Sankhyā : The Indian Journal of Statistics
2017, Volume 79-B, Part 1, pp. 60-75

Estimation of Actual Proportion of Loanwords in a Language

Kalyan Joshi and M. B. Rajarshi
Savitribai Phule Pune University, Pune, India

Abstract

We suggest a modification of Probability Proportional to Size With Replacement (PPSWR) sampling method to estimate the actual proportion of loanwords in a language. When the proposed modification to PPSWR sampling can be implemented, the modification leads to an estimator which performs better than the estimator in a PPSWR sampling design. Under the assumptions that some words have a relatively very high frequency whereas most of the words have low frequency, we show that the estimator under the PPSWR scheme is better than the estimator under the Simple Random Sampling Without Replacement (SRSWOR). The suggested procedure has been applied to estimate the usage-based (actual) proportion of the Persian and Arabic loanwords in contemporary Marathi.

AMS (2000) *subject classification.* Primary 62; Secondary 62D05.
Keywords and phrases. Confidence interval, Marathi, Persian-Arabic loanwords, Probability proportional to size sampling, Simple random sampling without replacement, Zipf's law

1 Introduction

Borrowing of words or accepting words from other languages is a common property to almost all languages. Words from other languages are borrowed for various reasons such as cultural influence (import), trade, political influence (impose) and unavailability of specific (technical/cultural borrowings) words in the language. Borrowing the words for new objects (like "computer") or for new concepts (like "email") are examples of technical borrowings. Recipient language is the language that acquires a loanword and donor language is the one which is the source of the loanword. A loanword can be defined as a word that is transferred from a donor language to a recipient language and is used in the recipient language. It has been observed that nouns are more easily borrowed. Verbs are borrowed less. A possible reason is that a parallel verb already existing in the recipient language conveys

the same meaning as in the donor language or can be used in the possibly new context with new conventions of meanings being adopted. Nouns refer to objects, procedures, processes and so on. Some of these are totally absent in the recipient language. In some cases, a loanword is introduced in a language because an author of a document is from the donor language and is trying to address to individuals of the recipient language in their language.

Marathi, the language of Maharashtra, a Western state of India, is influenced by Persian and Arabic languages, as during 1350-1650, different parts of Maharashtra were ruled by Muslim rulers from Delhi, Hyderabad and Vijapur. During the eighteenth century, Persian was almost the official language of Peshwas, the rulers of Maharashtra. (Since there is a lot of give and take between Persian and Arabic, from now onwards we consider combining loanwords from these two languages into a single group, for the sake of our discussion). We are interested in estimating the proportion of Persian-Arabic words in contemporary Marathi. Of course, Marathi as a language has an influence of Persian with respect to style and other characteristics as well.

We illustrate the performance of our proposed method using the Marathi words from the online sources, which will be clearly addressed in this paper. Also, it may be of some interest to readers to know the *Mahārāṣṭra Śabdakośa* (Date et al., 1932-1950) is the largest dictionary of Marathi words. It has approximately 120,000 words out of which about 2900 words are of Persian-Arabic origin. (A few words in Marathi are formed by combining a Marathi word with a Persian-Arabic (PA) word.) Thus, the percentage of PA words in Marathi is about 2.42. Such a proportion will be referred to as the dictionary proportion and denoted by π. The dictionary proportion may not represent the true influence of loanwords, since the actual or in-usage proportion of such words depends upon the frequencies with which these words occur in practice or in a given corpus.

To estimate the actual proportion of loanwords, we suggest a modification of the PPSWR sampling.

1.1. Zipf's Law (cf. Zipf (1949)). Consider a collection of N words, with f_i and R_i as the frequency and the rank of the i-th word respectively. It is assumed that the list is arranged with decreasing frequencies. Thus, the word with highest frequency has rank 1. Then

$$R_i * f_i = k,$$

where k is constant.

The frequency of the word which is ranked first is k and that of the word which is ranked second is $k/2$, In general, the word which has rank j has frequency k/j. More general versions of Zipf's law can be found in Woodroofe and Hill (1975). Zipf noted that the second most common word in the English language ("of") appears at approximately half the rate of the most common word ("the"). The third most common word ("to") appears at approximately one-third the rate of the most common word. However, it appears that the law is not valid for all the words, particularly those whose frequencies are low.

Zipf's law has found to have applications in areas other than natural language processing. It was possibly first noted by Willis who was dealing with frequencies of biological genera with specified number of species, cf. Hill (1970). Woodroofe and Hill (1975) develop an urn model to give a probabilistic explanation of the Zipf's law. The Zipf's law has been applied to a number of situations. Gabaix (1999) fits the Zipf's model to cities and their populations. Zipf's law is frequently applied to distribution of incomes, cf. (Hill, 1970) and (Woodroofe and Hill, 1975). Thus, the sampling scheme proposed herein has potential applications in these areas also.

A common feature of many linguistic corpuses is that there are few words which have a very high frequency and a large number of words have a very low frequency. We refer to this property as *Zipf's property*. We now discuss this property with reference to our corpus. Our corpus consists of Marathi words from e-editions of Marathi newspapers, Marathi entries in Wikipedia and the Marathi encyclopedia (17 volumes) published by the Government of Maharashtra, (http://www.vishwakosh.org.in (a website of the Government of Maharashtra))which is available on the internet. The corpus has 3,721,628 words in all. Frequencies of all word are computed first. Then, root word of each word was found by considerations of inflections of Marathi words (Nouns and verbs have a large number of inflections in Marathi). Additions of all inflections of a root word give the frequency of that root word. We then have 318,443 distinct (root) words together with their frequencies.

Table 1 gives the percentages of frequencies as explained of percentages of 318,443 distinct words. The top 5 % words explain 5 % frequency, the top 10 % words have 10 % frequency, and the top 50 % frequency is explained by the first 548 words only. Therefore, our corpus does not exactly satisfy the Zipf's law, but it certainly has the Zipf's property.

As it is clear from Table 1, most of the words have a small frequency. To estimate the proportion of PA words in Marathi, we propose to modify the PPSWR design in such a manner that high frequency words are always

Table 1: Zipf's property

Percentage of words	5	10	15	20	25	30	35	40
Number of distinct words	5	10	17	29	46	75	127	209
Percentage of words	45	50	60	70	80	90	100	
Number of distinct words	334	548	1523	4406	13,977	56,634	318,443	

included in the sample. Although different authors such as Hidiroglou (1986) discuss a sampling design wherein some units are included in the sample with probability one, he considers SRSWOR and selection of units to be included in the sample indirectly depends upon values of the variable of interest. In our paradigm, the selection of such units depends only upon values of the size variable which determine the inclusion probabilities in PPSWR design and for the application to the estimation of proportion of loanwords, the size variable corresponds to (known) frequencies or proportion of all the words.

For remaining words, we use PPSWR sampling. After getting a sample, we classify each word as a PA word or non-PA word. We demonstrate how to select the number of high frequency words for a given sample size and show that such a Modified PPSWR (MPPSWR) sampling design results in a significant reduction in variance of the estimators of the actual proportion, if Zipf's property holds reasonably well. We consider with replacement sampling schemes, as it is quite cumbersome to carry out a Probability Proportional to Size Without Replacement (PPSWOR) sampling with such a large N and the sample size n.

Our modification sampling scheme differs from PPSWR. The main feature of our modification is certain inclusion of some units in the sample. These units have dominant values of z, the size variable. The number of such units depends on the chosen sample size, the population size and the values of z variables. The remaining units of the population are included in the sample via a PPSWR design with adjusted values of the size variable.

For such a huge corpus, classifying each and every word is very time-consuming, so that, techniques based on sampling from finite populations come handy. Classifying each word in loanwords and non-loanwords (with respect to a given donor language) require sizable resources in terms of time and expertize, setting aside the linguistic ambiguities and controversies regarding few words.

The PPSWR sampling scheme and suggested estimator can be useful in similar linguistics and other areas also. For example, it can be used to estimate the proportion of nouns (or any grammatical category of a word) in actual usage. The proposed methodology may be of potential use in other situations wherein Zipf's property holds.

Rest of the paper is organized as follows. In Section 2, we describe the MPPSWR sampling procedure together with estimators under MPPSWR and PPSWR procedures. In Section 3, estimators based on MPPSWR, PPSWR and SRSWOR are compared and it is shown that if conditions which correspond to Zipf's property hold, the estimator of actual proportion of loanwords based on MPPSWR is better than estimators based on the remaining two procedures. We also describe a procedure to obtain an optimal manner in which the modification is to be carried out. We further show that if Zipf's property holds reasonably well, PPSWR design based estimator is better than the SRSWOR design based estimator. In Section 4, a study is reported which compares the three designs. This section also gives the details of our corpus of Marathi words. In Section 5, we describe the application of the proposed MPPSWR sampling design to estimate the actual proportion of PA words in Marathi and to obtain a confidence interval for the unknown proportion. In Section 6, we offer some concluding remarks. In the Appendix, we give proofs of some of the results used in the paper.

2 Sampling Schemes and Corresponding Estimators

2.1. Notation. Let us assume that the population is of size N. Let f_i denote the value of the size variable (frequency of the i^{th} unit), $i = 1, 2, \ldots, N$. Let L_i be a binary variable which takes the value 1, if the i^{th} unit has a specified characteristic of interest, and 0, otherwise. Let p_i be defined by $p_i = f_i / \sum_{i=1}^{N} f_i$, $\sum_{i=1}^{N} p_i = 1$. We want to estimate $Y = \sum_{i=1}^{N} L_i p_i$ (in our application, this corresponds to the actual proportion of loanwords in the language) rather than $\sum_{i=1}^{N} L_i/N$, the proportion of units with the specified character. Let $M = \sum_{i=1}^{N} f_i$ the total frequency of all the units.

2.2. Simple Random Sampling Without Replacement. Let n be the sample size. The estimator of Y is given by

$$\hat{y}_{wor} = \frac{N}{n} \sum_{i=1}^{n} L_i p_i. \tag{2.1}$$

Throughout, it is understood that in summations which refer to a sample, the sum over i and j refers to units selected on the i-th or the j-th draw. It

is easy to see that $\hat{y}_{wor}$ is unbiased estimator of Y. The variance of $\hat{y}_{wor}$ is given by

$$V(\hat{y}_{wor}) = \frac{N-n}{n(N-1)} \left[\sum_{i=1}^{N} L_i p_i^2 (N-1) - 2 \sum_{i=1}^{N} \sum_{j>i}^{N} L_i L_j p_i p_j \right]. \tag{2.2}$$

2.3. Probability Proportional to Size with Replacement. Let n denote the total number of draws. On *each* draw, independently of outcomes on earlier draws, the j^{th} unit is included in the sample with probability p_j, $j = 1, 2, \ldots, N$. A natural estimator of Y is given by

$$\hat{y}_{pps} = \frac{1}{n} \sum_{i=1}^{n} L_i. \tag{2.3}$$

Let t_j denote the number of times the j^{th} population unit is included in the sample. Then, the above estimator can be written as $\frac{1}{n} \sum_{j=1}^{N} L_j t_j$. The random vector $(t_1, t_2, \ldots, t_N)$ has a multinomial distribution with parameters n and $(p_1, p_2, \ldots, p_N)$, cf. (Cochran, 1977; Sukhatme et al., 1984). It follows that $E(\hat{y}_{pps}) = \sum_{j=1}^{N} L_j p_j$,i.e., $\hat{y}_{pps}$ is unbiased estimator of Y. From the properties of the multinomial distribution, it can be shown that the variance of $\hat{y}_{pps}$ is given by

$$V(\hat{y}_{pps}) = \frac{1}{n} [\sum_{i=1}^{N} L_i p_i (1-p_i) - 2 \sum_{i=1}^{N} \sum_{j>i}^{N} L_i L_j p_i p_j]. \tag{2.4}$$

cf. (Cochran, 1977; Sukhatme et al., 1984). The above variance can also be written as

$$V(\hat{y}_{pps}) = \frac{1}{n} \sum_{i=1}^{N} p_i (L_i - Y)^2. \tag{2.5}$$

Further,

$$v(\hat{y}_{pps}) = \frac{1}{n^2} [\sum_{i=1}^{n} L_i (1-p_i) - \frac{2}{n-1} \sum_{i=1}^{n} \sum_{j>i}^{n} L_i L_j]. \tag{2.6}$$

is an unbiased estimator of $V(\hat{y}_{pps})$.

2.4. Modification to Probability Proportional to Size with Replacement. In view of the Zipf's property, we propose that units with dominant frequencies be always included in the sample. To achieve this, we divide the

entire population into two groups. We include all the units from the first group in the sample, whereas, we apply a PPSWR sampling design to the second group. Without any loss of generality, we assume that $p_1 \geq p_2 \geq \cdots \geq p_{N-1} \geq p_N$. Formally, let N_1 and N_2 be the number of units in the first and second group respectively. We include all the units from the first group in the sample and write $n_1 = N_1$. Let n_2 denote the number of draws from the second group of $N_2 = N - N_1$ units. Frequencies of units in the second group are less than those of units in the first group. Let t_{2i} denote the number of times the $i-$th unit in the second group is included in the sample, $i = N_1 + 1, N_1 + 2, \ldots, N$. The random vector $t_{2i}, i = N_1+1, N_1+2, \ldots, N$ has a multinomial distribution with parameters n_2 and $\frac{1}{M_2}(f_{N_1+1}, f_{N_1+2}, \ldots, f_N)$, where $M_2 = \sum_{i=1}^{N_2} f_{2i}$.Let $M_1 = \sum_{i=1}^{N_1} f_i$. The proposed estimator of Y corresponding to MPPSWR design is defined by

$$\hat{y}_{mpps} = \sum_{i=1}^{n_1} L_i p_i + \frac{M_2}{n_2 M} \sum_{i=1}^{n_2} L_i. \tag{2.7}$$

Let us consider $E[\hat{y}_{mpps}] = \sum_{i=1}^{N_1} L_i p_i + E[\frac{M_2}{n_2 M} \sum_{i=1}^{n_2} L_i] = \sum_{i=1}^{N_1} L_i p_i + \frac{M_2}{n_2 M} \sum_{i=N_1+1}^{N} L_i E(t_{2i}) = \sum_{i=1}^{N_1} L_i p_i + \frac{1}{M} \sum_{i=N_1+1}^{N} L_i f_i = Y$. Thus, $\hat{y}_{mpps}$ is an unbiased estimator of Y. Let $Y_2 = \sum_{i=N_1+1}^{N} L_i p_i$. To obtain $V(\hat{y}_{mpps})$, we note that the first term of $\hat{y}_{mpps}$ is non-random and the second term is related to a mean of PPSWR sample. Hence after some simplification we have,

$$\begin{aligned} & V(\hat{y}_{mpps}) \\ & = \frac{M_2}{M} \frac{1}{n_2} \left(\sum_{i=N_1+1}^{N} p_i (L_i - Y)^2 - \frac{M_2}{M} (Y - \frac{Y_2 M}{M_2})^2 \right) \end{aligned} \tag{2.8}$$

The Result 1 of the Appendix gives an unbiased estimator of the $V(\hat{y}_{mpps})$.

3 Comparison of Estimators

3.1. Comparison of Estimators Under SRSWOR and PPSWR Schemes. From (2.2), (2.4) and Result 2 of the Appendix, we have

$$\begin{aligned} & V(\hat{y}_{wor}) - V(\hat{y}_{pps}) \\ & = (\frac{N-n+1}{n}) \sum_{i=1}^{N} L_i p_i^2 + \frac{2}{n} (1 - \frac{N-n}{(N-1)}) \sum_{i=1}^{N} \sum_{j>i}^{N} L_i L_j p_i p_j - \frac{1}{n} \sum_{i=1}^{N} L_i p_i \end{aligned} \tag{3.1}$$

We note that $\frac{2}{n}(1-\frac{N-n}{(N-1)})$ is always positive. Thus, the difference between the two variances is at least

$$\begin{aligned}&(\frac{N-n+1}{n})\sum_{i=1}^{N}L_ip_i^2-\frac{1}{n}\sum_{i=1}^{N}L_ip_i\\&=\frac{1}{n}[(N-n+1)\sum_{i=1}^{N}L_ip_i^2-\sum_{i=1}^{N}L_ip_i]\end{aligned}\tag{3.2}$$

It is seen that $\sum_{i=1}^{N}L_ip_i\leq 1$, since L_i is 0 or 1 and $\sum_{i=1}^{N}p_i=1$. Thus, (3.1) is non-negative if

$$\sum_{i=1}^{N}L_ip_i^2\geq\frac{1}{N-n+1}.\tag{3.3}$$

Now, suppose that there are K units in the population, which have the specified characteristic of interest. Let these units be denoted by $i_1,i_2,\ldots,i_K$. Then, the sum $\sum_{i=1}^{N}L_ip_i^2$ is larger than $\sum_{j=1}^{K}p_{i_j}^2$. If this sum is larger than $\frac{1}{N-n+1}$, it follows that $V(\hat{y}_{wor})\geq V(\hat{y}_{pps})$. Thus, we need to have p_i's for units with the specified characteristic to be large enough. We may point out that the actual difference between the two variances may be much larger, since there are additional terms which are, in general, positive. This condition is satisfied if some of the units with a specified property of interest have a high frequency. If all such units are very rare, it is possible that the SRSWOR is a better sampling design for estimation of $\sum L_ip_i$.

Our sufficient condition (3.3) is well compatible with Zipf's property. In the context of our application to linguistics, if words from another language are well integrated in the language of interest (which is the case of PA words and Marathi: in fact, a person with Marathi as the mother-tongue is not aware that some of the loanwords are not originally Marathi) and the Zipf's property holds well; therefore, we anticipate that the MPPSWR design to be better than SRSWOR design.

For example, let k be the loanword with the largest proportion p_k . If $p_k>\frac{1}{\sqrt{N-n+1}}$, the PPSWR estimator has a smaller variance than the SRSWOR estimator. In our specific application to Marathi linguistics, we classify the first K words as loanwords and non-loanwords and examine

whether $\sum_{j=1}^{K} p_{i_j}^2 > 1$. In the case of our corpus, discussed in Section 4, we verify this condition.

3.2. Comparison of Estimator of Y Under the PPSWR and MPPSWR Sampling Schemes. From (2.5), (2.8) and Result 3 of the Appendix, we have,

$$\begin{aligned} V(\hat{y}_{pps}) - V(\hat{y}_{mpps}) \\ = \frac{1}{n}\sum_{i=1}^{N_1} p_i(L_i - Y)^2 + (\frac{1}{n} - \frac{M_2}{M}\frac{1}{n_2}) \sum_{i=N_1+1}^{N} p_i(L_i - Y)^2 \\ + (\frac{M_2}{M})^2 \frac{1}{n_2}(Y - \frac{Y_2 M}{M_2})^2. \end{aligned} \tag{3.4}$$

The first and last terms are both positive. The second term is non-negative, if we can ensure that $(\frac{1}{n} - \frac{M_2}{n_2 M}) \geq 0$. Now, $(\frac{1}{n} - \frac{M_2}{n_2 M}) > 0$, if $\frac{n_2}{n} > \frac{M_2}{M}$ or if $\frac{n_1}{n} < \frac{M_1}{M}$. This condition is a consequence of Zipf's property which implies that the frequencies of first N_1 units are large enough, results to the ratio $\frac{M_1}{M}$ exceeds $\frac{n_1}{n}$.

We thus conclude that, if $\frac{n_1}{n} < \frac{M_1}{M}$, $V(\hat{y}_{pps}) - V(\hat{y}_{mpps}) > 0$ so that the MPPSWR scheme is better than the PPSWR scheme. This sufficient condition can be rewritten as $\frac{n_1}{n} < \frac{\sum_{i=1}^{N_1} f_{1i}}{\sum_{i=1}^{N} f_i}$.

It may be pointed out that the above requires us to know only the frequencies or proportions of all the words and no other information is required to decide n_1. Further, with a larger n, it is more likely that we can find n_1, as required.

3.3. Optimal Value of n_1 for the MPPSWR Scheme. To maximize gain due to MPPSWR scheme, we find n_2^* such that $(\frac{1}{n} - \frac{M_2}{n_2 M})$ is maximum , i.e.,

$$(\frac{1}{n} - \frac{M_2}{n_2 M}) \leq (\frac{1}{n} - \frac{M_2}{n_2^* M})$$

for all value of n_2. This also gives the optimal value $n_1^* = n - n_2^*$. In practice, one starts with values of n_1 equal to 1, 2, and so on and chooses the largest n_1 for which the above holds.

In some cases, it may turn out that $(\frac{1}{n} - \frac{M_2}{n_2 M})$ is negative for $n_1 = 1$ which means that modified PPSWR can not be implemented and in such a case, we continue to use PPSWR.

4 A Sampling Study

As discussed in Section 1, we have a corpus of $N = 318,443$ words together with their frequencies as our finite population. We first assign a dictionary proportion (π) of words and then randomly assign a value of 1 or 0 to each word, so that the proportion of words with 1 as the assigned value is π. In this case, $L_i = 1$ if the $i-$th word is a loanword, i.e., the $i-$th word is borrowed from Persian-Arabic languages. Thus, the parameter of interest is $\sum_{i=1}^{N} L_i p_i$, the proportion of loanwords in the entire corpus or in the actual use.

In our study of variances of the three estimators, we take $n = 5000$, the sample size in the case of without replacement sampling and the number of draws, in the case of with replacement sampling. Based on the frequencies of all the words, the number of units in the first group is obtained to be $n_1 = N_1 = 892$, by the procedure to obtain n_1 as described in Section 3.3. Thus, we always include the first n_1 units in the MPPSWR sample and from the second group of $N - n_1$ units, we have a PPSWR sample of size $n_2 = 5000 - 892$. Variances of the estimators as given by expressions (2.2), (2.4) and (2.8) under SRSWOR, PPSWR and MPPSWR designs are then computed. This exercise was carried out for different values of π in the range of (0,0.20).

It is seen from Table 2 that the MPPSWR estimator is more efficient than both the PPSWR and SRSWOR estimators. The conditions for a better performance of PPSWR as compared to SRSWOR and of MPPSWR as compared to PPSWR are easily satisfied in each case.

Figure 1 gives plots of variances of estimators under PPSWR and MPPSWR designs. It clearly brings out that the gain due to MPPSWR design as compared to PPSWR improves rapidly as the dictionary proportion π increases.

5 Application to Estimation of Actual Proportion of Persian-Arabic Loanwords in Marathi

In the present case, we have N=318,443 and $M = 37,21,628$. Further, we take n, the number of draws of the MPPSWR sample, to be 5000. Now, we find the first PA loanword and observe that its frequency is 40,122 , so that the corresponding $p_k = 0.010781$ which exceeds $\frac{1}{\sqrt{313,443}} = 0.001786$. Thus, in this case, from the discussion in section subsequent to (3.3) in Section 3, it follows that the variance of the usual estimator under PPSWR is less than the variance under SRSWOR. Based on the above mentioned calculations,

Table 2: Variances of PPS, SRSWOR and MPPS

True π	True Y	Variance of SRSWOR	Variance of PPSWR	Variance of MPPSWR
0.01	0.01078	1091	2.132	0.452
0.02	0.02010	1513	3.939	0.932
0.03	0.02366	366	4.620	1.538
0.04	0.02927	821	5.683	1.931
0.05	0.03822	935	7.352	2.384
0.09	0.07836	3339	14.444	4.157
0.10	0.08630	4729	15.771	4.328
0.06	0.08858	26474	16.147	2.698
0.08	0.09057	11342	16.474	3.700
0.07	0.09565	25133	17.301	3.240
0.12	0.12605	19700	22.032	5.330
0.13	0.12764	19362	22.269	5.587
0.11	0.13402	32846	23.212	4.803
0.15	0.15201	11622	25.780	6.492
0.14	0.15520	25376	26.222	5.980
0.18	0.16410	16127	27.434	7.228
0.16	0.17424	28216	28.776	6.667
0.19	0.18105	22055	29.655	7.700
0.17	0.19011	31668	30.793	6.924
0.20	0.20026	24802	32.031	7.956

To get actual variances from column numbers 3, 4 and 5, we multiply given numbers by 10^{-6}

we anticipate the variance of estimator under MPPSWR scheme to be far less than the estimators under the PPSWR scheme as well.

We now estimate Y, the actual proportion of PA loanwords in our corpus. The sample size is as before 5000; we draw a MPPSWR sample with $n_1 = 892$. Most of the words in the sample were classified into PA and non-PA categories by referring to (Kulkarni, 1993) and in some cases, experts were consulted. The estimate of Y is 0.050448 with a standard error of 0.001245. Thus, based on the corpus that we have, it is estimated that there are about 5 % PA words in the actual usage of contemporary Marathi. A 99 % large sample confidence interval for Y, based on the normal approximation to the estimator under MPPSWR design, is given by $(0.0472359, 0.05366)$, the dictionary proportion (0.0242) being far away.

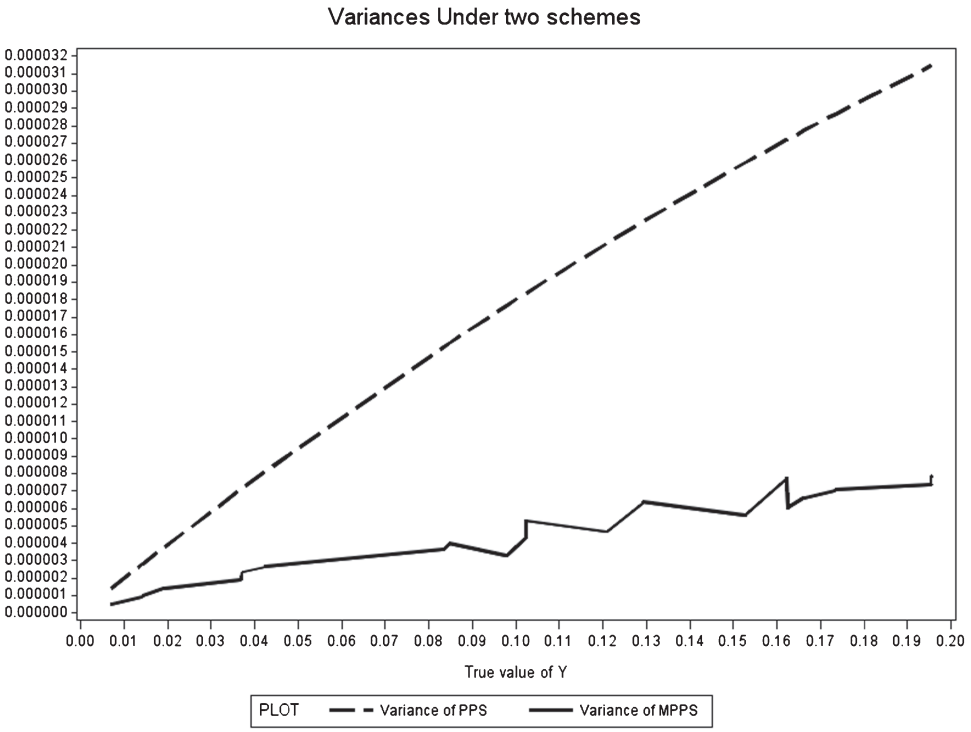

Figure 1: Variance of PPS and MPPS

6 Concluding Remarks

We have developed a modified PPSWR sampling scheme which gives a better estimator of the population proportion of a characteristic, when the population has a size variable which has Zipf's property. We need only the information regarding the size variable to implement the suggested modification. The modification involves guaranteed inclusion of dominant units, i.e., units with very large value of the size variable. Its implementation is very easy and depends only on known frequencies or proportions. When Zipf's property holds, such a design offers a design which is better than SRSWOR also. We have discussed in detail an application to natural language processing. Our methodology can be used to estimate the actual proportion of a specified word category such as adverbs, prepositions and so on. We hope to discuss such applications elsewhere.

As discussed in Section 1, the Zipf's property has been found to hold in situations other than linguistics and thus the MPPSWR scheme developed

herein can be applied to these situations also. For example, in the context of a population of towns (units) and their populations (size variable), one can apply the proposed scheme to estimate proportion of population which has a satisfactory utility available, such as public transport.

Acknowledgments. We sincerely thank the two reviewers, the Associate Editor and the Editor for helpful comments and suggestions which considerably improved presentation of the results. The second author wishes to thank the University Grants Commission, India for a research fellowship. Thanks are due to Dr Akanksha Kashikar for a careful reading of the manuscript and helpful comments.

Appendix

Result 1: An unbiased estimator of $V(\hat{y}_{mpps})$.

We note that

$$\begin{aligned}
&V(\hat{y}_{mpps})\\
&= (\frac{M_2}{M})^2 V(\frac{1}{n_2}\sum_{i=1}^{n_2} L_i)\\
&= (\frac{M_2}{M})^2 \frac{1}{n_2}\left(\sum_{i=N_1+1}^{N} L_i \frac{f_i}{M_2}(1-\frac{f_i}{M_2}) - 2\sum_{i=N_1+1}^{N}\sum_{j=N_1+1, j>i}^{N} L_i L_j \frac{f_i}{M_2}\frac{f_j}{M_2}\right)\\
&= \frac{1}{n_2}\left(\sum_{i=N_1+1}^{N} L_i p_i (\frac{M_2}{M} - p_i) - 2\sum_{i=N_1+1}^{N}\sum_{j=N_1+1, j>i}^{N} L_i L_j p_i p_j\right).
\end{aligned}$$

An unbiased estimator of the above variance is given by

$$v(\hat{y}_{mpps}) = (\frac{M_2}{Mn_2})^2\left(\sum_{i=1}^{n_2} L_i(1-\frac{f_i}{M_2}) - \frac{2}{n_2-1}\sum_{i=1}^{n_2}\sum_{j>i}^{n_2} L_i L_j\right).$$

A proof of the unbiasedness is given below. Consider

$$\begin{aligned}
&E(v(\hat{y}_{mpps}))\\
&= E\left((\frac{M_2}{Mn_2})^2[\sum_{i=1}^{n_2} L_i(1-\frac{f_i}{M_2}) - \frac{2}{n_2-1}\sum_{i=1}^{n_2}\sum_{j>i}^{n_2} L_i L_j]\right)\\
&= (\frac{M_2}{Mn_2})^2 E\left(\sum_{i=1}^{n_2} L_i(1-\frac{f_i}{M_2}) - \frac{2}{n_2-1}\sum_{i=1}^{n_2}\sum_{j>i}^{n_2} L_i L_j\right)
\end{aligned}$$

$$= (\frac{M_2}{Mn_2})^2\Bigg(\sum_{i=N_1+1}^{N} L_i E(t_{2i})(1-\frac{f_i}{M_2}) - \frac{2}{n_2-1}\sum_{i=N_2+1}^{N}\sum_{j=N_1+1,j>i}^{n_2} L_iL_jE(t_{2i}t_{2j})\Bigg).$$

Now, we know that $E(t_{2i}) = n_2\frac{f_i}{M_2}$ and $Cov(t_{2i},t_{2j}) = -n_2\frac{f_i}{M_2}\frac{f_j}{M_2}$, $i \neq j$. Thus,

$$\begin{aligned}
&E(v(\hat{y}_{mpps}))\\
&= (\frac{M_2}{Mn_2})^2\Bigg(\sum_{i=N_1+1}^{N} L_i n_2 \frac{f_i}{M_2}(1-\frac{f_i}{M_2}) - \frac{2}{n_2-1}\sum_{i=N_2+1}^{N}\sum_{j=N_1+1,j>i}^{n_2} L_iL_j n_2(n_2-1)\frac{f_i}{M_2}\frac{f_j}{M_2}\Bigg)\\
&= (\frac{1}{n_2})^2\left(\sum_{i=N_1+1}^{N} L_i n_2 \frac{f_i}{M}(\frac{M_2}{M}-\frac{f_i}{M}) - 2\sum_{i=N_2+1}^{N}\sum_{j=N_1+1,j>i}^{n_2} L_iL_j n_2\frac{f_i}{M}\frac{f_j}{M}\right)\\
&= \frac{1}{n_2}\left(\sum_{i=N_1+1}^{N} L_i p_i(\frac{M_2}{M}-p_i) - 2\sum_{i=N_1+1}^{N}\sum_{j=N_1+1,j>i}^{N} L_iL_j p_ip_j\right)\\
&= V(\hat{y}_{mpps}).
\end{aligned}$$

Result 2: Proof of equation (3.1). The variance of $\hat{y}_{mpps}$ can be rewritten as,

$$\begin{aligned}
&V(\hat{y}_{mpps})\\
&= (\frac{M_2}{M})^2\frac{1}{n_2^2}V(\sum_{i=1}^{n_2} L_i)\\
&= (\frac{M_2}{M})^2\frac{1}{n_2^2}V(\sum_{i=N_1+1}^{N} L_i t_{2i})\\
&= \frac{1}{n_2}\left(\frac{M_2}{M}\sum_{i=N_1+1}^{N} L_i p_i - Y_2^2\right)\\
&= \frac{M_2}{M}\frac{1}{n_2}\sum_{i=N_1+1}^{N} p_i(L_i - \frac{Y_2}{\frac{M_2}{M}})^2\\
&= \frac{M_2}{M}\frac{1}{n_2}\sum_{i=N_1+1}^{N} p_i(L_i - Y + Y - \frac{Y_2M}{M_2})^2
\end{aligned}$$

$$= \frac{M_2}{M}\frac{1}{n_2}\left(\sum_{i=N_1+1}^{N} p_i(L_i - Y)^2 - \frac{M_2}{M}(Y - \frac{Y_2 M}{M_2})^2\right).$$

Therefore, difference between variances under PPS and SRSWOR schemes is given by

$$\begin{aligned}
&V(\hat{y}_{wor}) - V(\hat{y}_{pps}) \\
&= (\frac{N-n}{(N-1)n})\left(\sum_{i=1}^{N} L_i p_i^2 (N-1) - 2\sum_{i=1}^{N}\sum_{j>i}^{N} L_i L_j p_i p_j\right) \\
&\quad - \frac{1}{n}\left(\sum_{i=1}^{N} L_i p_i (1-p_i) - 2\sum_{i=1}^{N}\sum_{j>i}^{N} L_i L_j p_i p_j\right) \\
&= (\frac{N-n}{n})\sum_{i=1}^{N} L_i p_i^2 - 2(\frac{N-n}{(N-1)n})\sum_{i=1}^{N}\sum_{j>i}^{N} L_i L_j p_i p_j \\
&\quad - \frac{1}{n}\sum_{i=1}^{N} L_i p_i + \frac{1}{n}\sum_{i=1}^{N} L_i p_i^2 + \frac{2}{n}\sum_{i=1}^{N}\sum_{j>i}^{N} L_i L_j p_i p_j \\
&= (\frac{N-n+1}{n})\sum_{i=1}^{N} L_i p_i^2 + \frac{2}{n}(1 - \frac{N-n}{(N-1)})\sum_{i=1}^{N}\sum_{j>i}^{N} L_i L_j p_i p_j \\
&\quad - \frac{1}{n}\sum_{i=1}^{N} L_i p_i,
\end{aligned}$$

which is equation (3.1) of the paper.

Result 3: Proof of equation (3.4). The difference between variances under PPSWR and MPPSWR schemes is given by

$$\begin{aligned}
&V(\hat{y}_{pps}) - V(\hat{y}_{mpps}) \\
&= \frac{1}{n}\left(\sum_{i=1}^{N_1} p_i(L_i - Y)^2 + \sum_{i=N_1+1}^{N} p_i(L_i - Y)^2\right) \\
&\quad - \frac{M_2}{M}\frac{1}{n_2}\left(\sum_{i=N_1+1}^{N} p_i(L_i - Y)^2\right) + \frac{M_2}{M}\frac{1}{n_2}\left(\frac{M_2}{M}(Y - \frac{Y_2 M}{M_2})^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{N_1} p_i(L_i - Y)^2 + (\frac{1}{n} - \frac{M_2}{M}\frac{1}{n_2})\sum_{i=N_1+1}^{N} p_i(L_i - Y)^2
\end{aligned}$$

$$+ (\frac{M_2}{M})^2 \frac{1}{n_2} (Y - \frac{Y_2 M}{M_2})^2,$$

which is equation (3.4) of the paper.

References

COCHRAN, W. G. (1977). *Sampling Theory Third Edition.* Wiley, New York.

DATE, Y.R., KARVE, C.G., CHANDORKAR, A. and DATAR, C.S. (1932-1950) *Mahārāṣṭra Śabdakośa* 7 volumes plus a supplement, Maharashtra Kosamandala Limited, Pune.

GABAIX, X. (1999) Zipf's Law and the Growth of Cities The American Economic Review 89 Papers and Proceedings of the One Hundred Eleventh Annual Meeting of the American Economic Association.

HIDIROGLOU, M. A. (1986). The construction of a self-representing stratum of large units in survey designs. *American Statistician* **40**, 27–31.

HILL, B. M. (1970). Zipf's Law and Prior Distributions for the Composition of a Population. *Journal of the American Statistical Association* **65**, 1220–1232.

KULKARNI, K. P. (1993). *Marāṭhī Vyutpatti Kośa (Marathi Etymological Dictionary)* Shubhada Saraswat Prakashan, Third Edition, Pune.

SUKHATME, P. V., SUKHATME, B.V., SUKHATME, S. and ASOK, C. (1984). *Theory of sample surveys with applications.* Indian Society for Agricultural Statistics, New Delhi.

WOODROOFE, M. and HILL, B. (1975). On Zipf's Law. *Journal of Applied Probability* **12**, 425–434.

ZIPF, G.K. (1949). *Human Behaviour and the Principles of Least Effort* Addison-Wesley.

Kalyan Joshi
M. B. Rajarshi
Department of Statistics,
Savitribai Phule Pune University,
Pune, 411007, India
E-mail: kalyan.joshi@gmail.com

Paper received: 24 December 2014.