BAUET JOURNAL
Published by
Bangladesh Army University of Engineering & Technology
Journal Homepage: http://journal.bauet.ac.bd
https://doi.org/10.59321/BAUETJ.V4I1.1

Bias Correction of Sample Proportion for Length or Size Biased Data

Noora N Saleh[1], Keya Rani Das[2] and A.H.M. Rahmatullah Imon[3*]

[1]Department of Mathematics, College of Education, Al-Qadisiyah University, Al Di-waniyah, Iraq
[2]Department of Statistics, Associate Professor, Bangabandhu Sheikh Mujibur Rahman Agricultural University, Salna, Gazipur 1706, Bangladesh
[3]Department of Mathematical Sciences, Professor, Ball State University, Muncie, IN 47306, USA

Abstract: In many areas of statistics it is not easy to collect samples which are entirely random in nature. When we collect data from nature by encounter and/or grab sampling we often observe that the collected data are length and/or size biased. The main limitation of this type of data is that they are biased. If we can estimate the bias factor then we can do bias correction and after that the resulting estimates are considered as almost equally reliable to their unbiased counterparts. Bias correction of mean for length and/or size biased data is available in the literature. This is of course a great development but often we are more interested to estimate the proportion rather than the mean. For example, we may be more interested to know the proportion of healthy plants in a garden or the proportion of products which comply the standard. In this paper we develop a bias corrected estimate of proportion for length and/or size biased data. We study few of their mathematical properties and its application in the real world situation.

Keywords: *Encounter sampling, Length bias, Size bias, Bias factor, Bias correction.*

Introduction: In statistics we want to know the population characteristics from the observed samples. The optimum methodology is to obtain a random sample so that the resulting estimates become unbiased. However, in practice it is really difficult to maintain the requirements for a probability sampling. Nonrandom data are more prevalent when we collect data from nature. For convenience researchers often collect data which are not entirely random. An excellent review of these nonrandom sampling techniques is available in [1] and [4]. Among them the encounter sampling has become very popular with the practitioners. We briefly discuss this sampling technique in section 2. The immediate consequence of adapting the encounter sampling is that the collected samples become length and/or size biased and the resulting estimators of parameters will be no longer unbiased. To combat this problem a number of bias correction methods have been suggested in the literature [1, 2, 5, 6, 7]. A bias correction technique for mean is suggested in the literature (see [1]) which is also briefly discussed in this section. So far as we know no bias corrected estimator of proportion exists in the literature. But we really need to know the correct estimate of the proportion because a satisfactory mean count does not necessarily guarantee that the observed units are individually flawless. In section 3, we develop a new estimate of bias corrected proportion. We study a couple of mathematical properties of this estimator as well. The application of the proposed method in a real life data is illustrated in section 4.

Encountered Data and Bias Correction : A large portion of data collected from the nature are encountered data. When an investigator goes into the field, observes and records what he/she encounters is called encountered data. The standard data collection techniques need observations to be taken at random from a prescribed frame, which is often not possible with environmental problems– we have instead to take whatever arises. For example, data are available when the sites are visited. Cyclone, tornado data are recorded when they occur. Another very important factor is accessibility to the data.

One immediate consequence that we may face in encounter sampling is the data become length or size biased. If we sample fish in a pond by catching them in a net, it is highly likely that smaller fishes will slip through the net and bigger fishes will be caught more. If we were to sample plants for their growth by line-intercept method (one type of encounter sampling), the similar problem may arise. This method is generally known as transect sampling method. In this case our data would consist of the lengths of plants crossed by the intercept line and hence the longer plants will have higher probability to be selected resulting length-bias as shown in Figure 1.

Article history:
Received 20 February 2023
Received in revised form 25 May 2023
Accepted 28 October 2023
Available online 15 November, 2023

Corresponding author details: A.H.M. Rahmatullah Imon
E-mail address: rimon@bsu.edu
Tel: +1 765 748 7967

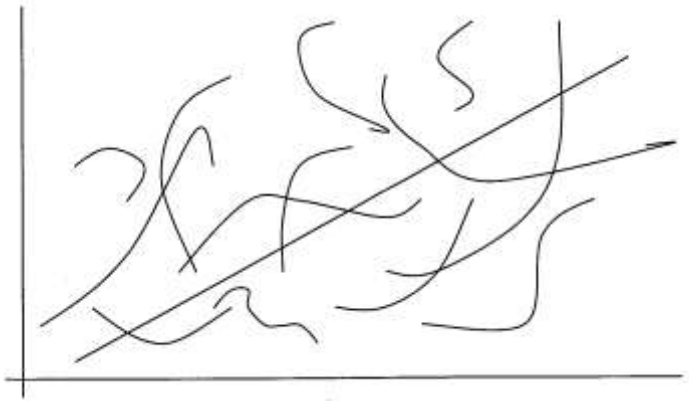

Figure 1: Encounter Sampling for Selecting Fibers

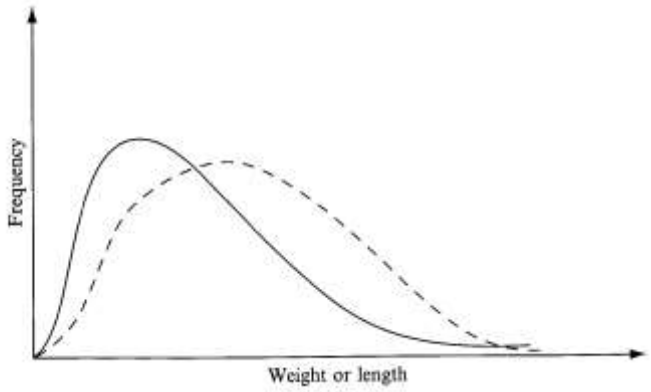

Figure 2: Original and Biased Probability Distributions

Our interest will be in the distribution of sizes, but the sampling methods just described are clearly likely to produce seriously biased results (as shown in Figure 2), It is worth mentioning that both of the figures 1 and 2 are taken from [4]. Here the original (unbiased) distribution is represented by the solid black line, and the biased distribution is represented by the dashed line. Suppose X is nonnegative continuous random variable with mean μ and variance σ^2, but what we actually sample is a random variable X^*. According to Figure 2, $X \geq X^*$ initially but with the increasing values of X things get reversed. For most of the cases, $X^* \geq X$ and thus we will typically overestimate the mean both in the fish and in the fiber examples, possibly to a serious extent.

Barnett [1] suggests a method of finding the amount of bias from the contaminated distribution. A special but popular case of the size-biased distribution has the p.d.f.

$$f^*(x) = \frac{xf(x)}{\mu} \tag{1}$$

The variable actually sampled has expected value

$$E(X^*) = \int \left[\frac{x^2 f(x)}{\mu}\right] dx = \mu\left(1 + \frac{\sigma^2}{\mu^2}\right) \tag{2}$$

The amount

$$\text{BF} = \left(1 + \frac{\sigma^2}{\mu^2}\right) \tag{3}$$

is known as the bias factor. It means if we take a random sample of size n, then the sample mean of the observed data $\overline{x}^*$ is biased upward by a factor $\left(1 + \frac{\sigma^2}{\mu^2}\right)$. Here the problem is that we do not know the true values of μ and σ^2. However, Barnett [1] proposed that the statistic

$$\text{EBF}\ (\overline{x}^*) = \overline{x}^* \frac{\sum_{i=1}^{n} \frac{1}{x_i}}{n} \tag{4}$$

provides an intuitively appealing estimate of the bias factor $\left(1+\frac{\sigma^2}{\mu^2}\right)$. Thus the bias corrected estimate of μ is given by

$$\overline{x} = \frac{\overline{x}^*}{\text{EBF}\,(\overline{x}^*)} \tag{5}$$

Bias Correction for Sample Proportion: In the previous section we have seen how to correct bias in a sample mean. But so far as we know there exists no bias correction method when we use sample proportion as an estimate of population proportion. One may argue that we can treat a sample proportion as a sample mean which takes only values 0 and 1. Definitely we can do that, but we cannot do the bias correction in the same way here as we see in (4) for the values 0, we need to compute 1/0, which is undefined. So we need a different bias correction technique for a sample proportion. Here we propose a new method for bias correction which is described in Theorem 1.

Theorem 1: Suppose X^* is a biased Bernoulli random variable with the probability of success p^*, while X is the bias corrected Bernoulli random variable with the probability of success p and this bias is defined in (1). If $\widehat{p}^* = \overline{x}^*$ is an estimate of p^* then for n independently and identically distributed (iid) observations the bias corrected estimate of p is

$$\widehat{p} = \frac{n\overline{x}^*-1}{n-1} \tag{6}$$

Proof: For Bernoulli random variable X, $\sum_{i=1}^{n} X_i$ is binomial with $\mu = np$ and $\sigma^2 = np(1-p) = \left(1-\frac{\mu}{n}\right)$.

Thus we obtain from (2)

$$\mu^* = \mu\left[1+\frac{\left(1-\frac{\mu}{n}\right)}{\mu^2}\right] = 1+\mu\left(\frac{n-1}{n}\right) \tag{7}$$

$$\Rightarrow \mu = \frac{n}{n-1}(\mu^*-1) \tag{8}$$

$$\Rightarrow p = \frac{np^*-1}{n-1} \tag{9}$$

If we estimate p^* by $\widehat{p}^* = \overline{x}^*$ then the bias corrected estimate of p is

$$\widehat{p} = \frac{n\overline{x}^*-1}{n-1}$$

and that completes the proof of Theorem 1.

Next we study some properties of the proposed bias-corrected estimator of proportion.

Corollary 1: For n iid observations $\widehat{p} = \frac{n\overline{x}^*-1}{n-1}$ is an unbiased estimate of p.

Proof: $E(\widehat{p}) = \frac{nE(\overline{x}^*)-1}{n-1} = \frac{\mu^*-1}{n-1} = \frac{\mu\left(\frac{n-1}{n}\right)}{n-1} = \frac{\mu}{n} = p$.

Corollary 2: For n iid observations $V(\widehat{p}) = \frac{n\sigma^{*2}}{(n-1)^2}$ where $\sigma^{*2} = V(X^*)$.

Proof: This proof is straight forward from the equation (6).

$$V(\widehat{p}) = V\left(\frac{n\overline{x}^*-1}{n-1}\right) = \frac{n^2V(\overline{x}^*)}{(n-1)^2} = \frac{n^2\frac{V(X^*)}{n}}{(n-1)^2} = \frac{n\sigma^{*2}}{(n-1)^2}$$

4 Example

Here we consider an example taken from [5] to see the usefulness of bias corrections. Table 1 presents lengths of 20 peas plants (pisum satiuvum) which are collected from Bangabandhu Sheikh Mujibur Rahman Agricultural University, Gazipur, Bangladesh using the transect sampling method as described in Figure 1.

Table 1: Lengths of Peas Plants.

Plant Length in cm (X^*)	$1/X^*$	Binomial version of X^*
40.1	0.0249377	1
45.2	0.0221239	1
40.1	0.0249377	1
88.2	0.0113379	1
77.3	0.0129366	1
74.1	0.0134953	1
71.1	0.0140647	1
74.5	0.0134228	1
31.4	0.0318471	0
35.0	0.0285714	0
35.2	0.0284091	0
35.4	0.0282486	0
66.2	0.0151057	1
65.2	0.0153374	1
54.2	0.0184502	1
45.2	0.0221239	1
45.2	0.0221239	1
56.3	0.0177620	1
98.0	0.0102041	1
99.2	0.0100806	1

At first we would like to check whether there is any bias for estimating the average length of peas plants. From this data we get $\overline{x}^*$ = 58.86 cm. The bias factor for the mean as defined in (4) is 1.1348 which looks substantially higher than 1. This makes perfect sense. Since the transect sampling is used here for selecting the plants, it is highly likely that relatively shorter plants have lower chance to be selected yielding a higher estimate of length of peas plants. The bias corrected mean as defined in (5) gives the average length of plants as 51.87 cm which is about 7 cm less than what we found before.

Another interesting feature of this data is that according to botanists on average 50 cm length of plants is acceptable growth but if it less than 40 cm they have a great concern about the health of plants. Since 16 out of 40 plants have lengths more than 40 cm, the estimated proportion of healthy plants in 0.80. But when we employ the bias corrected estimate of proportion as given in (6), this becomes 0.7895. This correction makes sense too. Since longer plants had higher chance to be selected, in the observed data it is likely that the proportion of lengthy plants will be higher than the true value as it is seen here as well.

5 Conclusions

In this paper the main objective was to propose a bias corrected estimator of proportion. We develop this new estimator and study few of its mathematical properties. We also present an example to demonstrate how this proposed method can be employed in the real world situation.

Acknowledgements

The authors express their thanks and gratitude to the anonymous reviewers for giving some useful suggestions that led to considerable improvement in the methodology and presentation of the results.

References

[1] V. Barnett. Environmental Statistics: Theory and Methods, Wiley, New York, 2004.
[2] P, Berg, H. Feldmann, H. J. Panitz. Bias correction of high resolution regional climate model data, Journal of Hydrology, 448-449 (2012) 80-92.
[3] J. Cai, L Haan, C. Zhou, Bias correction in extreme value statistics with index around zero, Extremes, 16 (2013), 173-201.
[4] A.H.M.R Imon. Introduction to Environmental Statistics, Nandita Prokash, Dhaka, 2016.
[5] A.H.M.R. Imon, K. Das. Analyzing length or size based data: A study on the lengths of pease plants, Malaysian Journal of Mathematical Sciences, 9 (2015) 1–20.
[6] F. Johnson, A. Sharma. What are the impacts of bias correction on future drought projections? Journal of Hydrology, 525 (2015) 472-485.
[7] H. Lin, S. D. Peddada. Analysis of compositions of microbiomes with bias, Nature Communications, 11 (2020) Article number: 3514.