Journal of Statistical Planning and Inference 157–158 (2015) 54–76

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Rigorous error control methods for estimating means of bounded random variables

Zhengjia Chen[a,*], Xinjia Chen[b]

[a] Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322, United States

[b] Department of Electrical Engineering, Southern University at Baton Rouge, LA 70813, United States

ARTICLE INFO

Article history:
Received 22 April 2013
Received in revised form 27 August 2014
Accepted 28 August 2014
Available online 6 September 2014

Keywords:
Rigorous error control method
Mean estimation
Bounded random variable

ABSTRACT

In this article, we propose rigorous sample size methods for estimating the means of random variables, which require no information of the underlying distributions except that the random variables are known to be bounded in a certain interval. Our sample size methods can be applied without assuming that the samples are identical and independent. Moreover, our sample size methods involve no approximation. We demonstrate that the sample complexity can be significantly reduced by using a mixed error criterion. We derive explicit sample size formulae to ensure the statistical accuracy of estimation.

1. Introduction

Many problems of engineering and sciences boil down to estimating the mean value of a random variable (Mitzenmacher and Upfal, 2005; Motwani and Raghavan, 1995). More formally, let X be a random variable with mean μ. It is a frequent problem to estimate μ based on samples $X_1, X_2, \dots, X_n$ of X, which are defined on a probability space $(\Omega, \mathscr{F}, \mathbb{P}_\mu)$, where the subscript in the probability measure $\mathbb{P}_\mu$ indicates its association with μ. In many situations, the information on the distribution of X is not available except that X is known to be bounded in some interval $[a, b]$. For example, in clinical trials, many quantities under investigation are bounded random variables, such as biomarker, EGFR, K-Ras, B-Raf, Akt, etc. (see., e.g., Arellano et al., 2012; Janik et al., 2010; Wang et al., 2012, and the references therein). Moreover, the samples $X_1, X_2, \dots, X_n$ may not be identical and independent (i.i.d). This gives rise to the significance of estimating μ under the assumption that

$$a \leq X_k \leq b \quad \text{almost surely for } k \in \mathbb{N}, \tag{1}$$

$$\mathbb{E}[X_k \mid \mathscr{F}_{k-1}] = \mu \quad \text{almost surely for } k \in \mathbb{N}, \tag{2}$$

where $\mathbb{N}$ denotes the set of positive integers, and $\{\mathscr{F}_k,\ k = 0, 1, \dots, \infty\}$ is a sequence of σ-subalgebras such that $\{\emptyset, \Omega\} = \mathscr{F}_0 \subset \mathscr{F}_1 \subset \mathscr{F}_2 \subset \dots \subset \mathscr{F}$, with $\mathscr{F}_k$ being generated by $X_1, \dots, X_k$. The motivation we propose to consider the estimation of μ under dependency assumption (2) is twofold. First, from a theoretical point of view, we want the results to hold under the most general conditions. Clearly, (2) is satisfied in the special case that $X_1, X_2, \dots$ are i.i.d. Second, from a practical standpoint, we want to weaken the independency assumption for more applications. For example, in the Monte Carlo estimation technique based on adaptive importance sampling, the samples $X_1, X_2, \dots$ are not necessarily independent. However, as demonstrated in page 6 of Gajek et al. (2013), it may be shown that the samples satisfy (2). An example of adaptive importance sampling is given in Section 5.8 of Fishman (1996) on the study of catastrophic failure.

[*] Corresponding author.
E-mail addresses: zchen38@emory.edu (Z. Chen), xinjia_chen@subr.edu (X. Chen).

http://dx.doi.org/10.1016/j.jspi.2014.08.007
0378-3758/© 2014 Elsevier B.V. All rights reserved.

An unbiased estimator for μ can be taken as

$$\overline{X}_n = \frac{\sum_{i=1}^{n} X_i}{n}.$$

Let $\varepsilon \in (0, 1)$ and $\delta \in (0, 1)$ be pre-specified margin of absolute error and confidence parameter, respectively. Since the probability distributions of $X_1, X_2, \ldots$ are usually unknown, one would use an absolute error criterion and seek the sample size, n, as small as possible such that for all values of μ,

$$\mathbb{P}_\mu \left\{ |\overline{X}_n - \mu| < \varepsilon \right\} > 1 - \delta \tag{3}$$

holds for all distributions having common mean μ. It should be noted that it is difficult to specify a margin of absolute error ε, without causing undue conservatism, for controlling the accuracy of estimation if the underlying mean value μ can vary in a wide range. To achieve acceptable accuracy, it is necessary to choose small ε for small μ. However, this leads to unnecessarily large sample sizes for large μ.

In addition to the absolute error criterion, a relative error criterion is frequently used for the purpose of error control. Let $\eta \in (0, 1)$ and $\delta \in (0, 1)$ be the pre-specified margin of relative error and confidence parameter, respectively. It is desirable to determine the sample size, n, as small as possible such that for all values of μ,

$$\mathbb{P}_\mu \left\{ \left| \overline{X}_n - \mu \right| < \eta |\mu| \right\} > 1 - \delta \tag{4}$$

holds for all distributions having common mean μ. Unfortunately, the determination of sample size, n, requires a good lower bound for μ, which is usually not available. Otherwise, the sample size n needs to be very large, or infinity.

To overcome the aforementioned difficulties, a mixed criterion may be useful. The reason is that, from a practical point of view, an estimate can be acceptable if either an absolute criterion or a relative criterion is satisfied. More specifically, let $\varepsilon > 0$, $\eta \in (0, 1)$ and $\delta \in (0, 1)$. To control the reliability of estimation, it is crucial that the sample size n is as small as possible, such that for all values of μ,

$$\mathbb{P}_\mu \left\{ |\overline{X}_n - \mu| < \varepsilon \text{ or } |\overline{X}_n - \mu| < \eta |\mu| \right\} > 1 - \delta \tag{5}$$

holds for all distributions having common mean μ.

In the estimation of parameters, a margin of absolute error is usually chosen to be much smaller than the margin of relative error. For instance, in the estimation of a binomial proportion, a margin of relative error $\eta = 0.1$ may be good enough for most situations, while a margin of absolute error may be expected to be $\varepsilon = 0.001$ or even smaller. In many applications, a practitioner accepting a relative error normally expects a much smaller absolute error, i.e., $\varepsilon \ll \eta$. On the other hand, one accepting an absolute error ε typically tolerates a much larger relative error, i.e., $\eta \gg \varepsilon$. It will be demonstrated that the required sample size can be substantially reduced by using a mixed error criterion.

Given that the measure of precision is chosen, the next task is to determine appropriate sample sizes. A conventional method is to determine the sample size by normal approximation derived from the central limit theorem (Chow et al., 2008; Desu and Raghavarao, 1990). Such an approximation method inevitably leads to unknown statistical error due to the fact that the sample size n must be a finite number (Fishman, 1996; Hampel, 1998). This motivates us to explore rigorous methods for determining sample sizes.

In this paper, we consider the problem of estimating the means of bounded random variables based on a mixed error criterion. The remainder of the paper is organized as follows. In Section 2, we introduce some martingale inequalities. In Section 3, we derive explicit sample size formulae by virtue of concentration inequalities and martingale inequalities. In Section 4, we extend the techniques to the problem of estimating the difference of means of two bounded random variables. Illustrative examples are given in Section 5. Section 6 provides our concluding remarks. Most proofs are given in Appendices.

2. Martingale inequalities

Under assumption (2), it can be readily shown that $\{X_k - \mu\}$ is actually a sequence of martingale differences (see, e.g. Doob, 1953; Willams, 1991, and the references therein). In the sequel, we shall introduce some martingale inequalities which are crucial for the determination of sample sizes to guarantee pre-specified statistical accuracy.

Define function

$$\psi(\varepsilon, \mu) = (\mu + \varepsilon) \ln \left(\frac{\mu + \varepsilon}{\mu} \right) + (1 - \mu - \varepsilon) \ln \left(\frac{1 - \mu - \varepsilon}{1 - \mu} \right)$$

for $0 < \varepsilon < 1 - \mu < 1$. Under the assumption that $0 \le X_k \le 1$ almost surely and (2) holds for all $k \in \mathbb{N}$, Hoeffding (1963) established that

$$\mathbb{P}_\mu \{ \overline{X}_n \ge \mu + \varepsilon \} < \exp(-n\psi(\varepsilon, \mu)) \quad \text{for } 0 < \varepsilon < 1 - \mu. \tag{6}$$

To see that such result is due to Hoeffding, see Theorem 1 and the remarks on page 18, the second paragraph, of his paper (Hoeffding, 1963). For bounds tighter than Hoeffding's inequality, see a recent paper (Bentkus, 2004).