Journal of the Korean Statistical Society 43 (2014) 643–652

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

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

The Student's t approximation to distributions of pivotal statistics from ranked set samples

Soohyun Ahn [a], Johan Lim [a,*], Xinlei Wang [b]

[a] Department of Statistics, Seoul National University, Seoul, Republic of Korea
[b] Department of Statistical Science, Southern Methodist University, Dallas, TX, USA

ARTICLE INFO

Article history:
Received 24 April 2013
Accepted 19 January 2014
Available online 3 February 2014

AMS 2000 subject classifications:
primary 62G30
secondary 62E17

Keywords:
Confidence interval
One sample test
Pivotal statistic
Ranked set sampling
t approximation

ABSTRACT

For data collected using ranked set sampling, pivotal statistics are commonly used in inferential procedures for testing the population mean, and their asymptotic distributions are used as surrogates of the underlying true distributions. However, the sample size of a ranked set sample (RSS) is often small so that the distribution of its pivotal statistic can deviate much from the limiting normality. In this paper, we propose to approximate the distribution by the Student's t distribution, of which the number of degrees of freedom (DF) is estimated from the data. We consider three estimators of the DF, two based on the Welch-type approximation (Welch, 1947) and the third simply given by the difference between the total sample size and the set size of the RSS. We numerically compare the corresponding approximate t distributions with the asymptotic normal distribution via simulation in two aspects: (i) the approximation error; and (ii) the coverage probability. We also apply the proposed Student's t approximation to tree height data in Platt et al. (1988) and Chen et al. (2003). Our results show that all the three approximate t distributions seem to be consistently better than the asymptotic normal distribution for the RSS under both perfect and imperfect ranking. We further give recommendations to practitioners based on the relative performance of the three DF estimators.

1. Introduction

Inference using order information receives much attention in the statistical literature. Ranked set sampling (RSS) is a cost-effective sampling method, which uses the rank (order) information of data to improve the estimation efficiency. It is particularly useful when measuring the characteristic of interest (say y) is expensive or difficult but recruiting sampling units and ranking them are relatively easy and inexpensive.

Like other sampling methods, one major goal in RSS is the inference about the mean μ of a population under study. The inferential procedures for the population mean strongly rely on pivotal statistics, which will be defined below. Suppose that we have RSS data in the form of

$$\{(y_i, r_i, k_i),\ i = 1, 2, \ldots, n\},$$

* Corresponding author.
E-mail address: johanlim@snu.ac.kr (J. Lim).

1226-3192/$ – see front matter
http://dx.doi.org/10.1016/j.jkss.2014.01.004

where the (judgment) rank of y_i among k_i independent observations is denoted by r_i. Generally, we assume $k_i = H$ for every $i = 1, 2, \ldots, n$ (i.e., H is the set size of the ranked set sample) so that $r_i \in \{1, \ldots, H\}$. Here, ranking errors could occur so that ranking can be imperfect. Let $\mathrm{I}(\cdot)$ denote the indicator function. Let $n_h = \sum_{i=1}^{n} \mathrm{I}(r_i = h)$ be the number of units with rank h, and so $n = \sum_{h=1}^{H} n_h$ is the total sample size of the RSS. Let $\widehat{\mu}^{\mathrm{rss}}$ denote the RSS mean estimator, namely

$$\widehat{\mu}^{\mathrm{rss}} = \frac{1}{H} \sum_{h=1}^{H} \bar{y}_{[h]},$$

where $\bar{y}_{[h]}$ is the sample mean of the hth rank stratum:

$$\bar{y}_{[h]} = \frac{\sum_{i=1}^{n} y_i \mathrm{I}(r_i = h)}{n_h}.$$

Then the pivotal statistic based on $\widehat{\mu}^{\mathrm{rss}}$ for testing $\mathcal{H}_0 : \mu = \mu_0$ is given by

$$\mathbf{T}_{\mathrm{p}} = \frac{\widehat{\mu}^{\mathrm{rss}} - \mu_0}{\sqrt{\frac{1}{H^2} \sum_{h=1}^{H} \frac{s^2_{[h]}}{n_h}}},$$

where $s^2_{[h]}$ is the sample variance of the hth rank stratum:

$$s^2_{[h]} = \frac{\sum_{i=1}^{n} \left(y_i - \bar{y}_{[h]}\right)^2 \mathrm{I}(r_i = h)}{n_h - 1}.$$

Here, we assume that there are at least two units measured in each rank stratum so that the in-stratum variance $\sigma^2_{[h]}$ is estimable; that is, for every $h \in \{1, \ldots, H\}$, $n_h > 1$.

The distribution of $\mathbf{T}_{\mathrm{p}}$ plays a key role in many RSS inferential procedures. In the literature, researchers used the standard normal distribution as its distribution (e.g., Chen, 2000; Ozturk, 1999; Sroka, Stasny, & Wolfe, 2006; Wang, Lim, & Stokes, 2008). To test $\mathcal{H}_0 : \mu = \mu_0$, Chen (2000) and Sroka et al. (2006) propose to reject the null hypothesis when $\left|\mathbf{T}_{\mathrm{p}}\right| > z_{\alpha/2}$, where $z_{\alpha/2}$ is $100(1 - \alpha/2)$th percentile of the standard normal distribution. Hence, the confidence interval can be expressed by

$$\widehat{\mu}^{\mathrm{rss}} \pm z_{\alpha/2} \sqrt{\frac{1}{H^2} \sum_{h=1}^{H} \frac{s^2_{[h]}}{n_h}}.$$

However, in practical situations, especially those requiring cost efficiency, it is typical that the sample size of a RSS is not large enough to support the normality of $\mathbf{T}_{\mathrm{p}}$ well.

In this paper, we approximate the distribution of $\mathbf{T}_{\mathrm{p}}$ by a Student's t distribution with appropriate degrees of freedom (DF), as in simple random sampling (SRS). We propose three estimates of the DF for the t approximation. The first two estimators are based on the Welch-type approximation (Welch, 1947), where estimating the DF (denoted ν) relies on estimating the in-stratum variances of (judgment) order statistics, $\sigma^2_{[h]} = \mathrm{var}\left(Y_i | r_i = h\right)$ for $h = 1, \ldots, H$. In our first estimator of the DF, we naturally plug in $\sigma^2_{[h]}$ using the sample variance $s^2_{[h]}$. For the second estimator, we compute $\sigma^2_{(h)}$ under the normality assumption to replace $\sigma^2_{[h]}$, which requires perfect ranking. Note that, normal distributions belong to location-scale families so that the in-stratum variances under perfect ranking (i.e., the variances of order statistics) satisfy $\sigma^2_{(h)} = \sigma^2 v_{(h)}$, where $v_{(h)}$ is the variance of the hth order statistic of the standard normal distribution, $h = 1, \ldots, H$, and σ^2 is the population variance. The third estimator simply sets the DF as $n - H$, the difference between the sample size and the set size of the RSS. It could be naively thought as the sample size minus the number of estimated parameters (in-stratum means). We denote the three estimators by $\widehat{\nu}_{\mathrm{samp}}$, $\widehat{\nu}_{\mathrm{norm}}$ and $\widehat{\nu}_{\mathrm{naive}}$, respectively.

This paper is organized as follows. In Section 2, we introduce the proposed t-approximation and three estimators of the DF in detail. In Section 3, we numerically study the performance of the proposed approximation and compare with the performance based on the asymptotic normality. We first compare the approximation error, and then consider the interval estimation of the population mean and compare the coverage probabilities based on the three approximate t distributions with those based on the standard normal distribution. The comparison is made for various settings; unbalanced/balanced design, perfect/imperfect ranking, and independent/correlated rankers. In Section 4, we apply the proposed procedures to estimate the average tree height using the tree data from Chen, Bai, and Sinha (2003) and Platt, Evans, and Rathbun (1988). In Section 5, we briefly summarize the paper and discuss the extension of the proposed approximations to the two sample problem.