Correction and Application of PPS Sampling Estimators Under Ranking

Authors: Wang Feng, Wang Feng

Date: 2018-09-26T00:00:00+00:00

Abstract

Inspired by the Zipf phenomenon exhibited by many entities, this paper proposes a sorted PPS sampling method and derives the modified Hansen-Hurwitz estimator along with its variance. In doing so, it resolves the long-standing issue in sampling survey practice regarding the lack of a clear methodology for determining how many important units should be directly included in the sample, providing both theoretical justification and a concrete determination method. Finally, through an illustrative example and a case study on China' s urban population sampling survey, the paper demonstrates the advantages of the modified Hansen-Hurwitz estimator and concludes with a summary and future research outlook.

Full Text

Modification and Application of PPS Sampling Estimator Under Rank Ordering

Wang Feng
(School of Statistics, Shanxi University of Finance and Economics, Taiyuan, Shanxi, 030006)

Abstract

Inspired by the Zipf phenomenon observed in many real-world contexts, this paper proposes a post-stratification PPS sampling method and derives a modified Hansen-Hurwitz estimator along with its variance. This approach addresses the long-standing practical issue in sampling surveys where important units are included directly in the sample without a clear methodology for determining how many such units should be included. We provide both the theoretical foundation and a specific method for determining this number. Finally, through an

illustrative example and a case study of China' s urban population sampling survey, we demonstrate the superiority of the modified Hansen-Hurwitz estimator and conclude with a summary and outlook for future research.

Keywords: sampling survey; PPS sampling; Zipf phenomenon
Classification Code: C811
Document Code: A

1 Introduction

In sampling practice, sampling units typically differ in size and thus in their significance within the population. Survey designers exploit these differences by assigning higher selection probabilities to important units and lower probabilities to less important ones, where importance is usually measured by unit size. This principle leads to probability proportional to size (PPS) sampling with replacement, where each unit' s selection probability is proportional to its size. PPS sampling has found widespread application in survey practice and has attracted continuous scholarly attention, resulting in numerous extensions and refinements.

Since Hansen and Hurwitz (1943) first proposed the theoretical foundations of PPS sampling [**?**], subsequent research has expanded the methodology considerably. Yates and Grundy (1953) examined PPS sampling within strata [**?**], Holmberg (1998) applied bootstrap methods to PPS sampling, Kim et al. (2013) studied variance estimation under two-stage systematic PPS sampling [**?**], and Patel and Bhatt (2016) developed model-based variance estimation for PPS sampling [**?**]. In the Chinese literature, Zou and Feng (1995) investigated the admissibility of PPS sampling with replacement within the class of sampling with replacement designs [**?**], Sun and Jiang (2002) studied rotation sampling for PPS samples [**?**], and Liu and Chen (2005) discussed extensions of the Hansen-Hurwitz estimator under MPPS sampling across different scenarios [**?**]. Additional contributions include [**?**].

A common thread in this literature is that unit size, as a measure of importance, serves as a crucial auxiliary variable in PPS sampling design. This raises a natural question: could we improve estimation by first ordering units by size and then designing the PPS sample accordingly?

The motivation for ordering stems from Zipf' s law. In 1935, Zipf analyzed the relative frequency of words in English and discovered that a few words appear with very high frequency. He established that the product of a word' s frequency and its rank remains approximately constant—an empirical regularity known as Zipf' s law [**?**]. This phenomenon reveals that a small proportion of units often account for a large share of frequency or total quantity, a property termed the Zipf phenomenon. Subsequent research has documented this phenomenon across diverse domains, including city populations, personal incomes, and agricultural production [**?**, **?**].

Inspired by this observation, we propose dividing the ordered population into two components: a set of "important" units (in the Zipf sense) that are included in the sample with certainty, and the remaining units sampled via conventional PPS sampling. This approach promises to improve estimation precision. While survey practitioners occasionally include important units directly in samples, the number of such units has typically been determined subjectively without a systematic methodology. This paper provides a theoretically grounded method for determining the number of important units to include with certainty and derives a modified Hansen-Hurwitz estimator for this setting.

The remainder of the paper proceeds as follows. Section 2 develops the modified PPS sampling estimator and its variance after partitioning the population. Section 3 derives conditions under which the modified estimator outperforms the traditional version. Section 4 presents our method for determining the number of important units. Section 5 demonstrates the advantages of our approach through an illustrative example and a case study of China' s urban population sampling survey. Section 6 concludes with a summary and directions for future research.

2 Modified PPS Sampling Estimator Under Ordering

Consider a finite population sorted by size variable, denoted as $Y_1, Y_2, \ldots, Y_N$, with corresponding size measures $M_1, M_2, \ldots, M_N$. Let $M_0 = \sum_{i=1}^{N} M_i$. Under conventional PPS sampling, the unbiased estimator for the population total is:

$$\hat{Y}_{HH} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{p_i}$$

where $p_i = M_i/M_0$. This is the well-known Hansen-Hurwitz estimator, shown as Equation (1).

The variance of this estimator is:

$$V(\hat{Y}_{HH}) = \frac{1}{n} \sum_{i=1}^{N} p_i \left(\frac{Y_i}{p_i} - Y \right)^2$$

which we denote as Equation (2). When $n > 1$, an unbiased estimator of this variance is given by:

$$\hat{V}(\hat{Y}_{HH}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{y_i}{p_i} - \hat{Y}_{HH} \right)^2$$

This is Equation (3).

Drawing on the Zipf phenomenon perspective, we partition the population into two groups. Group G_1 contains N_1 units that are included in the sample with certainty, so $n_1 = N_1$. Group G_2 contains the remaining N_2 units, from which we select n_2 units via conventional PPS sampling, where $n_2 = n - n_1$. For clarity, define:

$$Y_1 = \sum_{i=1}^{N_1} Y_i, \quad Y_2 = \sum_{i=N_1+1}^{N} Y_i, \quad M_1 = \sum_{i=1}^{N_1} M_i, \quad M_2 = \sum_{i=N_1+1}^{N} M_i$$

Treating G_2 as a separate population, Hansen and Hurwitz (1943) yield the unbiased estimator for G_2' s total as:

$$\hat{Y}_{2HH} = \frac{1}{n_2} \sum_{i=1}^{n_2} \frac{y_i}{p_{2i}}$$

where $p_{2i} = M_i/M_2$ for units in G_2.

Since G_1 is completely enumerated, it involves no random sampling. Combining both components, the unbiased estimator for the overall population total becomes:

$$\hat{Y}_{mHH} = Y_1 + \hat{Y}_{2HH}$$

We refer to this as the modified Hansen-Hurwitz estimator, shown as Equation (4). Note that Y_1 is not a random variable but a constant, while $\hat{Y}_{2HH}$ is a Hansen-Hurwitz estimator. Therefore, by the unbiasedness property of the Hansen-Hurwitz estimator, $\hat{Y}_{2HH}$ is an unbiased estimator of G_2's total, making $\hat{Y}_{mHH}$ an unbiased estimator of the population total.

Similarly, since the first component of the modified Hansen-Hurwitz estimator is non-random and the second component is a Hansen-Hurwitz estimator from PPS sampling, the variance of the modified estimator equals the variance of the second component alone:

$$V(\hat{Y}_{mHH}) = V(\hat{Y}_{2HH}) = \frac{1}{n_2} \sum_{i=N_1+1}^{N} p_{2i} \left(\frac{Y_i}{p_{2i}} - Y_2 \right)^2$$

This is Equation (5). Likewise, when $n_2 > 1$, Hansen and Hurwitz (1943) provide the unbiased variance estimator:

$$\hat{V}(\hat{Y}_{mHH}) = \frac{1}{n_2(n_2-1)} \sum_{i=1}^{n_2} \left(\frac{y_i}{p_{2i}} - \hat{Y}_{2HH} \right)^2$$

which we denote as Equation (6).

3 Variance Comparison Between PPS and Modified PPS Estimators

The variance of the Hansen-Hurwitz estimator under conventional PPS sampling, given in Equation (2), can be rewritten as:

$$V(\hat{Y}_{HH}) = \frac{1}{n}\left[\sum_{i=1}^{N_1} p_i \left(\frac{Y_i}{p_i} - Y\right)^2 + \sum_{i=N_1+1}^{N} p_i \left(\frac{Y_i}{p_i} - Y\right)^2\right]$$

For the modified Hansen-Hurwitz estimator, the variance from Equation (5) becomes:

$$V(\hat{Y}_{mHH}) = \frac{1}{n_2}\sum_{i=N_1+1}^{N} p_i \left(\frac{Y_i}{p_i} - Y_2\right)^2$$

The difference between these variances is:

$$V(\hat{Y}_{HH}) - V(\hat{Y}_{mHH}) = \frac{1}{n}\sum_{i=1}^{N_1} p_i \left(\frac{Y_i}{p_i} - Y\right)^2 + \left(\frac{1}{n} - \frac{1}{n_2}\right)\sum_{i=N_1+1}^{N} p_i \left(\frac{Y_i}{p_i} - Y\right)^2 - \frac{1}{n_2}(Y - Y_2)^2$$

This expression, labeled as Equation (7), comprises three components. The first and third terms are non-negative. If the second term is also positive—that is, if $\frac{p_2}{n_2} > \frac{1-p_1}{n-n_1}$—then $V(\hat{Y}_{HH}) > V(\hat{Y}_{mHH})$. Under this condition, the modified Hansen-Hurwitz estimator outperforms the traditional version.

4 Determination of the Number of "Important" Units ($n_1 = N_1$)

To identify the number of "important" units, we must find n_1 satisfying $\frac{p_2}{n_2} > \frac{1-p_1}{n-n_1}$. Since $n_2 = n - n_1$, this inequality becomes:

$$\frac{p_2}{n-n_1} > \frac{1-p_1}{n-n_1}$$

which simplifies to $p_2 > 1 - p_1$. However, because $p_2 = 1 - p_1$ by definition, we need to reconsider the condition. The actual requirement derived from Equation (7) is that the overall variance reduction be positive. This occurs when the reduction from sampling G_2 more intensively outweighs the variance introduced by the fixed component.

From the inequality analysis, if p_1 is so small that $np_1 < 1$, then no valid n_1 exists because n_1 cannot be less than 1. In such cases, the modified estimator

is not applicable, and the traditional Hansen-Hurwitz estimator should be used. Alternatively, one could increase the sample size to satisfy $np_1 > 1$, as illustrated in the examples below.

To determine n_1, we note that $np_1 > 1$ is a necessary condition. The variance reduction condition implies that we should minimize $\frac{p_2}{n_2}$. From sampling theory, when n_2 is minimized, $\frac{p_2}{n_2}$ is maximized, which ensures $V(\hat{Y}_{HH}) > V(\hat{Y}_{mHH})$. Therefore, we search among all possible $n_1 \in \{1, 2, 3, \ldots, n-1\}$ for the value that minimizes $\frac{p_2}{n_2}$. Let this optimal value be n_1^*.

The population units are ordered by size, so $z_1 > z_2 > \cdots > z_N$ where $z_i = Y_i/M_i$, and $p_1 > p_2 > \cdots > p_N$. As n_1 increases, n_2 decreases, and the sequence $\frac{p_2}{n_2}$ typically decreases initially then increases (a formal proof appears in the Appendix). Therefore, at $n_1 = n_1^*$, $\frac{p_2}{n_2}$ reaches its minimum, satisfying the condition for variance reduction. If $\frac{p_2}{n_2}$ increases monotonically, then $n_1^* = n-1$. If even $n_1 = 1$ fails to satisfy the initial condition, the modified estimator cannot be applied, and the traditional method should be used.

5 Empirical Analysis and Testing

We demonstrate the application of the modified Hansen-Hurwitz estimator through two examples, comparing its performance with the traditional estimator.

5.1 Textbook Example

This example, adapted from Feng, Ni, and Zou (1998) [**?**], concerns employee count surveys. Table 1 presents data for $N = 36$ units in a system, showing previous year employee counts (X_i) and current year counts (Y_i). Using X_i as the size measure M_i, we select $n = 11$ units via PPS sampling to estimate the total current-year employees Y.

Step 1: Sort units by size variable X_i in descending order. Calculate np_1 to verify the applicability condition. We obtain $p_1 = 0.163$ and $n = 11$, giving $np_1 = 1.79 > 1$. Note that the original textbook example used $n = 6$, which yields $np_1 < 1$ and fails to meet the condition for the modified estimator. In such cases, one could either use the traditional Hansen-Hurwitz estimator or increase the sample size to satisfy $np_1 > 1$. We use $n = 11$ here, which does not affect the comparison with the traditional estimator.

Step 2: Determine n_1. For each possible n_1, compute $p_1 = \sum_{i=1}^{n_1} p_i$, $n_2 = n-n_1$, and $\frac{p_2}{n_2}$, then identify the value minimizing $\frac{p_2}{n_2}$. Results appear in Table 2 .

Figure 1 [Figure 1: see original paper] shows that $\frac{p_2}{n_2}$ is minimized at $n_1 = 4$. Thus, the "important" units are the top four ranked units with IDs 4, 10, 18, and 11, which are included in the sample with certainty.

Step 3: After removing these n_1 units, select $n_2 = n - n_1$ units from the remaining population using PPS sampling. Using the code method, we obtain units 17, 25, 12, 23, 36, 15, and 3 from G_2.

Applying Equation (4) yields the modified Hansen-Hurwitz estimate, and Equation (6) provides its variance estimate. Following Feng, Ni, and Zou (1998), we repeat this three-step process to generate four samples:

- Sample I: 4, 10, 18, 11; 17, 25, 12, 23, 36, 15, 3
- Sample II: 4, 10, 18, 11; 24, 14, 19, 21, 7, 36, 25
- Sample III: 4, 10, 18, 11; 5, 24, 14, 1, 17, 3, 13
- Sample IV: 4, 10, 18, 11; 13, 24, 15, 25, 19, 21, 1

Table 3 compares the estimates and standard errors. The second and third columns show the modified Hansen-Hurwitz estimates and their standard errors, while the fourth and fifth columns present the traditional Hansen-Hurwitz estimates and standard errors for the same samples. In all cases, the modified estimator' s standard error is smaller, confirming the theoretical superiority. Note that the variance estimates in Feng, Ni, and Zou (1998) differ from ours, primarily because their sample size ($n = 6$) was smaller than ours ($n = 11$). However, comparing estimators on the same samples clearly shows the modified version' s advantage.

5.2 Chinese Urban Population Sampling Survey

We illustrate the modified estimator using data from 655 Chinese cities. Using 2010 census data, we order cities by population size and estimate the 2014 total population from a sample of 66 cities (approximately 10% of the population).

Step 1: Ordering by 2010 population yields $np_1 = 1.62 > 1$, satisfying the condition for the modified estimator.

Step 2: Compute $\frac{p_2}{n_2}$ for possible n_1 values. Table 4 shows the first ten values; $\frac{p_2}{n_2}$ increases thereafter. The minimum occurs at $n_1 = 3$, corresponding to the three largest cities: Chongqing, Shanghai, and Beijing.

Step 3: Select the remaining 63 units via PPS sampling. The full sample of 66 cities is listed in Table 5 .

Given the 2010 total urban population of 637,359,998, Equation (4) yields a 2014 estimate of 659,772,963 with an estimated standard error of 2,996,092. The traditional Hansen-Hurwitz estimator gives 658,451,503 with a standard error of 3,710,747—substantially larger. Repeated applications would confirm this pattern.

6 Summary and Outlook

The theoretical development and empirical validation demonstrate that the modified Hansen-Hurwitz estimator significantly outperforms the traditional ver-

sion. The ubiquity of the Zipf phenomenon suggests broad applicability for our method. When the condition $np_1 > 1$ is not satisfied, increasing the sample size can remedy the issue. Notably, this paper resolves the long-standing practical problem of determining how many "important" units to include with certainty by providing a clear methodology: select n_1 that minimizes $\frac{p_2}{n_2}$ among all feasible values. The two case studies confirm both the wide applicability and superiority of the modified estimator. Looking ahead, applying established empirical regularities to survey sampling promises further methodological advances and improved survey practice.

References

[1] Hansen M H, Hurwitz W N. On the Theory of Sampling from Finite Populations[J]. Annals of the Rheumatic Diseases. 1943, 70(12): 2111-2118.

[2] Yates F, Grundy P M. Selection without replacement from within strata with probability proportional to sue[J]. Journal of the Royal Statistical Society. 1953, 15(2): 253-261.

[3] Kim Y, Kim Y, Han H, et al. Efficiency of Variance Estimators for Two-stage PPS Systematic Sampling[J]. Korean Journal of Applied Statistics. 2013, 26(6): 1033-1041.

[4] Patel P A, Bhatt S. A Model-based Estimation of Finite population Variance under PPS Sampling[J]. Imperial Journal of Interdisciplinary Research. 2016, 2(4).

[5] Zou G H, Feng S Y. Admissibility of PPS Sampling with Replacement in the Class of Sampling with Replacement Designs[J]. Chinese Science Bulletin. 1995(08): 683-686.

[6] Sun S Z, Jiang T. Rotation Sampling for PPS Samples[J]. Application of Statistics and Management. 2002(04): 61-64.

[7] Liu J P, Chen G H. Extensions of the Hansen-Hurwitz Estimator under MPPS Sampling[J]. Statistical Research. 2005(05): 50-53.

[8] Chen G H, Cao W W. Sampling Estimation Methods and Application Research for Semiparametric Product Adjustment Models[J]. Application of Statistics and Management. 2017: 1-14.

[9] Li L L. Domain Estimation under Non-replacement Sample Augmentation Strategy Based on Brewer Sampling[J]. Application of Statistics and Management. 2017(04):

[10] Meng L B, Li E Q, Tian M Z. Construction of Confidence Intervals for Odds Ratio under Binomial Sampling Based on Saddlepoint Approximation[J]. Application of Statistics and Management. 2017(01): 85-102.

[11] Mi Z C, Li Y. Sampling Estimation for Capture-Removal Models for SNS Big Data[J]. Application of Statistics and Management. 2016(03):

[12] Zipf G. The Psycho-Biology of Language. An Introduction to Dynamic Philology[J]. Journal of Nervous & Mental Disease. 1935, 85(1): 93.

[13] Zhang Z Y. The Theoretical Basis and Practical Significance of Zipf' s Law[J]. Information Science. 1989(5): 62-66.

[14] Xu X Y. The Relationship between the 20/80 Rule and the Three Laws of Bradford, Zipf, and Lotka[J]. Library and Information Service. 2003(8): 39-42.

[15] Feng S Y, Ni J X, Zou G H. Sampling Survey Theory and Methods[M]. China Statistics Press, 1998.

Appendix: Proof that the $\frac{p_2}{n_2}$ Sequence Decreases Then Increases with n_1

Proof: For clarity, let $n_1 = 1, 2, \ldots, i, \ldots, n-1$ generate the sequence $a_{n_1} = \frac{p_2}{n_2}$:

$$a_i = \frac{\sum_{k=i+1}^{N} p_k}{n-i}, \quad i = 1, 2, \ldots, n-1$$

where p_k are decreasing and $p_k < 1$. When $n_1 = 1$, the condition for using the modified estimator requires $np_1 > 1$, so $p_1 > \frac{1}{n}$. Thus:

$$a_1 = \frac{1-p_1}{n-1}$$

For $n_1 = 2$:

$$a_2 = \frac{1-p_1-p_2}{n-2}$$

Examining the difference:

$$a_1 - a_2 = \frac{1-p_1}{n-1} - \frac{1-p_1-p_2}{n-2} = \frac{(n-2)(1-p_1) - (n-1)(1-p_1-p_2)}{(n-1)(n-2)}$$

For $a_1 > a_2$, we need $(n-2)p_2 > 1-p_1$. Since $p_1 > \frac{1}{n}$, this holds when $p_2 > \frac{1-p_1}{n-2}$. Given the ordering $p_1 > p_2 > \cdots$, this condition is typically satisfied for small n_1.

Similarly, for general i:

$$a_i - a_{i+1} = \frac{\sum_{k=i+1}^{N} p_k}{n-i} - \frac{\sum_{k=i+2}^{N} p_k}{n-i-1}$$

The sequence a_i decreases while p_{i+1} remains sufficiently large. However, as i increases, the numerator decreases faster than the denominator, eventually causing a_i to increase. The turning point occurs when p_{i+1} becomes small enough that the reduction in numerator is offset by the reduction in denominator.

Since $p_1 > p_2 > \cdots$ and $p_i < 1$, the sequence $\frac{p_2}{n_2}$ typically decreases initially then increases. If p_2 is already small, the sequence may increase monotonically. This completes the proof.

Note: Figure translations are in progress. See original paper for figures.