Research in the 2010 Sample Redesign.* Paper presented at the Proceedings of the American Statistical Association, Section on Survey Methods Research.

Ross, S. M. (2003). The inspection paradox. *Probability in the Engineering and Informational Sciences, 17*(1), 47-51.

Valliant, R., Dorfman, A. H., & Royall, R. M. (2000). *Finite population sampling and inference: a prediction approach*: John Wiley.

Yates, F. (1953). *Sampling methods for censuses and surveys*.