JP Journal of Biostatistics
Volume 1, Number 2, 2007, Pages 189-195
Published online: November 14, 2007
This paper is available online at http://www.pphmj.com
© 2007 Pushpa Publishing House

A SIMPLE TEST FOR DETECTION OF LENGTH-BIASED SAMPLING

OLCAY AKMAN[1], JINADASA GAMAGE[1], JASON JANNOT[2], STEVEN JULIANO[2], ANDREW THURMAN[1] and DOUGLAS WHITMAN[2]

[1]Department of Mathematics
Illinois State University
Normal, IL 61790, U. S. A.

[2]Department of Biological Sciences
Illinois State University
Normal, IL 61790, U. S. A.

Abstract

When the probability of selecting an individual in a population is proportional to its magnitude, it is called length biased sampling. Length-biased sampling (LBS) situations may occur in biological studies, clinical trials, reliability, queuing models, survival analysis and population studies where a proper sampling frame is absent. In such situations items are sampled at rate proportional to their length so that larger values of the quantity being measured are sampled with higher probabilities. In this note, we present a test to detect the presence of length-biasedness.

1. Introduction

Most biological field experiments involve observers who move along a selected path collecting samples of animals, insects or plants that are detected based on a pre-determined design. This method is commonly

2000 Mathematics Subject Classification: 62N01, 62N05.

Keywords and phrases: exponential, inverse Gaussian, lognormal, gamma, Monte-Carlo simulation.

All but the second authors are supported by NSF Grant DBI-0442412.

Received July 18, 2007

referred as the line-transect method. Line-transect methods have been used for many types of populations, such as bird, mammal, insect, and plant species (Thompson [6]). It is typical in such surveys that individuals that are closer to the observer's path have higher probability of being detected, not because they are abundant, but because their detectability probability is higher near the path. Additionally larger individuals are more likely to be observed thus causing the sampling distribution to be length-biased.

When observations are selected with probability proportional to their *length*, the resulting distribution is called *length-biased*. Statistical analysis based on length-biased samples has been studied in detail since the early 70's. The concept of length-biased sampling was mainly developed by Rao [5] and Zelen and Feinleib [9]. The length-biased distribution occurs naturally for some sampling plans in biometry, wildlife studies, and survival analysis, among others. When dealing with the problem of sampling and selection from a length-biased distribution, the possible bias due to the nature of data-collection process can be utilized to connect the population parameters to that of the sampling distribution. That is, biased sampling is not always detrimental to the process of inference on population parameters. Inference based on a biased sample of a certain size may yield more information than that given by an unbiased sample of the same size, provided that the choice mechanism behind the biased sample is known.

Examples of situations where the usual sample from a population of interest is not available due to the data having unequal probabilities of entering the sample are discussed in Zelen [8]. Among many others, Gupta and Akman [1, 2] examined statistical estimation for lifetime data under length-biased sampling plan, while Oluyede [4] studied the mathematical properties of length-biased experiments.

In this article, we consider a simple test to detect length-biased samples. Navarro et al. [3] studied how to detect biased samples and how to test a specific weighted model for sampling process. Our approach falls along the same lines, however we provide a distribution-free test which uses the sampled information only.

In Section 2, we derive a simple test for detecting presence of length-biasedness in a sample. In Section 3, we present an example using a data set from a biological study involving plasticity for juvenile development of lubber grasshoppers *Romelea Microptera*. In Section 3, we implement Monte Carlo simulations under various distributions to determine the nominal and observed coverage levels.

2. Derivation of the Test for Length-bias in a Sample of Observations

A distribution function G_F defined on $\mathbf{R}^+$ is called a *length-biased* distribution corresponding to a df F (also defined on $\mathbf{R}^+$), if

$$G_F(y) = \mu_F^{-1}\left\{\int_0^y x \cdot dF(x)\right\} \ \forall y \in \mathbf{R}^+,$$

where $\mu_F = \int_0^\infty x \cdot dF(x)$. Note that the Radon-Nikodym derivative of G_F with respect to F is given by

$$\frac{dG(x)}{dF(x)} = \frac{x}{\mu_F}.$$

Thus, the length-biased pdf can be written as

$$g(x) = \frac{x \cdot f(x)}{E(X)}.$$

Now let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with positive support, and let $f(.)$ be the pdf of the distribution to be sampled from. Also, let the corresponding length-biased density be $g(x) = \frac{x \cdot f(x)}{\mu_X}$, where μ_X is the mean of the original distribution. We consider the following procedure for testing

$$H_0 : f(x) = f(x),$$

$$H_1 : f(x) = g(x).$$

We reject the null hypothesis when

$$\lambda = \prod_{i=1}^{n} \frac{g(x_i)}{f(x_i)} > k \quad \text{for some } k > 0. \tag{1}$$

Since $g(x) = xf(x)/\mu_X$, reduces to

$$\lambda = \frac{x_1 \cdots x_n}{\mu_X^n}. \tag{2}$$

The corresponding test statistic is $T = X_1 \cdots X_n$ and the observed value of the test statistics is $t = x_1 \cdots x_n$. We reject the null hypothesis when the p-value $p = P(T > t)$ is less than the level of significance specified by the researcher. The p-value can be easily calculated using Monte Carlo simulation as follows: Simulate m samples from the distribution specified in the null hypothesis, namely the one with pdf $f_0(x)$ and compute the test statistic for each of these samples. Then the p-value will be

$$p = \frac{\text{Number of simulated } T > t}{m}. \tag{3}$$

In general, for any weight function $w(x)$, that is, $g(x) = w(x)f(x)/\mu_{w(X)}$ we construct the test statistic along the same lines as

$$\lambda = \frac{w(x_1) \cdots w(x_n)}{\mu_{w(X)}^n}. \tag{4}$$

Sampling distributions that are subjected to weight functions other than length-biased can also be seen in real life applications. For instance in wildlife sampling, studies that involve capture-recapture procedures may be prone to size-biased samples where the weight function is $w(x) = x^2$.

3. An Application

Most biological field experiments involve observers who move along a selected path collecting samples of animals, insects or plants that are detected based on a pre-determined design. Figure 1 depicts a typical line-transect sampling, where the observations collected are subject to size-biased sampling.

In this section, we consider an experiment conducted by Whitman and Thurman [7], where population of grasshoppers were first measured and then left in an isolated natural area for some time. A few days later, a different set of observers collected a sample using line-transect methods, and measured the grasshoppers in the sample. After comparison of the results with the known distribution of the original population, it was ascertained that this sampling procedure was length-biased and larger grasshoppers were over-represented in the sample.

Figure 2 contains the distribution of the true population as well as the samples observed in the experiment. It can clearly be seen that the larger grasshoppers are represented at a higher rate than they actually exist in the population. We implemented the test statistic given in (4) to detect possible length-biasedness by first drawing 1000 random samples of size 75, which is the size of the sample originally collected by biologists and by computing the test statistic from each of these samples as given in (3). We then compared the observed value of the test statistic with 1000 simulated ones to obtain the probability of type I error. In our example we observed this value as 0.01 indicating the presence of length-bias in the sample (under 5% level of significance).

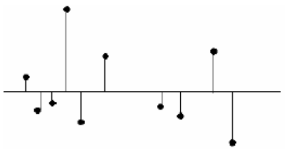

Figure 1. Line-transect sampling.

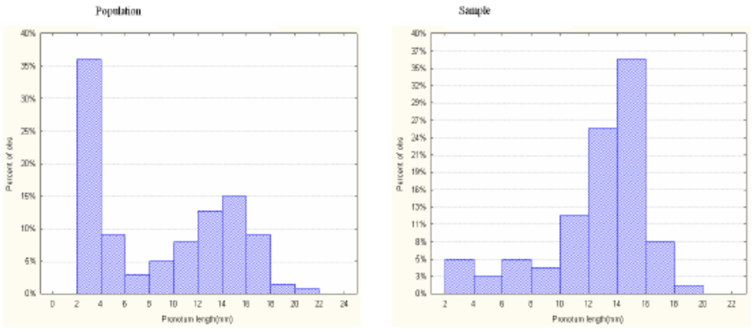

Figure 2. Population and sample measurements.

3.1. Simulation studies

In this section, we implement Monte Carlo simulations from various distributions using different weight functions to obtain the test performance. We simulate 10,000 samples from both original and biased (weighted) versions of the common distributions and note the percentage of the times that the test correctly identifies presence of bias in the sample. The following table contains these values. For instance, when the simulated sample is drawn from the exponential (original) distribution, we correctly identify it at 97% of the time, and when the sample is contaminated with length-biasedness $(g(x) = x)$, we correctly detect the presence of bias at 98% of the times, and so on.

Table 1. Observed bias detection level

	Analysis			
Data Sets (distribution type)	Exponential	Lognormal	Gamma	Inv. Gaussian
Original	.97	.94	.82	.96
$g(x) = x$ (length-biased)	.98	.96	.81	.92
$g(x) = x^{1/2}$	.94	.92	.83	.95
$g(x) = \exp(x)$	.92	.94	.84	.92

It may be important to note that the length-biased gamma distribution is again gamma distribution with only a slight parameter shift. Therefore both original and length-biased gamma densities have substantial overlap that makes it difficult for our (or any) test to identify them. This is why the correct coverage levels corresponding to gamma distribution are lower than that of other distributions. Having said that, the results of our Monte Carlo simulation for the other common lifetime distributions indicate that this simple test is able to detect presence of sampling-originated bias reasonably well and may be used as the first step in checking whether or not proper sampling procedures were employed.

References

[1] R. C. Gupta and O. Akman, On the reliability studies of a weighted inverse Gaussian model, J. Stat. Plann. Inference 48 (1995), 68-93.

[2] R. C. Gupta and O. Akman, Statistical inference based on the length-biased data for the inverse Gaussian distribution, Statistics 31 (1998), 325-337.

[3] J. Navarro et al., How to detect biased samples?, Biometrical J. 45(1) (2003), 91-112.

[4] Broderick O. Oluyede, On inequalities and selection of experiments for length-biased distributions, Probab. Engrg. Inform. Sci. 13 (1999), 169-185.

[5] C. R. Rao, A natural example of a weighted distribution, Amer. Stat. 31 (1977), 24-26.

[6] S. K. Thompson, Sampling, 2nd ed., Wiley, NY, 2002.

[7] D. Whitman and A. Thurman, Sampling Experiment with Romelea Microptera, Normal, IL, 2006.

[8] M. Zelen, Length-biased sampling and biomedical problems, Biometric Society Meeting, Dallas, Texas, 1974.

[9] M. Zelen and M. Feinleib, On the theory of screening for chronic diseases, Biometrika 56 (1969), 601-614.

■