Journal of the Korean Statistical Society 45 (2016) 176–187

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

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

Exponentially tilted empirical distribution function for ranked set samples

Saeid Amiri [a,*], Mohammad Jafari Jozani [b], Reza Modarres [c]

[a] Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA, USA
[b] Department of Statistics, University of Manitoba, Winnipeg, MB, Canada, R3T 2N2
[c] Department of Statistics, The George Washington University, Washington DC, USA

ARTICLE INFO

Article history:
Received 8 January 2015
Accepted 17 September 2015
Available online 21 October 2015

AMS 2000 subject classifications:
primary 62F40
secondary 62G30

Keywords:
Distribution function
Exponential tilting
Imperfect ranking
Ranked set sample

ABSTRACT

We study nonparametric estimation of the distribution function (DF) of a continuous random variable based on a ranked set sampling design using the exponentially tilted (ET) empirical likelihood method. We propose ET estimators of the DF and use them to construct new resampling algorithms for unbalanced ranked set samples. We explore the properties of the proposed algorithms. For a hypothesis testing problem about the underlying population mean, we show that the bootstrap tests based on the ET estimators of the DF are asymptotically normal and exhibit a small bias of order $O(n^{-1})$. We illustrate the methods and evaluate the finite sample performance of the algorithms under both perfect and imperfect ranking schemes using a real data set and several Monte Carlo simulation studies. We compare the performance of the test statistics based on the ET estimators with those based on the empirical likelihood estimators.

© 2015 The Korean Statistical Society. Published by Elsevier B.V. All rights reserved.

1. Introduction

Ranked set sampling (RSS) is a powerful and cost-effective data collection technique that is often used to collect more representative samples from the underlying population when a small number of sampling units can be fairly accurately ordered without taking actual measurements on the variable of interest. RSS is most effective when obtaining exact measurement on the variable of interest is very costly, but ranking the sampling units is relatively inexpensive. RSS finds applications in industrial statistics, environmental and ecological studies as well as medical sciences. For recent overviews of the theory and applications of RSS and its variations see Wolfe (2012) and Chen, Bai, and Sinha (2004).

Ranked set samples can be either balanced or unbalanced. An unbalanced ranked set sample (URSS) is one in which the ranked order statistics are not quantified the same number of times. To obtain an URSS of size n from the underlying population we proceed as follows. Let n sets of sampling units, each of size k, be randomly chosen from the population using a simple random sampling (SRS) technique. The units of each set are ranked by any means other than the actual quantification of the variable of interest. Finally, one and only one unit in each ordered set with a pre-specified rank is measured. Let m_r be the number of measurements on units with rank r, $r \in \{1, \ldots, k\}$ such that $n = \sum_{r=1}^{k} m_r$. Suppose $X_{(r)j}$ denotes the measurement on the jth unit with rank r. The resulting URSS of size n from the underlying population is denoted by $\mathbf{X}_{\text{URSS}} = \{\mathcal{X}_1, \ldots, \mathcal{X}_n\}$, where the elements of the rth row $\mathcal{X}_r = (X_{(r)1}, X_{(r)2}, \ldots, X_{(r)m_r})$ are independently

[*] Corresponding author.
E-mail address: saeid.amiri1@gmail.com (S. Amiri).

http://dx.doi.org/10.1016/j.jkss.2015.09.004
1226-3192/© 2015 The Korean Statistical Society. Published by Elsevier B.V. All rights reserved.

and identically distributed (i.i.d.) from $F_{(r)}, r = 1, \ldots, k$ and $F_{(r)}$ is the DF of the rth order statistic. Moreover, $X_{(r)j}$s are independent for $r = 1, \ldots, k$ and $j = 1, \ldots, m_r$. Note that if $m_r = m, r = 1, \ldots, k$, then URSS reduces to the balanced RSS. The DF of URSS is

$$\widehat{F}_{q_n}(t) = \frac{1}{n}\sum_{r=1}^{k}\sum_{j=1}^{m_r} I(X_{(r)j} \leq t) = \sum_{r=1}^{k} q_{m_r}\widehat{F}_{(r)}(t), \tag{1}$$

where $n = \sum m_r$ and $q_{m_r} = m_r/n$. As it is shown in Chen et al. (2004), when $n \longrightarrow \infty$, and $q_{m_r} \longrightarrow q_r$, for $r = 1, \ldots, k$, we have $\widehat{F}_{q_n}(t) \longrightarrow F_q(t)$, where

$$F_q(t) = \sum_{r=1}^{k} q_r F_{(r)}(t). \tag{2}$$

One can easily see that $F_q(t)$ is not equal to the underlying DF $F(t)$, unless $q_r = 1/k, r = 1, \ldots, k$, showing that the EDF based on the URSS data does not provide a good estimate of the underlying distribution F. The properties of the EDF of the balanced and unbalanced RSS are studied in Stokes and Sager (1988) as well as Chen et al. (2004).

In this paper, we use the empirical likelihood method as a nonparametric approach for estimating F. To this end, we propose two methods to estimate F using the exponentially tilted (ET) technique. The proposed estimators can be used as standard tools for practitioners to estimate the standard error of any well-defined statistic based on RSS or URSS data and to make inferences about the characteristics of interest of the underlying population. Another interesting problem in this direction is to develop efficient resampling techniques for URSS data, as in many cases the exact or the asymptotic distribution of the statistics based on URSS data are not available or they are very difficult to obtain (e.g. Chen et al., 2004). Akin to the methods of Modarres, Hui, and Zhang (2006) and Amiri, Jafari Jozani, and Modarres (2014), the new ET estimators of F are used to construct new resampling techniques for URSS data. We study different properties of the proposed algorithms. For a hypothesis testing problem, about the underlying population mean, we show that the bootstrap tests based on the ET estimators are asymptotically normal and exhibit a small bias of order $O(n^{-1})$ which are desirable properties.

The outline of the paper is as follows. In Section 2, we present ET estimators of F based on the URSS data. Section 3 considers two methods for resampling RSS and URSS data based on the ET estimators of F. We provide justifications for validity of these methods for a hypothesis testing problem about the population mean. Section 4 describes a simulation study to compare the finite sampling properties of the proposed methods with parametric bootstrap and some existing resampling techniques for testing a hypothesis about the population mean. We consider both perfect and imperfect ranking scenarios, three different distributions and five RSS designs. We compare the performance of our proposed methods with the one based on the empirical likelihood method studied in Liu, Lin, and Zhang (2009) as well as Baklizi (2009). In Section 5, we apply our methods for a testing hypothesis problem using a real data set consisting of the birth weight and seven-month weight of 224 lambs along with the mother's weight at time of mating. Section 6 provides some concluding remarks.

2. Exponential tilting of DF

Exponential tilting of an empirical likelihood is a powerful technique in nonparametric statistical inference. The impetus of this approach is the use of the estimated DF subject to some constraints rather than the EDF. ET methods find applications in computation of bootstrap tail probabilities (Efron & Tibshirani, 1993), point estimation (Schennach, 2007), estimation of the spatial quantile regression (Kostov, 2012), Bayesian treatment of quantile regression (Schennach, 2005), small area estimation (Chaudhuri & Ghosh, 2011) and Calibration estimation (Kim, 2010), among others.

Let $\mathbf{X} = \{X_1, \ldots, X_n\}$ be a generic sample of size n from F and suppose $F_n(x) = \sum_{i=1}^{n}\frac{1}{n}\mathbb{I}(X_i \leq x)$ is the EDF of $\mathbf{X}$ which places empirical frequencies (weights) $1/n$ on each X_i. Consider an estimator $\widetilde{F}_p(x) = \sum_{i=1}^{n} p_i\mathbb{I}(X_i \leq x)$ of F which assigns weights p_i instead of $1/n$ to each X_i. To obtain the ET estimator of F, we minimize an aggregated distance between the empirical weights $1/n$ and p_i subject to some constraints on the p_i's. More specifically, one chooses a distance $d(\widetilde{F}_p, F_n) = \sum_{i=1}^{n} d(p_i, \frac{1}{n})$ and minimizes $d(\widetilde{F}_p, F_n)$ subject to $\sum_{i=1}^{n} p_i = 1$ and some other constrains such as $g(\mathbf{X}, \theta_0) = \sum_{i=1}^{n} p_i g(X_i, \theta_0) = 0$, using the following Lagrangian multiplier method

$$d(\widetilde{F}_p, F_n) - \lambda g(\mathbf{X}, \theta_0) - \alpha\left(\sum_{i=1}^{n} p_i - 1\right), \tag{3}$$

where $g(\mathbf{X}, \theta_0)$ is often imposed under the null hypothesis in a testing problem or any other conditions that one needs to account for in practice. Note that the minimization in (3) can also be done by minimizing the distance between $\widetilde{F}_p(x)$ and any target estimator $F_{\widehat{p}}(x) = \sum_{i=1}^{n} \widehat{p}_i\mathbb{I}(X_i \leq x)$ other than the EDF $F_n(t)$.

The choice of the discrepancy function $d(\cdot, \cdot)$ for the aggregated loss $d(\widetilde{F}_p, F_n)$ in (3) leads to different ET estimators of F. Since $F_n(x)$ is the nonparametric maximum likelihood estimator of F under the Kullback–Leibler distance subject to the

Download English Version:

https://daneshyari.com/en/article/1144557

Download Persian Version:

https://daneshyari.com/article/1144557

Daneshyari.com