Journal of Statistical Planning and Inference 141 (2011) 809–816

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

A note on ranked-set sampling using a covariate

Jesse Frey

Department of Mathematical Sciences, Villanova University, Villanova, PA 19085, USA

ARTICLE INFO

Article history:
Received 2 October 2009
Received in revised form
17 May 2010
Accepted 4 August 2010
Available online 10 August 2010

Keywords:
Admissibility
Judgment post-stratification
Nonparametric regression

ABSTRACT

Ranked-set sampling (RSS) and judgment post-stratification (JPS) use ranking information to obtain more efficient inference than is possible using simple random sampling. Both methods were developed with subjective, judgment-based rankings in mind, but the idea of ranking using a covariate has received a lot of attention. We provide evidence here that when rankings are done using a covariate, the standard RSS and JPS mean estimators no longer make efficient use of the available information. We first show that when rankings are done using a covariate, the standard nonparametric mean estimators in JPS and unbalanced RSS are inadmissible under squared error loss. We then show that when rankings are done using a covariate, nonparametric regression techniques yield mean estimators that tend to be significantly more efficient than the standard RSS and JPS mean estimators. We conclude that the standard estimators are best reserved for settings where only subjective, judgment-based rankings are available.

1. Introduction

Ranked-set sampling (RSS), proposed by McIntyre (1952, 2005), is a sampling method that improves on simple random sampling (SRS) by taking advantage of auxiliary ranking information. To obtain a balanced ranked-set sample, one begins by specifying a set size m. One then draws m independent simple random samples (sets) of size m and ranks the units in each set from smallest to largest. The ranking may be done either by judgment or by using an easily available covariate, and it need not be perfectly accurate. One then selects for measurement the unit ranked smallest in the first set, the unit ranked second-smallest in the second set, and so on. This process yields a sample of m independent values. To obtain a larger sample, one repeats the process for n independent cycles. The sample then consists of $N \equiv nm$ independent values, with n values from units ranked smallest, n values from units ranked second-smallest, and so on. If the rankings are perfect, these values are independent order statistics from the parent distribution. Otherwise, they are independent judgment order statistics.

In some statistical problems, it is helpful to allow the number of measured units with each rank to vary from one rank to another. In this case, one may use unbalanced RSS. One simply specifies a set size m and a vector $(n_1,\ldots,n_m)$, where $n_i > 0$ is the number of units with in-set rank i to be selected for measurement. The sample then consists of $N \equiv \sum_{i=1}^m n_i$ independent judgment order statistics. If $\overline{Y}_i$ is the sample mean for the measured values from units given rank i, then the standard RSS nonparametric mean estimator is $\overline{Y}_{RSS} = (1/m)\sum_{i=1}^m \overline{Y}_i$, which is unbiased under either balanced or unbalanced RSS.

Another variation on balanced RSS is judgment post-stratification (JPS), proposed by MacEachern et al. (2004). To collect a JPS sample of size N using set size m, one first selects a simple random sample of size N. Each of the N units is measured, and some additional ranking information is also collected. For each of the N measured units, one obtains an additional

E-mail address: jesse.frey@villanova.edu

0378-3758/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2010.08.002

$m-1$ independent units to create a set of size m. One then ranks the m units from smallest to largest, noting the rank of the measured unit. The full data set then consists of the N measured values and their associated ranks. As in unbalanced RSS, the number of units with each rank may vary from one rank to another. However, while the number n_i of measured values with rank i is fixed in advance in RSS, it is random in JPS. In fact, the vector $(n_1,\ldots,n_m)$ has a multinomial distribution with mass parameter N and probability vector $(1/m,\ldots,1/m)$. JPS tends to be somewhat less efficient than balanced RSS, but it offers increased flexibility. One advantage is that since JPS is based on a simple random sample, researchers retain the option of using SRS-based methods. In addition, JPS allows rankers to declare ties (MacEachern et al., 2004). The standard JPS nonparametric mean estimator is $\overline{Y}_{JPS}$, the average of $\overline{Y}_i$ over all ranks i such that $n_i>0$.

Both RSS and JPS were developed with subjective, judgment-based rankings in mind, but the idea of ranking using a covariate has received a lot of attention. For example, Chen et al. (2006) discussed how RSS with covariate-based rankings can be used in estimating a population proportion, and Wang et al. (2006) discussed how one can implement JPS using more than one covariate. We provide evidence here that when rankings are done using a covariate, the estimators $\overline{Y}_{JPS}$ and $\overline{Y}_{RSS}$ no longer make efficient use of the available information. Others such as Ridout and Cobby (1987) and Wang et al. (2008) have noted that when covariates are used in the ranking process, incorporating the covariates into the estimation process should lead to improved estimators. We expand on this observation in two ways. We first show that when covariates are used to do the rankings in JPS and unbalanced RSS, estimators like $\overline{Y}_{JPS}$ and $\overline{Y}_{RSS}$ that do not use the covariate information are actually inadmissible under squared error loss. We then propose some alternate nonparametric estimators that do incorporate the covariate information, and we show that these estimators tend to be more efficient than $\overline{Y}_{JPS}$ and $\overline{Y}_{RSS}$.

In Section 2, we show that when rankings are done using a covariate, $\overline{Y}_{JPS}$ and $\overline{Y}_{RSS}$ are inadmissible under squared error loss. The argument that we use also shows that several other estimators in the literature are inadmissible when the rankings are done using a covariate. Kvam and Samaniego (1993) showed that in balanced RSS, $\overline{Y}_{RSS}$ is inadmissible when sampling from specific parametric families of distributions. Our results here are less general in that they apply only when the ranking is done using a covariate, but more general in that they require neither parametric assumptions nor perfect rankings. In Section 3, we show that when rankings are done using a covariate, nonparametric regression techniques yield estimators that tend to be significantly more efficient than $\overline{Y}_{JPS}$ and $\overline{Y}_{RSS}$. We give our conclusions in Section 4.

2. Inadmissibility results

Suppose that a sample is obtained using either JPS or unbalanced RSS, with the rankings done using a covariate. It then follows that for the N units that are measured, both the variable of interest Y and the ranking variable X are known, while for the $N(m-1)$ units used only for ranking, X is known, but Y is unknown. Using this data, we obtain better estimators through conditioning. Specifically, we condition on (i) the set S_1 of units used for ranking and (ii) the set $S_2\subset S_1$ of units chosen for measurement, and we define $\tilde{Y}_{JPS}=E[\overline{Y}_{JPS}|S_1,S_2]$ and $\tilde{Y}_{RSS}=E[\overline{Y}_{RSS}|S_1,S_2]$. By the identities $E[A]=E[E[A|B]]$ and $V(A)=V(E[A|B])+E[V(A|B)]$ (Casella and Berger, 2002, p. 342), $\tilde{Y}_{JPS}$ and $\tilde{Y}_{RSS}$ have the same expected values as $\overline{Y}_{JPS}$ and $\overline{Y}_{RSS}$, but variances no larger than those of $\overline{Y}_{JPS}$ and $\overline{Y}_{RSS}$. The quantities $\tilde{Y}_{JPS}$ and $\tilde{Y}_{RSS}$ also depend only on the data, as we argue below.

Theorem 1. *The expectations $E[\overline{Y}_{JPS}|S_1,S_2]$ and $E[\overline{Y}_{RSS}|S_1,S_2]$ depend only on the data.*

Proof. Consider the JPS case first. Suppose that we want to compute $E[\overline{Y}_{JPS}|S_1,S_2]$. We are conditioning on the set of Nm units used for ranking and the set of N units chosen for measurement. Thus, the only random element is how the unmeasured units are grouped into sets with the measured units. We obtain $E[\overline{Y}_{JPS}|S_1,S_2]$ by considering all $(Nm-N)!/((m-1)!)^N$ ways to create N sets of size m by assigning $m-1$ unmeasured units to each measured unit. For each of these equally likely assignments, we note the rank associated with the measured value and compute the estimate $\overline{Y}_{JPS}$. Averaging these estimates over all such assignments gives $E[\overline{Y}_{JPS}|S_1,S_2]$.

Now consider the RSS case. The vector of sample sizes $(n_1,\ldots,n_m)$ is given, and we want to compute $E[\overline{Y}_{RSS}|S_1,S_2]$. We are conditioning on the set of Nm units used for ranking and the set of N of units chosen for measurement. We may also assume without loss of generality that we have fixed ahead of time which ranked unit to select from each of the N sets. Thus, the only random element is how the Nm units are split into sets. We obtain $E[\overline{Y}_{RSS}|S_1,S_2]$ by considering all $(Nm)!/(m!)^N$ ways to create the N sets. For each of these equally likely assignments, we note the rank associated with the measured value and whether the units in S_2 are the ones that would be chosen for measurement. For each assignment where the set of units selected for measurement matches the set S_2 on which we are conditioning, we compute $\overline{Y}_{RSS}$. Averaging these estimates gives $E[\overline{Y}_{RSS}|S_1,S_2]$.

Since $\tilde{Y}_{JPS}$ and $\tilde{Y}_{RSS}$ depend only on the data, they are estimators. They have variances as small or smaller than the original estimators, and as long as there is some choice of S_1 and S_2 for which the distribution of $\overline{Y}_{JPS}$ or $\overline{Y}_{RSS}$ given S_1 and S_2 is not degenerate, they improve on the original estimator. For balanced RSS, the distribution of $\overline{Y}_{RSS}$ given S_1 and S_2 is always degenerate, but in any JPS setting with $N>2$ or any unbalanced RSS setting, there are choices of S_1 and S_2 such that the distribution of $\overline{Y}_{JPS}$ or $\overline{Y}_{RSS}$ given S_1 and S_2 is not degenerate. Thus, the new estimators always offer an improvement for JPS with $N>2$ and for unbalanced RSS, but not for balanced RSS. When N and m are small, it is possible to list out all of the possible assignments mentioned in the proof of Theorem 1. When the sample is larger, $\tilde{Y}_{JPS}$ and $\tilde{Y}_{RSS}$ may be approximated