DIGITALCOMMONS
@WAYNESTATE

Journal of Modern Applied Statistical Methods

Volume 12 | Issue 1 Article 10

5-1-2013

The Length-Biased Versus Random Sampling for the Binomial and Poisson Events

Makarand V. Ratnaparkhi
Wright State University, makarand.ratnaparkhi@wright.edu

Uttara V. Naik-Nimbalkar
Pune University, Pune, India

Recommended Citation

Ratnaparkhi, Makarand V. and Naik-Nimbalkar, Uttara V. (2013) "The Length-Biased Versus Random Sampling for the Binomial and Poisson Events," *Journal of Modern Applied Statistical Methods*: Vol. 12: Iss. 1, Article 10.
Available at: http://digitalcommons.wayne.edu/jmasm/vol12/iss1/10

Journal of Modern Applied Statistical Methods
May 2013, Vol. 12, No. 1, 54-57

1538 – 9472/13/$95.00

The Length-Biased Versus Random Sampling for the Binomial and Poisson Events

Makarand V. Ratnaparkhi
Wright State University,
Dayton, OH

Uttara V. Naik-Nimbalkar
Pune University,
Pune, India

The equivalence between the length-biased and the random sampling on a non-negative, discrete random variable is established. The length-biased versions of the binomial and Poisson distributions are discussed.

Key words: Length-biased data, weighted distributions, binomial, Poisson, convolutions.

Introduction

The occurrence of so-called length-biased data has been documented by many researchers. No formal mathematical definition of length-biasedness exists; however, if for collected data the probability of inclusion of an observation in a sample is proportional to the magnitude of an observation then the data is referred to as the length-biased data. Further, if the probability of inclusion depends on a certain known function (weight function) of an observation then it is referred to as the size-biased data.

The realization of such data as a sample, as opposed to the realization of a well-defined random sampling procedure, appeared for the first time in a paper by Fisher (1934). In particular, the data collected for estimating the proportion of individuals having certain rare genetic traits such as albinism are length-biased. However, in statistical literature, the term length-biased data was introduced during the 1960s. Fisher (1934) referred to the data collection procedure for counting the number of albino children in a family as the method of ascertainment, which clearly is not a random sampling procedure. For the analysis of the collected data he considered the modifications of binomial distribution (probabilities), which was appropriate at that time. The methodology for such discrete data analysis was formalized by Rao (1965) who introduced the concept of weighted distributions (or in informal nomenclature, the distorted distributions). Rao also considered the weighted distributions for the analysis of various discrete data sets. Recently, Gao, et al. (2011) found that the length-biased Poisson (events) data arise in bioinformatics. There are a number of fields such as ecology, geology, medical science and engineering where the data is length-biased and the researchers have used the weighted distribution for the analysis of such data.

This study follows Ratnaparkhi and Naik-Nimbalkar (2012) regarding the length-biased data arising in oil filed exploration and seeks to associate length-biased data with a random sampling procedure. It is hoped that these efforts will lead to new methods for building appropriate models for length-biased data and for the estimation of the related parameters using such data. Since, the general discussion of length-biasedness (regardless of whether the data is on continuous or discrete random variables is more involved, here, the discussion is limited to the discrete random variables. Further, the binomial and Poisson distributions are commonly used for modeling discrete data. Thus, the derivation of results and the discussion are restricted to these distributions; however, the extensions and modifications to other discrete distributions are possible.

Makarand V. Ratnaparkhi is a Professor of Statistics. Email him at: makarand.ratnaparkhi@wright.edu. Uttara V. Naik-Nimbalkar is a Professor of Statistics. Email her at: uvnaik@stats.unipune.ac.in.

Length-Biased Data versus Random Sample Data: Examples

Example 1: Estimation of Proportion of Families with Albino Children (Fisher, 1934)

An albino child is located and his/her family is included in a sample. Therefore, larger the number of albino children in a family higher is the probability of inclusion of the family in the sample. Let *X*= the number of albino children in a family. Clearly, the data on *X* is length-biased. The binomial probabilities are considered for the analysis if the observed frequencies of the values of *X*.

Example 2: Estimation of the Population of Moose Using Aerial Transect Sampling

Moose, in general, stay and move as groups. The group is located (aerial survey) and the numbers of animals in a group are recorded by a team on the ground. For locating the group, at least one animal from the group must be visible while traveling on a randomly selected transect in the aerial survey of the population. The Poisson distribution is considered for the analysis of these data. Note that larger the number of animals in group, the higher is the probability of sighting at least one animal and hence the length-biasedness in the collected data.

Example 3: Length-Biased Data on RNA Sequences (Gao, et al., 2011)

The quantification of transcripts (RNA-sequences) and related properties, such as the differential expression of the transcripts, arise in microarray data analysis. The differential expression of longer transcripts is more likely to be identified than that of a shorter transcript with the same effect size. This situation is described as length-biasedness in the related data. For the analysis of such data, the Poisson distribution is considered as a model.

Methodology

The derivations of the results for this study are related to the distributions arising in the analysis of length-biased data on discrete random variables. These are based on certain definitions and the probability generating function of the random variables. The derivations are either simple or are the known properties of the distributions of random variables. Therefore, the detailed proofs are omitted for brevity. For future reference, the notations and definitions used herein are listed below.

- *X*: A random variable representing the populations of interest (the original random variable) *X* is assumed to be a non-negative, non-degenerate discrete random variable.
- *Y*, Z_1, Z_2: Other random variables that will be introduced as needed.
- $f_X(x;\theta)$: The pdf of the original distribution of a random variable *X* where θ is a scalar or a vector of parameters.
- $g_X(x;\theta)$: The pdf of the length-biased version of $f_X(x;\theta)$.
- LBS: Length-biased sampling

The pdf of the length-biased version of $f_X(x;\theta)$ is given by

$$g_X(x;\theta) = xf_X(x;\theta)/E[X]\ ,\ E[X] < \infty. \tag{1}$$

For a more general case of (1), *x* is replaced by $w(x) > 0$, and $E[X]$ by $E[w(X)] < \infty$.

- $G_X(t) = E[t^X]$:The probability generating function (pgf) of X. The subscript of $G(t)$ denotes the pgf of the corresponding variable.
- $B(n, p)$: The binomial distribution with parameters *n* and *p*.
- $LBB(m, p)$: The length-biased version of $B(n, p)$.
- $P(\lambda)$: The Poisson distribution with parameter λ.
- $LBP(\lambda)$: The length-biased version of the Poisson distribution.

Results

Convolution of the Poisson and Degenerate Random Variables

Suppose that the original distribution of a random variable $X \sim P(\lambda)$ with pdf

$$f_X(x;\lambda) = e^{-\lambda}\lambda^x / x!\,,\quad x = 0,1,2,\ldots,\infty. \tag{2}$$

then using (1), the pdf of the length biased Poisson distribution is given by

$$g_X(x;\lambda) = e^{-\lambda}\lambda^{(x-1)} /(x-1)!\,,\quad x = 1,2,\ldots,\infty. \tag{3}$$

Further, if $Z_1 \sim P(\lambda), z_1 = 0,1,2,\ldots,\infty,$ then $G_{Z_1}(t) = \exp(-\lambda(1-t))$. And, if Y has a degenerate distribution at $Y = 1$ with $G_Y(t) = t$, then, $G_{(Z_1+Y)}(t) = t\exp(-\lambda(1-t))$, which is the pgf of (3). Thus, the length-biased version of the Poisson distribution is a convolution of the Poisson random variable and the degenerate distribution of *Y*. Further, this result shows that the length-biased version of *X* has the same distribution as that of (Z_1 + *Y*).

This observation, interpreted in terms of sampling, implies that the random sampling on the original Poisson variable denoted by, Z_1, and the degenerate random variable *Y* is equivalent to the so-called *LBS* on random variable *X*.

It is known that the chance mechanism, namely, the well-defined Poisson process, leads to the pdf given by (2) and a well-defined *P*(λ) exists as a corresponding probability model. However, no known stochastic process can be associated that will lead to the pdf (3), the so-called the *LBP*(λ) distribution; therefore, in a strict sense the pdf in (3) is an artificial mathematical construct if the above convolution result is not associated for deriving (3). The related implication is: because Z_1 and *Y* are random variables, the random sampling on Z_1 and *Y* is well defined and, as a result, the random sampling on a length-biased random variable *X* is in order. Thus, the data on *X* is a realization of a random sample on *X*= (Z_1 + *Y*) and there seems to be no need to refer to the data as a length-biased sample. Regarding the theory of statistics, this provides a theoretical basis for length-biased data as a random sample. Note that, in practice, there is a need for providing a practical justification for introducing variable *Y* and its interpretation; however, such interpretation must come from the description of the research problem form where such data is arising in practice.

For practical purposes, if it is known that while sampling for the original random variable *X* the probability of inclusion of an observation is proportional to the magnitude of the observation, a researcher should define the random variables Z_1 and *Y* appropriately and then collect sample data on these variables independently, for example, take a sample first on *Y* followed by a sample on Z_1 and then define $X = Z_1 + Y$ for practical purposes. It is likely that such a sampling plan is more involved in terms of its execution and planning as compared to the practice of collecting the data on *X* and calling it a length-biased variable without any justification with reference to the theory of probability. But, it provides a theoretical base for the collected data and a general theoretical justification to the statistical inference based of these data.

Amari (1985) introduced the role of differential geometrical properties of statistical manifolds, which are well defined objects, in estimation theory. Clearly, the manifolds of the family of distributions given by (2) and (3) are not the same. Therefore, due care is needed when using (3) for estimating parameter λ when it is known that it is associated with a different manifold corresponding to the family of distributions given by (2) and not the one associated with distribution (3).

In view of the discussion related to the Poisson distribution, it seems reasonable to consider other discrete distributions for which the above conclusions are applicable. In general, it can be shown that these conclusions are applicable to the commonly used original discrete distributions with support **Ω** = {0, 1, 2, …, ∞}. For example, it was observed that the method based on the convolution of random variables was useful for generating the length-biased version of the original geometric random variable defined on **Ω**.

Convolution of the Binomial and Degenerate Random Variables

As noted, the length-biased binomial distribution also arises as a model for the length-biased data on the original discrete random variable *X*; therefore, it seems reasonable to obtain the results, similar to those obtained for the Poisson distribution, for the binomial case and to further interpret these results in view of applications.

Suppose that the original distribution of a random variable $X \sim B(n,p)$ with pdf:

$$f_X(x;n,p) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x=0,1,2,\ldots n. \tag{4}$$

Using (1) the pdf of the length-biased binomial distribution is given by:

$$g_X(x;n,p) = \binom{n-1}{x-1} p^{x-1} (1-p)^{n-x}, \quad x=1,2,\ldots n. \tag{5}$$

If $Z_2 \sim B(n,p), z_2 = 0,1,2,\ldots,n,$ then $G_{Z_2}(t) = (q+pt)^n$. If Y has a degenerate distribution at Y=1 with $G_Y(t) = t$, then $G_{(Z_2+Y)}(t) = t\,(q+pt)^n$, which is not the pgf of (5). Thus, the length-biased version of the binomial distribution of random variable is not a convolution of a binomial distribution and the degenerate distribution of Y. Thus, the situation is different than for $P(\lambda)$ and $LBP(\lambda)$; further, the length-biased sampling on the original random variable *X* is not equivalent to the random sampling on (Z_2+Y).

It is not possible to associate a natural chance mechanism with the length-biased binomial distribution – either as repetitions of Bernoulli trials or as a convolution of random variables. Therefore, the *LBB* (*m*, *p*) is a mathematical construct lacking statistical significance. Under such a situation, as opposed to using *LBB* distribution it might be necessary to investigate other distributions as a model for the length-biased data on (original) binomial random variable *X*. The lack of convolution property and, hence, the non-availability of corresponding chance mechanism for the *LBB* distribution appears to be due to the finite support for binomial random variable *X*. Therefore, a similar situation will be present in other distributions with finite support. This raises the following question: Can the so-called length-biased binomial distribution be considered as a model for the so-called length-biased binomial data when it is known that the support of the random variable of interest is finite? At this stage, the exact answer is unknown; therefore, it is left to the reader.

Conclusion

It has been shown that, at least for some situations where data are so-called length-biased, the convolution property could be used for providing a more rigorous treatment for modeling such data within the framework of random sampling and thereby justifies the probability models such as a Poisson distribution. Length-biased data on binomial random variables needs further investigation in data analysis rather than using the mathematical constructs such as the weighted binomial distributions. The extension of results obtained in this study is not straightforward; in general, for length-biased data on continuous random variables, for example, survival data arising in medical science. However, it is worth investigating such cases in view of the number of practical situations where the collected data are on continuous random variables.

References

Amari, S. I. (1985). *Differential-geometrical methods in statistics*. New York, NY: Springer-Verlag.

Fisher, R. A. (1934). The effects of the methods of ascertainment upon the estimation of frequencies. *Annals of Eugenics*, *6*, 13-25.

Gao, L., et al. (2011). Length bias correction for RNA-seq gene set analyses. *Bioinformtics*, *27*,662-669.

Rao, C. R. (1965). On discrete distributions arising out of methods of ascertainment. *Sankhya A*, *27*, 311-324.

Ratnaparkhi, M. V., & Naik-Nimbalkar, U. V. (2012). Length-biased lognormal distribution and its application in the analysis of data from the oil field exploration studies. *Journal of Modern Applied Statistical Methods*, *11*(*1*), 255-260.