Computational Statistics and Data Analysis 55 (2011) 3183–3196

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Some variants of adaptive sampling procedures and their applications

Raghu Nandan Sengupta[a,*], Angana Sengupta[b]

[a] *Department of Industrial & Management Engineering, Indian Institute of Technology Kanpur, Kanpur - 208 016, India*
[b] *Delhi Public School Kalyanpur Kanpur, Kanpur - 208 017, India*

ARTICLE INFO

Article history:
Received 19 November 2009
Received in revised form 24 May 2011
Accepted 29 May 2011
Available online 12 June 2011

Keywords:
Decision analysis
Normal
Exponential
Gamma
Extreme value
Sequential sampling
Loss functions
Squared error loss
Linear exponential loss
Bounded risk
Simulation
Business applications

ABSTRACT

Sequential analysis as a sampling technique facilitates efficient statistical inference by considering less number of observations in comparison to the fixed sampling method. The optimal stopping rule dictates the sample size and also the statistical inference deduced thereafter. In this research we propose *three* variants of the already existing multistage sampling procedures and name them as (i) Jump and Crawl (JC), (ii) Batch Crawl and Jump (BCJ) and (iii) Batch Jump and Crawl (BJC) sequential sampling methods. We use the (i) *normal*, (ii) *exponential*, (iii) *gamma* and (iv) *extreme value* distributions for the *point estimation* problems under *bounded risk* conditions. We highlight the efficacy of using the right adaptive sampling plan for the *bounded risk* problems for these *four* distributions, considering *two* different loss functions, namely (i) *squared error* loss (SEL) and (ii) *linear exponential* (LINEX) loss functions. Comparison and analysis of our proposed methods with existing sequential sampling techniques is undertaken and the importance of this study is highlighted using extensive theoretical simulation runs.

1. Introduction

Multistage sampling procedure or sequential analysis as a tool was first developed by Wald along with Wolfowitz (Wald and Wolfowitz, 1945) for a more efficient industrial quality control. The same approach was independently developed at the same time by Alan Turing, to test the hypothesis whether different messages coded by German enigma machines should be connected and analyzed together. Wald (1947) developed this branch of statistics using the concept of sequential probability ratio test (SPRT). In course of time different variations of multistage sampling techniques have been developed such as those proposed by Hall (1981, 1983), Liu (1997), Mukhopadhyay (1990), Mukhopadhyay and Solanky (1991), Ray (1957) and Stein (1945). As a sampling technique this method is efficient as it considers less number of observations in comparison to the fixed sampling method. Few good references are, Ghosh et al. (1997), Ghosh and Sen (1991), Govindarajulu (2004), Mukhopadhyay et al. (2004), Mukhopadhyay and de Silva (2009), Mukhopadhyay and Solanky (1994), Schmitz (1993), Siegmund (1985), Wald (1947) and Zacks (2009). In recent times use of sequential analysis as a technique has found wide applications in areas like regime switching models (Carvalhoa and Lopes, 2007; Bollen et al., 2000), study of optimal clinical trials and novel therapies (Cui et al., 2009; Salvan, 1990; Orawo and Christen, 2009;

* Corresponding author. Tel.: +91 512 259 6607; fax: +91 512 259 7553.
E-mail address: raghus@iitk.ac.in (R.N. Sengupta).

0167-9473/$ – see front matter
doi:10.1016/j.csda.2011.05.020

Prado, in press; Todd et al., in press), analysis of financial time series (Bai and Perron, 2002; Alp and Demetrescu, 2010), etc.

In this paper we propose *three* different variants of the existing multistage sampling procedures available in the literature, and name them as (i) Jump and Crawl (JC); (ii) Batch Crawl and Jump (BCJ) and (iii) Batch Jump and Crawl (BJC) sequential sampling methodologies. We consider the *normal*, *exponential*, *gamma* and *extreme value* distributions separately to show the advantage of using the right adaptive sampling plan. We solve the *bounded risk* point estimation problems for (i) μ, i.e., the *location* parameter, (ii) λ, which is again the *location* parameter, (iii) α, the *shape* parameter and finally (iv) $E(X) = \mu + \gamma_E\sigma$, (where μ and σ are the *location* and *scale* parameters respectively, while γ_E is Euler's constant) for the (i) *normal*, (ii) *exponential*, (iii) *gamma* and (iv) *extreme value* distributions respectively. The *two* different loss functions used for the *bounded risk* problem formulations are the *squared error* loss (SEL) and the *linear exponential* (LINEX) loss functions. To solve these *bounded risk* point estimation problems we use our proposed sequential sampling plans, under some specified constraints for these distributions. Finally we test the validity of our models using extensive simulation runs.

The paper is organized as follows. In Section 2 we describe the concepts of a loss function and *bounded risk* along with the brief descriptions of few existing sequential sampling schemes. Our proposed models are then discussed in Section 3, while Section 4 deals with the background for simulation runs conducted for the *four* distributions. Section 5 gives the detailed analysis related to all the *four* sets of simulations and finally we conclude this paper with our comments in Section 6.

2. Loss functions, bounded risk, sequential sampling methodologies

2.1. Loss functions

In point estimation problem our aim is to find an estimator, T_n, based on a collection of sample of observations, $X_1, X_2, \ldots, X_n$, to estimate the parameter, θ. T_n should be unbiased as well consistent, but this may be unlikely because of the randomness of the sample. Imposition of consistency and unbiasedness may not always lead to a unique estimate. To overcome this problem we utilize the idea of a non-negative metric, called the *loss function*, $L(T_n, \theta)$, where, $L(T_n, \theta) = f(\Delta)$ is usually a function of $\Delta = (T_n - \theta)$, which is the error of estimation. The accuracy of the loss function is measured by its corresponding *risk* function, $R(T_n, \theta) = E[L(T_n, \theta)]$ and our main concern is to minimize this *risk* by a proper choice of the estimator, T_n. Ideally we try to find the optimal estimator, T_n^*, for which the risk, $R\left(T_n^*, \theta\right) \leq R\left(T_n, \theta\right)\ \forall\theta \in \Theta$. The optimal estimator thus obtained is called the *minimum risk estimator* (MRE) with respect to that *particular* loss function only. Practically for a particular loss function, we may not be able to find, T_n^*, as the value of $R(T_n, \theta)$ usually depends on the sample size, n, and the parameter, θ.

Thus the general risk function may be expressed as $[L(T_n, \theta) + c(n)]$, which is a sum of *two* terms. The first term, $L(T_n, \theta)$, is non-increasing in nature with respect to n. While the second term, which is the cost of sampling is given by $c(n)$ and depends on the sample size, n. Furthermore this second term is a non-decreasing function with respect to the sample size, n. Hence it may become difficult to find the optimal value of n, such that the *risk*, $E[L(T_n, \theta) + c(n)]$, is minimized or made as low as possible. Before we discuss the method to find the optimal estimate corresponding to the minimum risk, we give examples of different types of loss functions available in the literature.

Loss functions are of various types, few examples of it being the (i) squared error loss (SEL) function, (ii) weighted squared loss function, (iii) linear loss function, (iv) non-uniform linear loss function, (v) 0-1 loss function, (vi) balanced loss function (BLF), Zellner (1994), (vii) squared exponential loss function and (ix) linear exponential (LINEX) loss function, Zellner (1986). Without going into the detail of each we discuss about the SEL and LINEX loss functions only, as our analysis for all the *four* distributions is based on the risks associated with these *two* loss functions.

Squared error loss (SEL) loss function: The SEL is of the form $L(T_n, \theta) = (T_n - \theta)^2$ and is the most widely used loss function. It is used in estimation problems when unbiased estimators of θ are considered, since the risk, $R(T_n, \theta) = E[L(T_n, \theta)] = E[(T_n - \theta)^2]$ is the *mean square error* (MSE) of T_n, which reduces to the variance of T_n subject to unbiasedness. The corresponding optimal estimator, if it exists, is called the *minimum variance unbiased* (MVU) estimator. It must be remembered that the weighted squared error loss $[w(\theta)(\theta - T_n)^2]$ is a variant of the SEL, where the choice of $w(\theta)$, the weight, depends on the specific value of θ, where $\theta \in \Theta$, (Θ being the parameter space).

Linear exponential loss function: An important point which SEL ignores is the fact that overestimation and underestimation of θ may be of unequal importance in many situations. A loss function, which takes care of this is the linear exponential loss function (LINEX), (Zellner, 1986), which is an asymmetric convex loss function given by $L\left(T_n, \theta\right) = b\left[e^{a(T_n - \theta)} - a\left(T_n - \theta\right) - 1\right]$, where a and b are the *shape* and *scale* parameters respectively. One can easily see that for $a > 0$, the convex loss increases almost exponential (linearly) for positive (negative) values of error, Δ. Therefore, overestimation is of more serious concern than underestimation, while for $a < 0$, the trend is just the opposite. It is quite interesting to note that as $a \to 0$, $L\left(\Delta\right) = b\left[1 + \frac{a\Delta}{1!} + \frac{a^2\Delta^2}{2!} + \cdots - a\Delta - 1\right] \approx b\left[\frac{a^2}{2}\Delta^2\right] + o\left(a^2\right)$, i.e., the LINEX loss reduces to the SEL function for *small* values of a. The value of b does not affect the property of the loss function, but only scales the nature of the LINEX loss function without affecting its shape.