COMSTA: 6322 Model 3Gsc pp. 1–13 (col. fig: NIL)

ARTICLE IN PRESS

Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Q1 Confidence intervals through sequential Monte Carlo

Q2 Ivair R. Silva *

Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA, USA
Department of Statistics, Federal University of Ouro Preto, Ouro Preto, MG, Brazil

ARTICLE INFO

Article history:
Received 13 March 2016
Received in revised form 30 July 2016
Accepted 31 July 2016
Available online xxxx

Keywords:
Confidence coefficient
Coverage probability
Hypothesis testing
Scan test

ABSTRACT

Usually, confidence intervals are built through inversion of a hypothesis test. When the analytical shape of the test statistic distribution is unknown, Monte Carlo simulation can be used to construct the interval. In this direction, a sequential Monte Carlo method for interval estimation is introduced. The method produces intervals with guaranteed confidence coefficients. Because in practice one always needs to establish a truncation on the number of simulations, a simple rule of thumb is offered for choosing the number of simulations as a function of desired upper bounds for the coverage probability. As a novelty in the literature, the sequential Monte Carlo method presents equivalence with the conventional Monte Carlo test. In terms of performance, the superiority of the proposed method is illustrated for two different problems, estimation of gamma distribution means, and estimation of population sizes based on mark-recapture sampling. An example of application for real data is offered for relative risk estimation following the circular spatial scan test.

1. Introduction

A well-known method to obtain exact confidence intervals is the inversion of a hypothesis test method (Casella and Berger, 2002, p. 420). For this, let T denote the test statistic, t_0 to denote a realized value of T, and $F_T(t|\theta)$ to represent the probability distribution of T, where θ is a one-dimensional parameter to be inferred. If $F_T(t|\theta)$ is decreasing in θ for fixed t, an exact $100 \times (1-2\alpha)\%$ confidence interval for θ, say $(\hat{\theta}_l, \hat{\theta}_u)$, is obtained by solving $F_T(t_0|\theta = \hat{\theta}_l) = 1-\alpha$ and $F_T(t_0|\theta = \hat{\theta}_u) = \alpha$, with $\alpha \in (0, 0.5)$. The reasoning is similar for the case where $F_T(t|\theta)$ is increasing in θ.

If the analytical shape of $F_T(t|\theta)$ is unknown, or if the solution $(\hat{\theta}_l, \hat{\theta}_u)$ requires intractable calculations, one of the three options can be tried: (i) asymptotic approximations for $F_T(t|\theta)$. Well-established approaches are the Delta method (Casella and Berger, 2002, p. 243) and the asymptotic efficiency of maximum likelihood estimators (Casella and Berger, 2002, p. 472). The validity of the asymptotic approach has to be checked for each problem and, naturally, does not hold when the sample size is small; (ii) if samples of T can somehow be generated for a fixed value of θ, then Monte Carlo (MC) simulation can provide exact confidence intervals for many types of problems. In general, the validity of MC methods depends on certain assumptions about the shape of $F_T(t|\theta)$ (Preacher, 2012; DiCiccio and Efron, 1996; Efron, 1998) and; (iii) if MC methods are infeasible, a nonparametric bootstrap method can provide a good solution (DiCiccio and Efron, 1996; Chernick, 1999). The nonparametric bootstrap confidence interval is not exact, and its performance depends on each problem. The present manuscript introduces a method of the type (ii).

* Correspondence to: Department of Statistics, Federal University of Ouro Preto, Ouro Preto, MG, Brazil.
E-mail address: ivairest@gmail.com.

http://dx.doi.org/10.1016/j.csda.2016.07.017
0167-9473/

Please cite this article in press as: Silva, I.R., Confidence intervals through sequential Monte Carlo. Computational Statistics and Data Analysis (2016), http://dx.doi.org/10.1016/j.csda.2016.07.017

MC confidence intervals can be broadly separated into two different types, the percentile approach, divided in nonparametric and parametric bootstrap (Efron, 1979; Buckland, 1980; Efron, 1981; Buckland, 1982, 1983; Casella and Berger, 2002; Preacher, 2012), and the method based on inverting a hypothesis test.

In the case of the percentile approach, methods of high performance are possible when dedicated to deal with specific problems. This is the case, for example, of the methods developed to infer the indirect effects in mediation analysis (Preacher, 2012). A Markov chain Monte Carlo approach was proposed by Rzhetsky and Morozov (2001) in order to obtain confidence intervals for the substitution-rate variation in proteins. Interesting applications for population size estimation of animals through capture–recapture were used by Buckland (1980) and by Buckland (1982). Tyralis et al. (2013) proposed a MC method to construct confidence intervals for functions of a parameter in the cases where $F_T(t|\theta)$ is known but has a complicated structure. The noise introduced by the Monte Carlo variability was not analytically explored by Tyralis et al. (2013), but a simulation study showed that the method performs well for some members of the exponential distribution family.

Concerning the methods based on inverting a hypothesis test, it merits to cite the proposal of Bølviken and Skovlund (1996). Their method is based on a randomized procedure that requires the creation of an auxiliary statistic with same distribution as T, say T^*, which by its turn has to be computable for each observed $T^* = t_0^*$ and for each θ in the parameter space. If the Monte Carlo variation is ignored (equivalent to $m \to \infty$), the interval $(\hat{\theta}_l^*, \hat{\theta}_u^*)$ is exact. But, creativity is required on the user's part to find T^*, and a general rule to guide in the search for T^* was not provided.

In general, the performance of MC methods is strongly dependent on the validity of specific conditions about the shape of $F_T(t|\theta)$, but conditions are difficult to check in practice because, usually, $F_T(t|\theta)$ is unknown when MC is needed. Aiming to solve this problem, this paper introduces the sequential MC method. The sequential MC method has already been suggested by Silva (2014) in the proceedings of the 8th International Conference on Applied Mathematics, Simulation, Modeling (ASM'14), and here the method is evaluated through further analytical derivations, numerical study of its performance and an application for real data. Sequential MC makes use of the well-known 'bisection method', a numerical procedure extensively used for finding roots of equations (Autar et al., 2011). We can propose the following three important advantages of the sequential MC method in comparison to former methods:

- The sequential MC method has the ability to establish the finite maximum number of Monte Carlo simulations needed to achieve a specific precision on the coverage probability, which only requires feasibility of Monte Carlo simulation.
- Former methods are applicable only when the statistic used for constructing the interval is an estimator of θ. Unlike, with sequential MC one can proceed to get an interval estimate even when the statistic is not an estimator of θ, like the likelihood ratio test statistics, for example.
- As a novelty in the literature, conclusions drawn from the proposed method will always agree with conclusions from the conventional MC hypothesis testing. Such duality is enjoyed when the exact confidence interval is feasible. Now, with the sequential MC method such convenient property is extended for the Monte Carlo approach.

The notation $F_T(t|\theta)$ depends on θ only, but, if Monte Carlo simulation is feasible, all results are valid in the presence of two or more nuisance parameters. As a sufficient condition for validity of the analytical properties deduced in this paper, $F_T(t|\theta)$ is assumed monotone (increasing or decreasing) in θ for each fixed t. But note: monotonicity in θ is the basic condition for having real-line intervals even when exact solution is feasible. Thus, monotonicity is not really a limitation of the sequential MC method, but a condition for defining confidence intervals in general.

This material is organized in the following way: next section offers a brief overview on the main procedures for constructing confidence intervals when exact inference is unpleasant or computationally unfeasible. Section 3 describes the sequential MC method and Section 4 describes the unification of testing and interval estimation followed by Monte Carlo designs. Section 5 presents a simulation study showing that, for estimation of a gamma distribution mean, the sequential MC outperforms conventional methods in terms of accruing the desired coverage probabilities. For the mark-recapture problem, Section 6 shows that, in comparison to one of the prominent former methods, called percentile method, the performance of the sequential MC is superior. Section 7 presents an example of how to use the method to construct confidence intervals for the relative risk associated to a spatial cluster detected by the circular scan test. Section 8 closes the paper with some last comments.

2. Overview of former methods

This section describes three of the prominent resampling/Monte Carlo methods for finding confidence intervals: bootstrap; BC-bootstrap; and the percentile. In addition, a brief introduction about Monte Carlo testing is offered as it shall be necessary later.

Let $\vec{W} = (W_1, \ldots, W_N)$ denote a random sample containing information about a one-dimensional parameter θ, and let $f_W(w|\theta)$ denote the probability density function (or probability function in the discrete case) of W_i, $i = 1, \ldots, N$. Also, let T denote a statistic for θ, and t_0 to denote a realized value of T.

2.1. Bootstrap

The nonparametric percentile, also called nonparametric bootstrap, and here simply called by 'bootstrap', is based on sampling N points from the random sequence $W_1, \ldots, W_N$ B times. Thus, a sequence of $(B - 1)$ bootstrap statistics