ISSN 0971-0388

Aligarh Journal of Statistics
Vol. 43(2023) 139-148

Application of Modified PPSWR to Estimate the Actual Proportion of Indeclinable Words of Hitopadesha

Anupama Goyal, Pooja Choraria, Versha Dwivedi, P. C. Gupta
[Received on March, 2022. Accepted on April, 2023]

ABSTRACT

Probability Proportional to Size with Replacement (PPSWR) Sampling offers an important role in the sampling theory. K. Joshi and M.B. Rajarshi (2017) proposed Modified Probability Proportional to Size with Replacement Sampling, which lead to the estimators with higher efficiency in case where data follow Zipf's Law. Hitopadesha, a linguistic corpus in Sanskrit language consists of some words having very high frequency than other words hence the data of word frequency follows Zipf's Law. In this paper, we illustrate the performance of MPPSWR Sampling to estimate the actual proportion of Indeclinable words of Hitopadesha and compare the estimators with PPSWR Sampling.

1. Introduction

In linguistic, the smallest component with a practical meaning is a word. In this paper, we are considering special type of words known as Indeclinable words (or Avikari shabd). These are the words that do not change themselves no matter in which form like adverbs, post position, interjection or conjunctions, we are using them. Every language has some set of indeclinable words which may be in the form of connectives, exclamation or collapse too.

Sanskrit, the language of primeval India has a history of about 3500 years. It is the holy language of Hinduism. Most works of Hindu philosophy as well as few

: Anupama Goyal
Email: anupamagoyal6198@gmail.com

Extended authors information available after reference list of the article.

of the principal text of Jainism and Buddhism are written in Sanskrit. In the early 1st millennium CE (Common Era), Sanskrit migrated to Southeast Asia, Central Asia and parts of East Asia along with Buddhism and Hinduism, emerging as a high cultured language. It is an old Indo-Aryan language so it has an eminent place in Indo-European studies.

Hitopadesha is an Indian text in Sanskrit language consisting of literary genre having both human and animal characters. People know less about its origin but it is believed to be from 12th century. It is composed by Narayana probably between 800 to 950CE. The Hitopadesha has several versions of books available. The shortest version has 655 verses while longest has 749 verses. So, it is quite tedious to work on whole book, therefore we are using sampling techniques to estimate the actual proportion of indeclinable words.

Joshi *et al.* (2017) suggested 'MPPSWR sampling to estimate the actual proportion of Persian-Arabic loanwords in Marathi language where some units are selected in the sample with probability 1'.

In our study, we are interested in estimating the actual proportion of indeclinable words in a Sanskrit book named "Hitopadesha". Of course, Sanskrit as a language has some set of indeclinable words. So, we demonstrate the working of a sampling method proposed by Joshi *et al.* (2017) using the Sanskrit words from the book Hitopadesha.

If the units in the population varying in their size, then applying SRS to select the sample is not adequate, as it does not give any importance to the size of the units. To get more efficient estimators of the population parameters, the probability of selection may be assigned in proportion to the size of the unit, when the size of the units varies and the study variate is highly correlated with the unit size. So, we use a sampling procedure known as Probability Proportional to Size (PPS) sampling. Under this sampling procedure we have two methods of selecting a sample. First is cumulative total method and second is Lahiri's method. Here, we are using cumulative total method to draw our sample with the help of random number table given in the book Sukhatme *et al.* (1984).

We get the list of our population units along with the size of the units on Wiktionary (online source) of total words i.e., 20218 in which 9259 distinct

words are included and we are selecting a sample of size 370 from the population of size 9259 (Sukhatme, 1984).

Zipf's Law (cf. Zipf's (1949)): Zipf's Law is a statistical distribution in certain data sets, like occurrence of words in a linguistic corpus, in which frequencies of certain words are inversely proportional to their ranks. Initially, George Kingsley Zipf (1935) highlighted this property of words. This law examines words frequencies. It can be understood by an example of the most common word in English "the", which in a typical text appears about one-tenth of the time (rank1); next most common word is "of" (rank 2) which appears about one-twentieth of the time. Frequencies decline sharply in this type of distributions, as the rank number increases. So, a large number occur rarely and small number of items appears very often.

"Consider a collection of N words, with f_i and R_i as the frequency and rank of the i^{th} word respectively. As the list is arranged with decreasing frequency thus, rank 1 is given to the word with highest frequency. Then,

$$R_i * f_i = k_i \tag{1.1}$$

where k is a constant."

Except working of language, we can apply Zipf's Law in several fields. It is used by Willis in his study as he considered the several species and noted down the occurrence of biological genera follows Zipf's Law, c.f. Hill (1970). Woodroofe *et al.* (1975) frame a model of urn through which he explained the probabilistic behaviour of Zipf's Law. Gabaix (1999) applied Zipf's Law on the cities as units and their populations as corresponding frequencies. (Hill, 1970) and (Woodroofe *et al*., 1975) applied Zipf's Law to the distribution of income. Cunha et al. (1995) observed that the data on frequency of website visited also follows the Zipf's Law. Hence, we can say that the proposed sampling scheme by Joshi *et al*. has wide variety of applications.

Many linguistics corpuses have a common feature of having some words with high frequency and many words with low frequency. Such behaviour of words follows Zipf's Law. Now, this property can be seen with respect to our variable under study of the population corpus.

The collection of writings consists of Sanskrit words from book "Hitopadesha" listed in Wiktionary which is available on internet. The corpus having 9259

distinguished words in all, of which frequencies i.e., size of the units is available along with sampling units. Here we can see that a large number of words are occurring with very less frequency.

In our study, we are using MPPSWR for estimating the actual proportion of indeclinable words for study population, that the units with high frequency will always be included in the sample. We select such units by considering the size of the population units as it determines the probability of getting selected in the sample and for estimating the proportion of indeclinable words. Rest of the words in the sample are drawn by PPSWR Sampling. In our corpus, we show that if the data follows Zipf's Law the MPPSWR Sampling results in lesser variance of the estimators. The units in the sample are drawn with replacing the units drawn in previous draws as PPSWOR Sampling with large N is difficult in practical application.

In this sampling scheme, units with higher occurrence are certainly included in the sample i.e., first group. The number of first group units with higher occurrence is calculated using the population size, the corresponding size of the population units and chosen sample size. PPSWR Sampling is used to include the remaining population with adjusted values of the size variables.

Similarly, we can use suggested estimators in PPSWR Sampling for several languages as well as in other fields also. Gi Sung Lee (2012) estimate a rare sensitive attribute in PPS measures using Poisson distribution, Sajinder Singh (2003) suggested new techniques to calibrate estimators of variance using PPSWR. Joshi *et al.* (2017) estimated the proportion of loanwords in Marathi language using MPPSWR Sampling.

We aim to study the applications of MPPSWR Sampling scheme as there exists several situations where the population follows Zipf's Law viz. vary in their size and, the size variable depends upon the characteristic under study. Applying PPSWR Sampling in these cases will not give adequate results then MPPSWR Sampling. So, we aim to establish the applicability of MPPSWR Sampling by implementing it to other examples of Zipf's Law such as other linguistic corpus, cities and its population, website visiting frequency and other fields.

The modified method also has a very vast field of application for example, it can be useful in similar languages to estimate the proportion of nouns in actual usage.

One of the applications in linguistic is numerically illustrated using real data in this paper to estimate the actual proportion of Indeclinable words of book "Hitopadesha".

Illustration: As mentioned above that Gabaix (1999) fits the Zipf's Law on cities and their population. Suppose we want to estimate the actual proportion of population living in a state having under-5 mortality to at least as low as 25 per 1,000 live births i.e., the goal 3 target in Sustainable development goals by 2030. The districts in the state are first arranged in their decreasing order of their population. If the population follows the Zipf's Law approximately then MPPSWR Sampling technique can be applied. In this case, districts are considered as units and their population as the size of the sampling units.

2. Sampling Scheme

Notations: Let N be the population size and f_i be the frequency of corresponding i^{th} unit, $i = 1,2, \dots, N$. Let L_i be the indicator function such that

$$L_i = f(x) = \begin{cases} 1, & if\ i^{th}\ unit\ is\ indeclinable\ word \\ 0, & Otherwise \end{cases}$$

And $$p_i = \frac{f_i}{\sum_{i=1}^{N} f_i} \quad ; \sum_{i=1}^{N} p_i = 1$$

Here, we want to estimate the actual proportion of indeclinable words

$$Y = \sum_{i=1}^{N} L_i p_i \tag{2.1}$$

(Rather than $\sum_{i=1}^{N} L_i/N$, proportion of units). And

$$M = \sum_{i=1}^{N} f_i \tag{2.2}$$

is the total frequency.

PPSWR: Let n be the sample size. In PPSWR Sampling each draw is independent with the previous draw, and included in the sample with probability p_j of selection, where j is the sampling unit, $j = 1, 2, \dots, N$. Estimator of Y is given by

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

Now, j^{th} unit is included in the sample t_j times. So,

$$\hat{y}_{pps} = \frac{1}{n}\sum_{i=1}^{N} L_j t_j \tag{2.4}$$

And vector $(t_1, t_2, \ldots, t_N)$ follows multinomial distribution with parameter $(p_1, p_2, \ldots, p_N)$ and n.

$$E(\hat{y}_{pps}) = E\left(\frac{1}{n}\sum_{i=1}^{N} L_i\, t_i\right)$$

$$E(\hat{y}_{pps}) = \frac{1}{n}\sum_{i=1}^{N} L_i\, E(t_i)$$

$$E(\hat{y}_{pps}) = \sum_{i=1}^{N} L_i p_i \tag{2.5}$$

Also, variance is given by,

$$V(\hat{y}_{pps}) = \frac{1}{n}\left[\sum_{i=1}^{N} L_i p_i\,(1-p_i) - 2\sum_{i=1}^{N}\sum_{j>i}^{N} L_i L_j p_i p_j\right] \tag{2.6}$$

Variance can also be written as,

$$V(\hat{y}_{pps}) = \frac{1}{n}\left[\sum_{i=1}^{n} p_i\,(L_i - Y)^2\right] \tag{2.7}$$

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

$$v(\hat{y}_{pps}) = \frac{1}{n^2}\left[\sum_{i=1}^{n} L_i(1-p_i) - \frac{2}{n-1}\sum_{i=1}^{n}\sum_{j>i}^{n} L_i L_j\right] \tag{2.8}$$

MPPSWR: If the data follows the Zipf's Law then we include higher frequency units in the sample with probability 1. Hence the population after arranging in descending order of their frequency is divided in two parts. One part of higher frequency units of the population is included with fixed probability 1 and the rest of the sample is drawn from the remaining population units using PPSWR Sampling.

Suppose N_1 is the number of sampling units selected in the sample with probability 1 and N_2 is the remaining population. The total sample size n is also divided in two parts such that $n_1 = N_1$ and the remaining sample of size n_2 is drawn with PPSWR Sampling from population $N_2 = N - N_1$. Let the number of times i^{th} unit is selected from the second part be represented by t_{2i} $(i = N_1 + 1, N_1 + 2, \ldots, N)$. t_{2i} will follow multinomial distribution with parameters n_2 and $\frac{1}{M_2}\left(f_{N_1+1}, f_{N_1+2}, \ldots, f_N\right)$, $M_2 = \sum_{i=1}^{N_2} f_{2i}$, $M_1 = \sum_{i=1}^{N_2} f_i$ such that $M = M_1 + M_2$.

According to MPPSWR Sampling, estimate of Y is given by,

$$\hat{y}_{mpps} = \sum_{i=1}^{n_1} L_i p_i + \frac{M_2}{n_2 M}\sum_{i=1}^{n_2} L_i \tag{2.9}$$

Now,

$$E(\hat{y}_{mpps}) = \sum_{i=1}^{N_1} L_i p_i + E\left[\frac{M_2}{n_2 M}\sum_{i=1}^{n_2} L_i\right]$$

$$= \sum_{i=1}^{N_1} L_i p_i + \frac{M_2}{n_2 M}\sum_{i=N_1+1}^{N} L_i\, E(t_{2i})$$

$$= \sum_{i=1}^{N_1} L_i p_i + \frac{1}{M}\sum_{i=N_1 1}^{N} L_i f_i$$

$$= Y \qquad (2.10)$$

Hence, $\hat{y}_{mpps}$ is an unbiased estimator of Y.

Since, first term of the estimator is non-random. So, the variance only depends on the second term. Let,

$$Y_2 = \sum_{i=N_1+1}^{N} L_i p_i \qquad (2.11)$$

So, variance is given by

$$V(\hat{y}_{mpps}) = \frac{1}{n_2}\sum_{i=N_1+1}^{N} L_i p_i \left(\frac{M_2}{M} - p_i\right) - 2\sum_{i=N_1+1}^{N}\sum_{j=N_1+1>i}^{N} L_i L_j p_i p_j \qquad (2.12)$$

And,

$$V(\hat{y}_{mpps}) = \frac{M_2}{n_2 M}\sum_{i=N_1+1}^{N} p_i (L_i - Y)^2 - \frac{M_2}{M}\left(Y - \frac{Y_2 M}{M_2}\right)^2 \qquad (2.13)$$

The estimate of variance is,

$$v(\hat{y}_{mpps}) = \left(\frac{M_2}{n_2 M}\right)^2 \left[\sum_{i=1}^{n_2} L_i \left(1 - \frac{f_i}{M_2}\right) - \frac{2}{n_2 - 1}\sum_{i=1}^{n_2}\sum_{j>i}^{n_2} L_i L_j\right] \qquad (2.14)$$

Comparison of PPSWR and MPPSWR

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

$$+ \left(\frac{M_2}{M}\right)^2 \frac{1}{n_2}\left(Y - \frac{Y_2 M}{M_2}\right)^2 \qquad (2.15)$$

We observe that the first and last terms are positive and to maximize the gain second term should be positive i.e., $\left(\frac{1}{n} - \frac{M_2}{M n_2}\right) > 0$. If the term $\left(\frac{1}{n} - \frac{M_2}{M n_2}\right)$ is maximum then the gain will be maximum.

Hence the value of $n_2 = n_2^*$ should be such that $\left(\frac{1}{n} - \frac{M_2}{Mn_2}\right) \leq \left(\frac{1}{n} - \frac{M_2}{Mn_2^*}\right)$. Also $\left(\frac{1}{n} - \frac{M_2}{Mn_2}\right)$ should be positive for $n_1 = 1$ else MPPSWR can't give better results than PPSWR.

3. Numerical Illustration

In this study, we consider the book "Hitopadesha" having $N = 9259$ words with total frequency 20218. We wish to estimate the actual proportion of indeclinable words in the book. We draw a sample of size $n = 370$ from the total population of distinct words.

First, the total population is arranged in decreasing order of their frequencies and the optimal value i.e., $N_1 = 35$ is obtained by optimising the function $\left(\frac{1}{n} - \frac{M_2}{Mn_2}\right)$. The first part of size $N_1 = n_1$ is completely included in the sample that contains high frequency units and the remaining population is drawn from the second part of the population viz, $N_2 = N - N_1 = 9259 - 35 = 9223$ using PPSWR Sampling.

The PPSWR sample is drawn using cumulative total method by replacing the units drawn in previous draws. Then we assign $L_i = 1$, if the word is an indeclinable word and $L_i = 0$, if the word is not indeclinable for the complete sample of size n. This is done using the aid of the book "Vyakaran Siddhant Komji". Thus, we want to estimate $\sum_{i=1}^{N} L_i p_j$, the proportion of indeclinable words in the book as a whole.

The estimates for the population parameter are calculated using the MPPSWR Sampling proposed by Joshi *et al.* (2017). The results are then compared with estimates using PPSWR Sampling.

In our study of proportion of indeclinable words, $N = 9259$, $M = 20218$ and $n = 370$. The samples of PPSWR and MPPSWR are both classified as indeclinable and declinable words and assigned 1 and 0 respectively. At 95% of confidence interval, the estimate of population parameter Y using PPSWR Sampling is 0.113514 with variance 0.00027123618 while using MPPSWR Sampling the estimate of Y is 0.177774262 with variance is 0.000117498554.

4. Conclusion

Joshi *et al.* (2017) suggested a MPPSWR Sampling scheme for the data distribution where the population parameter depends upon the size of the sampling units. Also, the data follows Zipf's Law where size of the highest frequent unit is approximately double to the second most frequent unit and so on. If the data follows Zipf's Law then some units of higher frequency are selected in the sample with probability 1. The certain inclusion of higher frequency units in the sample decreases the variance of the estimate in case of MPPSWR Sampling as compared with PPSWR Sampling.

In our study, the frequency data of distinct words contained in "Hitopadesha" follows Zipf's Law. So, we applied MPPSWR to estimate the population parameter i.e., actual proportion of indeclinable words. As the data follows Zipf's Law, we conclude that MPPSWR gives much better results of the population parameters then PPSWR.

References

Cochran, W.G. (1999): Sampling techniques, *Third Edition*, Wiley, India.

Cunha, C.R., Bestavros, A. and Crovella, M.E. (1995): Characteristics of WWW client-based traces. *Boston University Computer Science Department, Technical Report BUCS-1995-010.*

Gabiax, X. (1999): Zipf's law and the growth of cities. *The American Economic Review*, **89(2)**, 129-132.

Gi-Sung Lee, Daiho Uhm & Jong-Min Kim (2014): Estimation of a rare sensitive attribute in probability proportional to size measures using Poisson distribution, *Statistics*, 48:3, 685-709.

Hill, B. M. (1970): Zipf's law and prior distribution for the composition of a population. *Journal of the American Statistical Association*, **65**, 1220-1232.

Joshi, K. and Rajashi, M. B. (2017): Estimation of actual proportion of loanwords in a language. *Sankhya-B*, **79(I)**, 60-75.

Singh, S. (2003): Use of auxiliary information: probability proportion to size and with replacement (PPSWR) Sampling. *Advanced Sampling Theory with Applications*, 295-348.

Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S. and Ashok,C., (1984): Theory of sample surveys with applications, *Indian society of agriculture statistics*, New Delhi.

Woodroofe, M. and Hill, B. (1975): On Zipf's law. *Journal of applied probability* **12**, 425-434.

Zipf, G. K. (1935): The psycho-biology of language: an introduction to dynamic philology. *Houghton Mifflin*.

Zipf, G. K. (1949): Human behaviour and the principle of least effort. *Cambridge, MA: Addison-Wesley*.

Appendix: Sanskrit frequency lists- Wiktionary, https://en.wiktionary.org/wiki/Appendix:Sanskrit_Word_Frequency_of_Hitopadesha

Authors and Affiliations

Anupama Goyal[1], Pooja Choraria[2], Versha Dwivedi[3] and P C Gupta[4]

Pooja Choraria
Email: pooja.choraria@iisuniv.ac.in

Versha Dwivedi
Email: vershadwivedi21@gmail.com

P C Gupta
Email: pcgupta44@yahho.co.in

[1,3] Department of Statistics, Panjab University, Chandigarh
[2] Department of Statistics, IIS (deemed to be UNIVERSITY)
[4]Retd. Prof. V.N. South Gujarat University, Surat (Gujarat)