Estimating Variances of Strata in Ranked Set Sampling

Min Chen, Johan Lim *

Abstract

Ranked set sampling (RSS) is an established cost-effective sampling method. In RSS, the variance of observations in each ranked set plays an important role in finding an optimal design for unbalanced RSS and in inferring the population mean. The empirical estimator (i.e., the sample variance in a given ranked set) is most commonly used for estimating the variance in the literature. However, the empirical estimator does not use the information in the entire data over different ranks. Further, it is highly variable when the sample size is not large enough, as is typical in RSS applications. In this paper, we propose a plug-in estimator for the variance of each stratum, which is more efficient than the empirical one. The estimator uses a result in order statistics which characterizes the cumulative distribution function (CDF) of the rth order statistics $F_{(r)}(x)$ as a function of the population CDF $F(x)$. We analytically prove the asymptotic normality of the proposed estimator. We further apply it to estimate the standard error of the RSS mean estimator. Both our simulation and empirical study show that our estimators consistently outperform existing methods.

Keywords: cumulative distribution function, judgment post-stratification, order statistics, plug-in estimator, population mean estimator, variance estimation.

1 Introduction

Ranked set sampling (RSS) is an established cost-effective sampling method. It is useful in situations where the characteristic of interest is expensive to measure, but sampling units can be easily gathered and ranked by some means not requiring quantification. Theoretical development shows that the RSS estimator of population mean is at least as efficient as the estimator from a simple random sample (SRS) of equal sample size. We refer readers to Chen

*Min Chen is Postdoctoral Associate, Department of Epidemiology and Public Health, Yale University, New Haven, CT 06510, USA. Johan Lim is Associate Professor, Department of Statistics, Seoul National University, Seoul 151-747, Korea.

et al. (2003) and the references therein for an overview of RSS. For recent developments in RSS, including judement poststratification that is a variation of RSS, please see, e.g., MacEachern et al. (2004), Wang et al. (2006), Óztúrk (2008), etc.

For RSS experiments, the variance of observations in each ranked set (or stratum) plays an important role. In designing unbalanced RSS, the allocation of units to different strata is a key issue to get an efficient estimate of the population mean. The Neyman allocation assigns a given number of units to each stratum to minimize the variance of the RSS mean estimator (say $\hat{\mu}$). To be specific, it assigns n_r units to the rth stratum, where $n_r \propto \sigma_{(r)}$ and $\sigma^2_{(r)}$ is the variance of the rth stratum. Also note that the standard error (SE) of $\hat{\mu}$ is a function of the $\sigma^2_{(r)}$'s, which is needed to make statistical inference about population mean. Thus, the estimation of the $\sigma^2_{(r)}$'s is the first but a very important step for both problems. The most commonly used estimator of $\sigma^2_{(r)}$ is simply the sample variance of observations in the rth stratum. However, this empirical estimator of $\sigma^2_{(r)}$ only relies on the units from the rth stratum, and does not utilize the information from the entire data set over different strata. Consequently, it has a large variance when the sample size is small, as is typical in many RSS applications that have cost constraints.

In this paper, we propose a new estimator of $\sigma^2_{(r)}$ that is more efficient than the empirical estimator. The proposed estimator is motivated by a well-known result in order statistics, which characterizes the cumulative distribution function (CDF) $F_{(r)}(x)$ as a function of the population CDF $F(x)$. Thus, once we can estimate the population CDF $F(x)$ reliably, $F_{(r)}(x)$ and its variance $\sigma^2_{(r)}$ can be estimated from the relationship between $F_{(r)}(x)$ and $F(x)$.

The remainder of the paper is organized as follows. In Section 2, we introduce new estimators for $F_{(r)}(x)$ and $\sigma^2_{(r)}$, which we refer to as plug-in estimators. We analytically derive the asymptotic properties for the proposed estimators and numerically examine their performance. For symmetric distributions, to achieve better performance, we also propose a modified plug-in estimator. In Section 3, we apply the new estimators to construct the SE of

the RSS mean estimator $\hat{\mu}$, in which we consider both binary and non-binary data. Section 4 illustrates our estimators using tree data from Platt et al. (1988). Section 5 concludes the paper by a brief discussion.

2 Estimating stratum variance $\sigma^2_{(r)}$

2.1 Plug-in estimators

Consider a ranked set sample (RSS) of size N from a population with CDF $F(x)$:

$$X_{(1)1}, \ldots, X_{(1)n_1}, \cdots, X_{(H)1}, \ldots, X_{(H)n_H},$$

where H is the number of ranked sets, n_r is the number of observations in the stratum of rank r and $N = \sum_{r=1}^{H} n_r$. We assume that judgment ranking is perfect. Thus, $X_{(r)i}$ is the rth smallest among its H comparison units, and its CDF, denoted by $F_{(r)}(x)$, is that of the rth order statistic among H samples.

Our new estimators of the CDF $F_{(r)}(x)$ and the corresponding variance $\sigma^2_{(r)}$ for stratum r are motivated by the fact that $F_{(r)}(x)$ is a function of $F(x)$. That is,

$$F_{(r)}(x) = \mathrm{I}_{F(x)}(r, H - r + 1), \tag{1}$$

where $\mathrm{I}_p(a, b)$ is the incomplete beta function

$$\mathrm{I}_p(a, b) = \frac{1}{\mathrm{Beta}(a, b)} \int_0^p t^{a-1}(1-t)^{b-1} dt, \quad \text{and} \quad \mathrm{Beta}(a, b) = \int_0^1 t^{a-1}(1-t)^{b-1} dt.$$

Two special cases of the identity (1) are those with $r = 1$ and H. For $r = 1$,

$$F_{(1)}(x) = \mathrm{I}_{F(x)}(1, H) = \frac{1}{\mathrm{Beta}(1, H)} \int_0^{F(x)} (1-t)^{H-1} dt = 1 - \{1 - F(x)\}^H,$$

and for $r = H$,

$$F_{(H)}(x) = \mathrm{I}_{F(x)}(H, 1) = \frac{1}{\mathrm{Beta}(H, 1)} \int_0^{F(x)} t^{H-1} dt = F(x)^H.$$

Given an estimate of $F(x)$ (say $\widehat{F}$), we can propose a new plug-in estimator of $F_{(r)}(x)$ using (1), namely,

$$\widehat{F}^{\mathrm{PI}}_{(r)}(x) = \mathrm{I}_{\widehat{F}(x)}(r, H-r+1).$$

In RSS studies, estimation of the population CDF is well established in the literature. To list a few, Stokes and Sager (1988) considered the empirical CDF from RSS data. Kvam and Samaniego (1994) studied a nonparametric maximum likelihood estimator and proposed an EM algorithm to compute it. A kernel estimator was studied by Chen (1999). More recently, Óztúrk (2007) has proposed an estimator using the stochastic ordering among the $F_{(r)}(x)$'s, which is robust to ranking error.

From the estimator $\widehat{F}^{\mathrm{PI}}_{(r)}(x)$, we can further define a plug-in estimator $\widehat{V}^{\mathrm{PI}}_{(r)}$ for the variance of the rth stratum, $\sigma^2_{(r)}$. Let $\widehat{\mu}_{(r)}$ be an estimate of the mean of the rth stratum. In the case when X, the variable of interest, is discrete and has a support $\mathcal{X}$, the plug-in estimator can be given by

$$\widehat{V}^{\mathrm{PI}}_{(r)} = \sum_{x\in\mathcal{X}} (x-\widehat{\mu}_{(r)})^2 \left[\widehat{F}^{\mathrm{PI}}_{(r)}(x+1) - \widehat{F}^{\mathrm{PI}}_{(r)}(x)\right]. \tag{2}$$

In the case when X is continuous, we have

$$\widehat{V}^{\mathrm{PI}}_{(r)} = \int \left(x-\widehat{\mu}_{(r)}\right)^2 \widehat{f}^{\mathrm{PI}}_{(r)}(x)\, dx, \tag{3}$$

where

$$\widehat{f}^{\mathrm{PI}}_{(r)}(x) = \frac{1}{\mathrm{Beta}(r, H-r+1)} \left[\widehat{F}(x)\right]^{r-1} \left[1-\widehat{F}(x)\right]^{H-r} d\widehat{F}(x).$$

In (2) or (3), we choose $\widehat{\mu}_{(r)}$ to be the isotonic regression estimator, which is given by

$$\widehat{\mu}_{(r)} = \max_{s\le r} \min_{t\ge r} \sum_{g=s}^{t} \frac{n_g \bar{Y}_{(g)}}{n_{st}}, \quad n_{st} = \sum_{g=s}^{t} n_g.$$

We could choose $\widehat{\mu}_{(r)}$ to be the sample mean $\bar{Y}_{(r)}$ of the rth stratum. However, based on our numerical experience, we find that using the isotonic version has better performance. This

is not surprising, since the isotonic estimator imposes the built-in ordering in the means of the strata.

Finally, we remark that both $\widehat{F}^{\mathrm{PI}}_{(r)}(x)$ and $\widehat{V}^{\mathrm{PI}}_{(r)}$ only require estimation of the population CDF, which is not necessarily from a RSS sample. This provides a great benefit in problems of optimally designing unbalanced RSS experiments. We can easily obtain our estimates and the optimal allocation for an unbalanced design from simple random samples, which we commonly encounter in the previous literature.

2.2 Asymptotics

Here we obtain the asymptotic properties of the plug-in estimators, $\widehat{F}^{\mathrm{PI}}_{(r)}(x)$ and $\widehat{V}^{\mathrm{PI}}_{(r)}$. In what follows, we use $\widehat{F}_{(r)n}$ and $\widehat{V}_{(r)n}$ to emphasize that they depend on the sample sizes $n = (n_1, \ldots, n_H)$ and $N = \sum_{r=1}^{H} n_r$.

We assume that $\widehat{F}_n(x)$, the estimate of the population CDF, satisfies the central limit theorem (CLT). That is, for fixed x,

$$\sqrt{N}\left[\widehat{F}_n(x) - F(x)\right]$$

converges in distribution to a Gaussian random variable $\sigma(x)Z$, where Z is a standard normal random variable. This assumption, which we refer to as (A), is true for most of existing estimators of $F(x)$ including those mentioned in Section 2.1. For example, consider the empirical estimator proposed by Stokes and Sager (1988),

$$\widehat{F}^{\mathrm{SS}}_n(x) = \frac{1}{H}\sum_{r=1}^{H} \widehat{F}^{\,\mathrm{E}}_{(r)n}(x),$$

where

$$\widehat{F}^{\,\mathrm{E}}_{(r)n}(x) = \frac{1}{n_r}\sum_{i=1}^{n_r} \mathrm{I}\left(X_{(r)i} \le x\right), \quad r = 1, \ldots, H$$

and $I(\cdot)$ is the indicator function. For fixed x, we know from the functional CLT that

$$\sqrt{n_r}\left\{\widehat{F}^{\,\mathrm{E}}_{(r)n}(x) - F_{(r)}(x)\right\} = \sqrt{n_r}\left\{\frac{1}{n_r}\sum_{i=1}^{n_r}\left[\mathrm{I}(X_{(r)i} \le x) - F_{(r)}(x)\right]\right\} \tag{4}$$

converges in distribution to a normal distribution with mean 0 and variance $F_{(r)}(x)(1 - F_{(r)}(x))$. By summing up the asymptotic results of each $\widehat{F}^{\,\mathrm{E}}_{(r)n}(x)$ in (4), we have

$$\begin{aligned}\sqrt{N}\left\{\widehat{F}^{\mathrm{SS}}_n(x) - F(x)\right\} &= \frac{1}{H}\sum_{r=1}^{H}\sqrt{\frac{N}{n_r}}\sqrt{n_r}\left\{\frac{1}{n_r}\sum_i\left[\mathrm{I}(x_{(r)i} \le x) - F_{(r)}(x)\right]\right\}. \\ &\to \frac{1}{H}\sum_{r=1}^{H}(1/\sqrt{q_r})\sqrt{F_{(r)}(x)\left[1 - F_{(r)}(x)\right]}Z_r,\end{aligned}$$

where $q_r = n_r/N$ and $Z_1, \ldots, Z_H$ are IID from the standard normal distribution. Thus,

$$\sqrt{N}\left[\widehat{F}^{\mathrm{SS}}_n(x) - F(x)\right]$$

converges in distribution to a normal random variable with mean 0 and variance

$$\sigma^2(x) = \frac{1}{H^2}\sum_{r=1}^{H}\frac{1}{q_r}F_{(r)}(x)\left[1 - F_{(r)}(x)\right]. \tag{5}$$

We now prove the central limit theorem for $\widehat{F}^{\mathrm{PI}}_{(r)n}(x)$, which further gives the consistency of $\widehat{F}^{\mathrm{PI}}_{(r)n}(x)$ and $\widehat{V}^{\mathrm{PI}}_{(r)n}$ as well. Let the function $\mathrm{G}_r(F(x))$ be

$$\mathrm{G}_r(F(x)) = \mathrm{I}_{F(x)}(r, H - r + 1).$$

Then, for fixed x,

$$\widehat{F}^{\mathrm{PI}}_{(r)n}(x) = \mathrm{G}_r(\widehat{F}_n(x)),$$

and

$$\begin{aligned}\sqrt{n_r}\mathrm{U}_r(x) &\equiv \sqrt{n_r}\left[\mathrm{G}_r(\widehat{F}_n(x)) - \mathrm{G}_r(F(x))\right] \\ &\approx \sqrt{\frac{n_r}{N}}\mathrm{G}'_r(F(x))\sqrt{N}\left[\widehat{F}_n(x) - F(x)\right],\end{aligned}$$

where

$$\mathbf{G}'_r(F(x)) = \frac{1}{\mathrm{Beta}(r, H - r + 1)}F(x)^{r-1}\left[1 - F(x)\right]^{H-r}.$$

Thus, under (A), $\sqrt{n_r}\mathrm{U}_r(x)$ converges to a Gaussian random variable with mean 0 and variance

$$\tau_r^2 = q_r \mathbf{G}_r'(F(x))^2 \sigma^2(x),$$

where $\sigma^2(x)$ is defined in (A). Here, $\sigma^2(x)$ depends on the choice of $\widehat{F}_n(x)$. For example, if $\widehat{F}_n(x) = \widehat{F}_n^{\mathrm{SS}}(x)$, it is equal to that in (5). Finally, noting that $\widehat{\mu}_{(r)}$ is a consistent estimator of $\mu_{(r)}$, we have

$$\widehat{V}_{(r)n}^{\mathrm{PI}} - V_{(r)} \quad = \quad \int (x - \mu_{(r)})^2 d\left\{\widehat{F}_{(r)n}^{\mathrm{PI}}(x) - F_{(r)}(x)\right\} - \left(\widehat{\mu}_{(r)} - \mu_{(r)}\right)^2,$$

which converges to 0 in probability, if $\mathrm{E}|X_{(r)}|^{2+\delta}$ is finite for some $\delta > 0$.

2.3 Numerical Studies

In this section, we implement numerical studies to investigate the efficiency of the proposed plug-in estimator of $\sigma_{(r)}^2$ under various settings. We generate balanced RSS samples from six different distributions: standard normal, uniform (0,1), gamma with shape parameter 5 and scale parameter 1, standard exponential and standard log-normal as in MacEachern et al. (2002) and Wang et al. (2008), plus Poisson with mean 5. We set the number of ranked sets $H = 2, 3, 4, 5, 10$, and the number of units in each stratum $m = 1, 2, 3, 4, 5, 20$. Under each setting, the relative efficiency (RE) is estimated from 20,000 replicates. Here, relative efficiency (RE) is defined as the ratio of the mean squared errors (MSE) between two estimators. We use the Stokes and Sager estimator $\widehat{F}^{\mathrm{SS}}$ of the population CDF to calculate the estimators $\widehat{V}_{(r)}^{\mathrm{PI}}$ for diffirent strata.

Figures 1-3 report the simulated relative efficiency of $\widehat{V}_{(r)}^{\mathrm{PI}}$ to $s_{(r)}^2$, the sample variance of the rth stratum. Here, we only report results for normal, lognormal and Poisson data, due to the limit of space. The REs of $\widehat{V}_{(r)}^{\mathrm{PI}}$ to $s_{(r)}^2$ are larger than 1 in all the cases we consider, meaning that $\widehat{V}_{(r)}^{\mathrm{PI}}$ performs uniformly better than $s_{(r)}^2$. In many cases, the values of REs are higher than 2, and could be much higher as in the lognormal distribution. This means

substantial improvement could be obtained. Among the six distributions, the REs for the Poisson case are relatively small, compared to the other five continuous distributions. Also, for the continuous distributions, we find that the REs for $r = 1$ (the smallest rank) among H strata tend to be particularly higher than those for the other ranks when the sample size m of each stratum is small. Overall, it is hard to find a relationship between RE and H or between RE and m while fixing one of them.

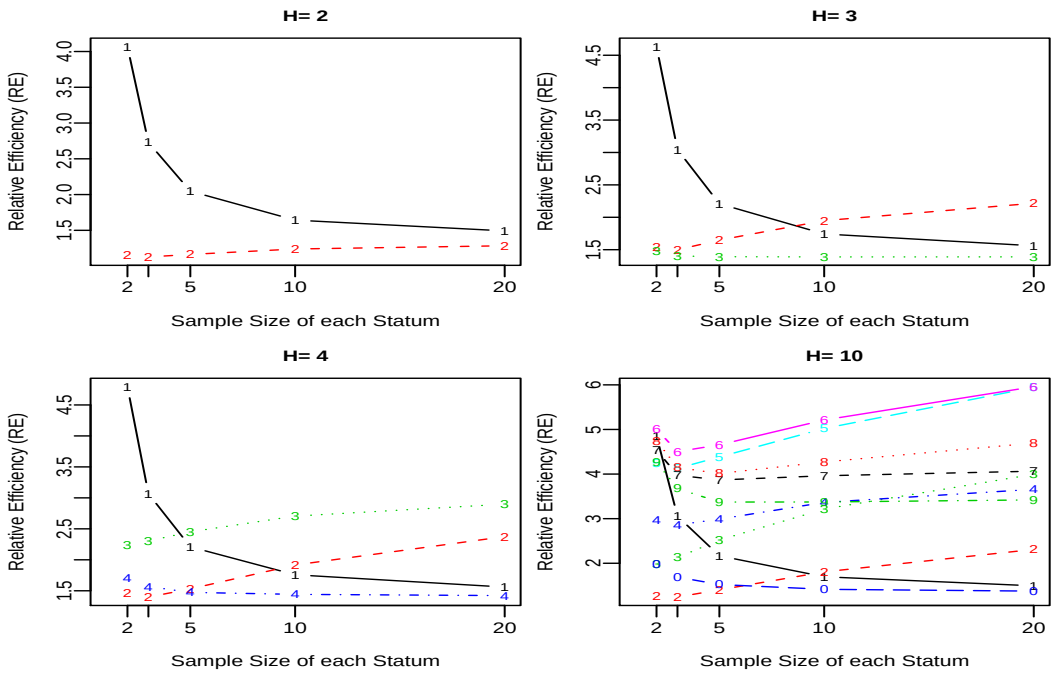

Figure 1: Simulated relative efficiency of $\widehat{V}^{\mathrm{PI}}_{(r)}$ over $s^2_{(r)}$ for normal data.

We further investigate the biases and variances of $s^2_{(r)}$ and $\widehat{V}^{\mathrm{PI}}_{(r)}$. For each distribution, we compute the estimates for the variance of each stratum from RSS samples with $H = 3$ and $m = 2, 3, 5, 10$. Figure 4 reports results from normal data, where the left panels shows boxplots for $\widehat{V}^{\mathrm{PI}}_{(r)}$ and the right ones for $s^2_{(r)}$. The red line marks the true value of $\sigma^2_{(r)}$ and the "*" marks the sample mean of the estimates. The figure shows that $s^2_{(r)}$ is unbiased but has a larger variance. On the other hand, $\widehat{V}^{\mathrm{PI}}_{(r)}$ tend to underestimate $\sigma^2_{(r)}$ for small r, and overestimate $\sigma^2_{(r)}$ for larger r. However, it has a much smaller variance than $s^2_{(r)}$ and overall,

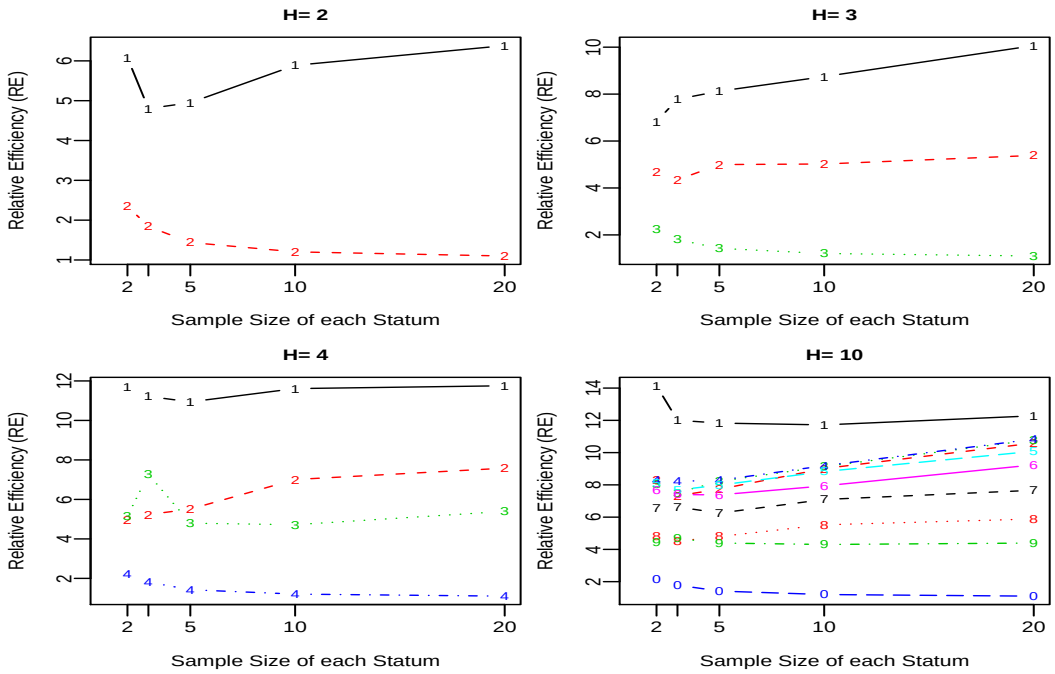

Figure 2: Simulated relative efficiency of $\widehat{V}^{\mathrm{PI}}_{(r)}$ over $s^2_{(r)}$ for log-normal data.

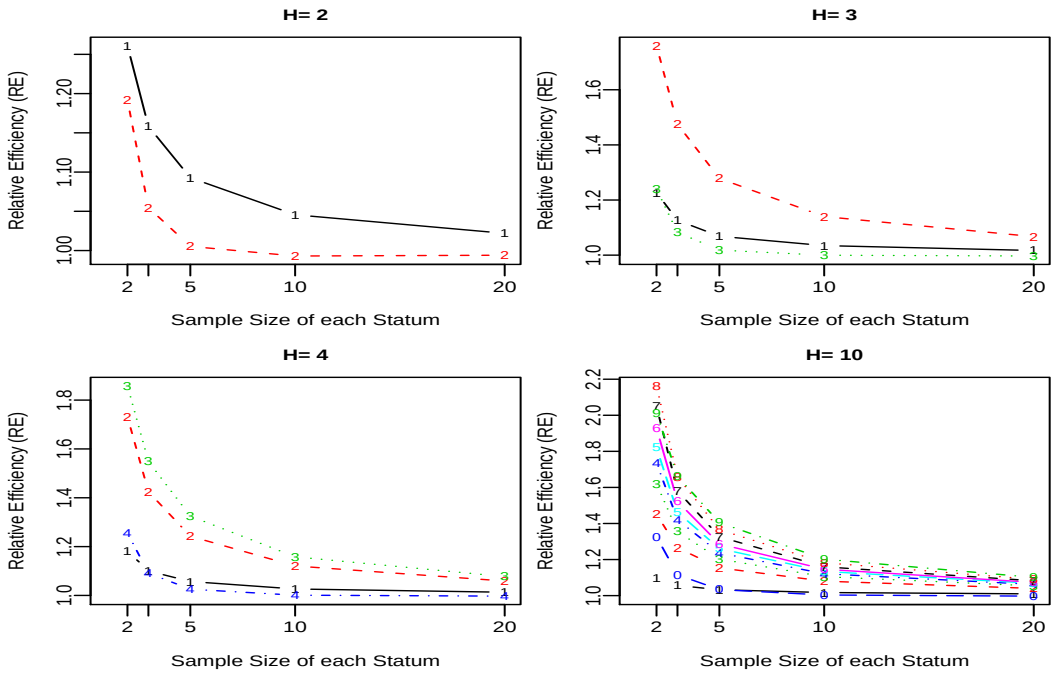

Figure 3: Simulated relative efficiency of $\widehat{V}^{\mathrm{PI}}_{(r)}$ over $s^2_{(r)}$ for Poisson data.

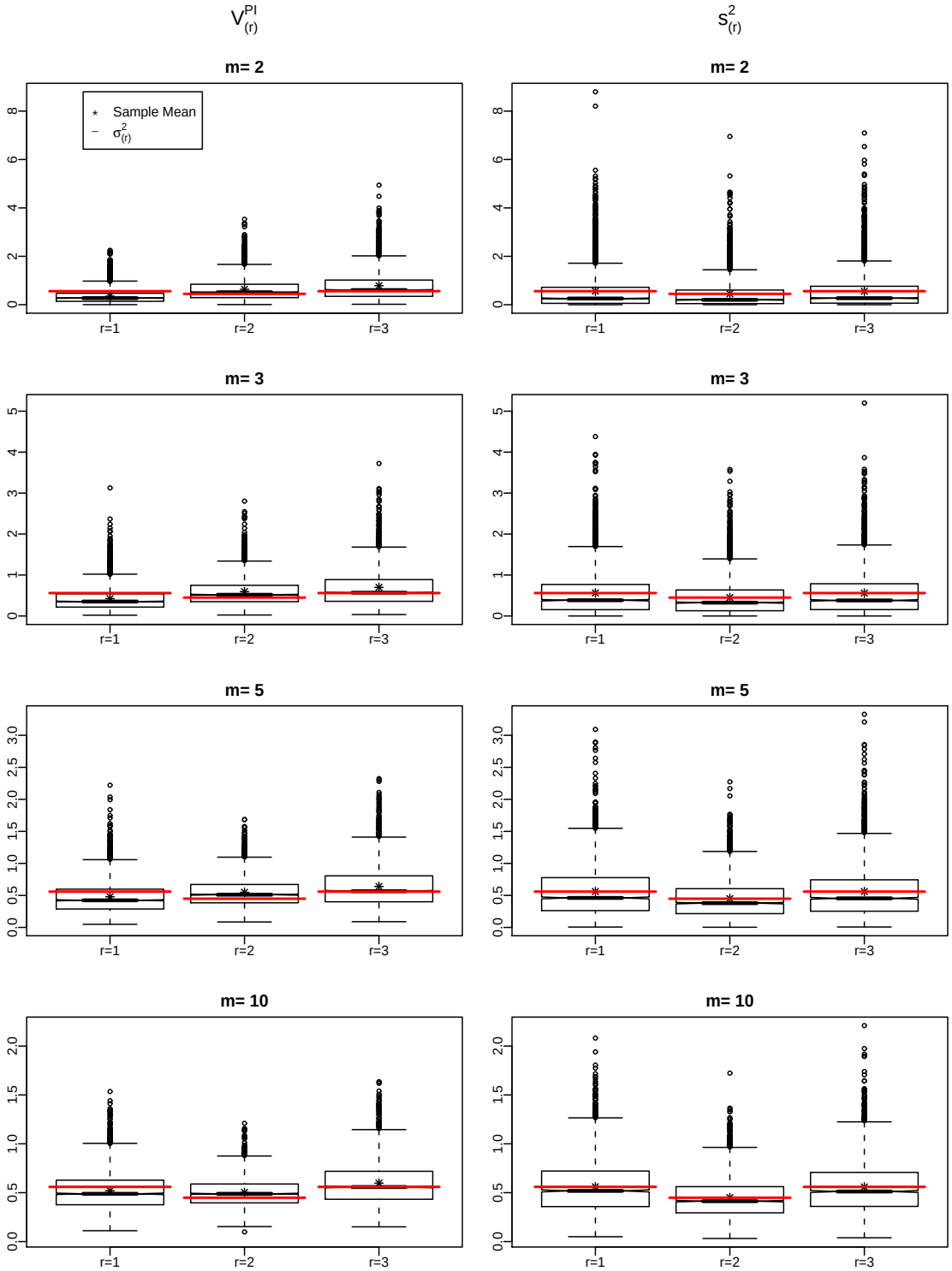

Figure 4: Boxplots of estimates of $\sigma^2_{(r)}$ for normal data with $H = 3$ and $m = 2, 3, 5, 10$.

its MSE is smaller. The results for the other distributions lead to the same observations as in the normal case.

2.4 Modified Estimators for Symmetric distributions

We first show that, when the population distribution is symmetric with finite mean μ, $\sigma^2_{(r)} = \sigma^2_{(H-r+1)}$ for $r = 1, \ldots, H$. Suppose that $X_1, \ldots, X_H$ are IID from a symmetric distribution with a density function $f(x)$, satisfying $f(\mu - x) = f(\mu + x)$. Let $F(x)$ and $F_{(r)}(x)$ be the population CDF and the CDF of the rth order statistics. We know

$$F(\mu - x) + F(\mu + x) = 1$$

from the symmetry. Then,

$$\begin{aligned}
F_{(r)}(\mu - x) &= \frac{1}{\text{Beta}(r, H-r+1)} \int_0^{F(\mu-x)} t^{r-1}(1-t)^{H-r} dt \\
&= \frac{1}{\text{Beta}(r, H-r+1)} \int_{1-F(\mu-x)}^{1} (1-s)^{r-1} s^{H-r} ds \\
&= 1 - \frac{1}{\text{Beta}(r, H-r+1)} \int_0^{F(\mu+x)} (1-s)^{r-1} s^{H-r} ds \\
&= 1 - F_{(H-r+1)}(\mu + x).
\end{aligned}$$

So we have

$$\begin{aligned}
\mu_{(r)} &= \int_{-\infty}^{+\infty} x dF_{(r)}(x) \\
&= -\int_{-\infty}^{+\infty} (\mu - t) dF_{(r)}(\mu - t) \\
&= 2\mu - \int_{-\infty}^{+\infty} x dF_{(H-r+1)}(x) \\
&= 2\mu - \mu_{(H-r+1)},
\end{aligned}$$

and

$$\begin{aligned}
\sigma^2_{(r)} &= \int_{-\infty}^{+\infty} \left(x - \mu_{(r)}\right)^2 dF_{(r)}(x) \\
&= -\int_{-\infty}^{+\infty} \left(\mu - t - \mu_{(r)}\right)^2 d[1 - F_{(H-r+1)}(\mu + t)] \\
&= \int_{-\infty}^{+\infty} \left(\mu_{(H-r+1)} - x\right)^2 dF_{(H-r+1)}(x) \\
&= \sigma^2_{(H-r+1)}.
\end{aligned}$$

It is reasonable to expect that our plug-in estimator $\widehat{V}^{\mathrm{PI}}_{(r)}$ of $\sigma^2_{(r)}$ satisfies the symmetric property. However, the results in Section 2.3 clearly show that, for symmetric distributions, the performance of $\widehat{V}^{\mathrm{PI}}_{(r)}$ is not equal to that of its counterpart $\widehat{V}^{\mathrm{PI}}_{(H-r+1)}$. Recall that $\widehat{V}^{\mathrm{PI}}_{(r)}$ tends to have a positive bias when r is close to H, but a negative bias when r is close to 1. Thus, we propose a modified estimator based on our plug-in estimator by taking the average of the corresponding top and bottom strata,

$$\tilde{V}^{\mathrm{PI}}_{(r)} = \tilde{V}^{\mathrm{PI}}_{(H-r+1)} = \frac{\widehat{V}^{\mathrm{PI}}_{(r)} + \widehat{V}^{\mathrm{PI}}_{(H-r+1)}}{2},$$

for $r = 1, 2, \ldots, [H/2]$. By doing so, the bias tends to be canceled partially while the variance is still similar. Overall, we expect that the MSE of this modified estimator is smaller than that of the original one.

We numerically compare the performance of the modified estimator $\tilde{V}^{\mathrm{PI}}_{(r)}$ to the original plug-in estimator $\widehat{V}^{\mathrm{PI}}_{(r)}$. We consider three symmetric distributions, uniform$(0, 1)$, standard normal, and t with degrees of freedom 5. All the other settings are the same as those in Section 2.3. We compute the REs of $\tilde{\sigma}^2_{(r)}$ to $\widehat{\sigma}^{2,\mathrm{PI}}_{(r)}$ and plot them in Figure 6 (for normal data only). The figure confirms that $\tilde{V}^{\mathrm{PI}}_{(r)}$ are uniformly better than $\widehat{V}^{\mathrm{PI}}_{(r)}$. The results from the other two symmetric distributions support the conclusion, too.

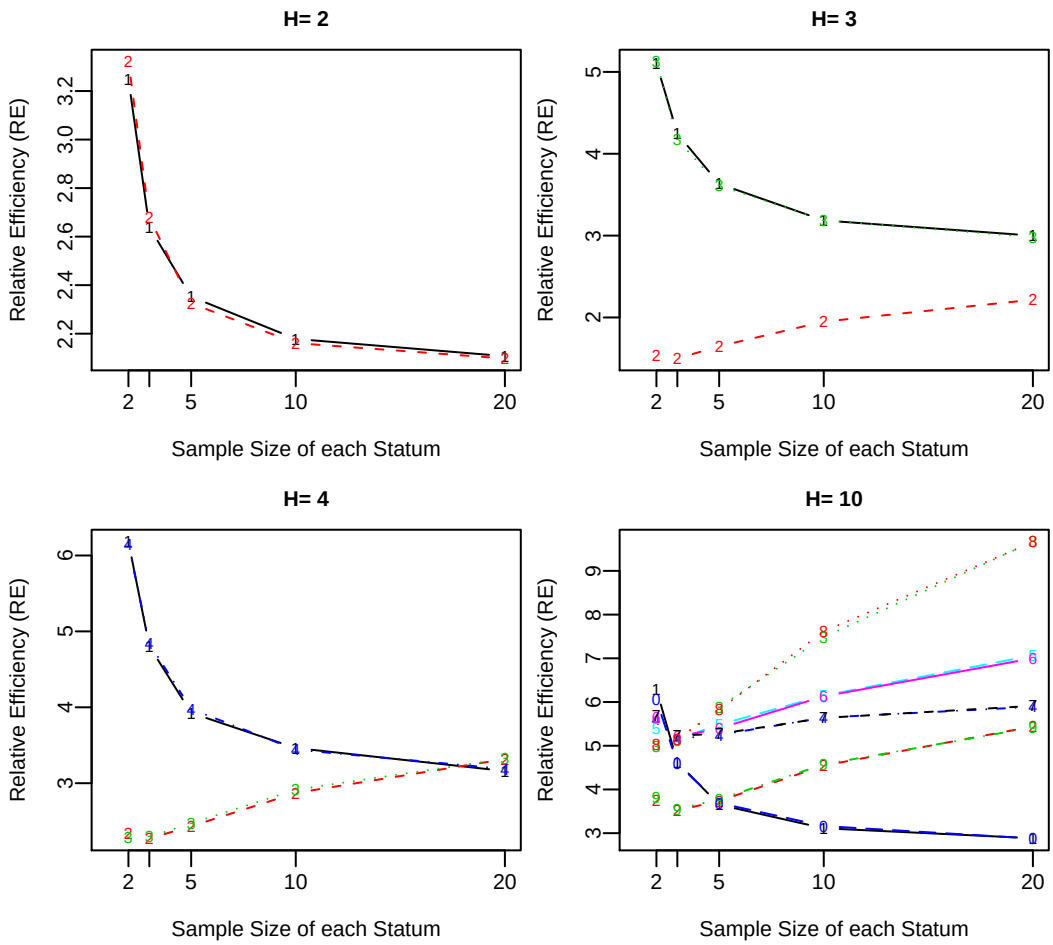

Figure 5: Simulated relative efficiency of the modified plug-in estimator $\tilde{V}^{\mathrm{PI}}_{(r)}$ over the original plug-in estimator $\widehat{V}^{\mathrm{PI}}_{(r)}$ for normal data.

3 Estimating the SE of the RSS mean estimator

3.1 Non-binary cases

Here, we assume the variable of interest X is non-binary. We propose a new estimator for the standard error (SE) of the RSS mean estimator using our plug-in estimator of $\sigma^2_{(r)}$. In a RSS experiment, the population mean is estimated by

$$\widehat{\mu} = \frac{1}{H}\sum_{r=1}^{H}\frac{\sum_{i=1}^{n_r} X_{(r)i}}{n_r}$$

and its variance is

$$V_\mu \equiv \mathrm{var}(\widehat{\mu}) = \frac{1}{H^2} \sum_r \frac{\sigma^2_{(r)}}{n_r}.$$

The above variance is commonly estimated using the sample variance $s^2_{(r)}$ of each stratum, namely

$$\widehat{V}^{\mathrm{E}}_\mu = \frac{1}{H^2} \sum_{r=1}^{H} \frac{s^2_{(r)}}{n_r}, \tag{6}$$

This empirical estimator is known to be unreliable when the sample size n_r is small for some r.

As an alternative to the empirical estimator $\widehat{V}^{\mathrm{E}}_\mu$, we apply $\widehat{V}^{\mathrm{PI}}_{(r)}$ to estimating (6). The proposed new estimator of $\mathrm{var}(\widehat{\mu})$ becomes

$$\widehat{V}^{\mathrm{PI}}_\mu = \frac{1}{H^2} \sum_{r=1}^{H} \frac{\widehat{V}^{\mathrm{PI}}_{(r)}}{n_r}.$$

It is easy to show from the consistency of $\widehat{V}^{\mathrm{PI}}_{(r)}$ for every r that the estimator $\widehat{V}^{\mathrm{PI}}_\mu$ is consistent, too. To show its efficiency, we again numerically compare the REs of $\widehat{V}^{\mathrm{PI}}_\mu$ to $\widehat{V}^{\mathrm{E}}_\mu$. We use the same simulation settings as in Section 2.3. The estimated REs are plotted in Figure 6. Again, we find that the REs are larger than 1 in every case we consider. We also find that REs decreases to 1 as m increases in the exponential, Gamma, lognormal and Poisson cases (heavier-tail distributions) while REs sometimes increase for uniform and normal distributions. Here, a random variable X with density $f(x)$ has a heavier tail than a random variable Y with density $g(y)$ if $g(|t|)/f(|t|)$ decreases to 0 as $|t|$ increases.

3.2 The binary case

When the variable of interest X is binary, the ranking can be performed based on a correlated concomitant variable (Lacayo et al., 2002). Chen et al. (2005) propose to rank sample units according to estimated probabilities of success from a logistic regression model. Below we study the SE of the RSS estimator for the population proportion, say p. An application of unbalanced RSS in estimating p can be find in Chen et al. (2006). The RSS proportion

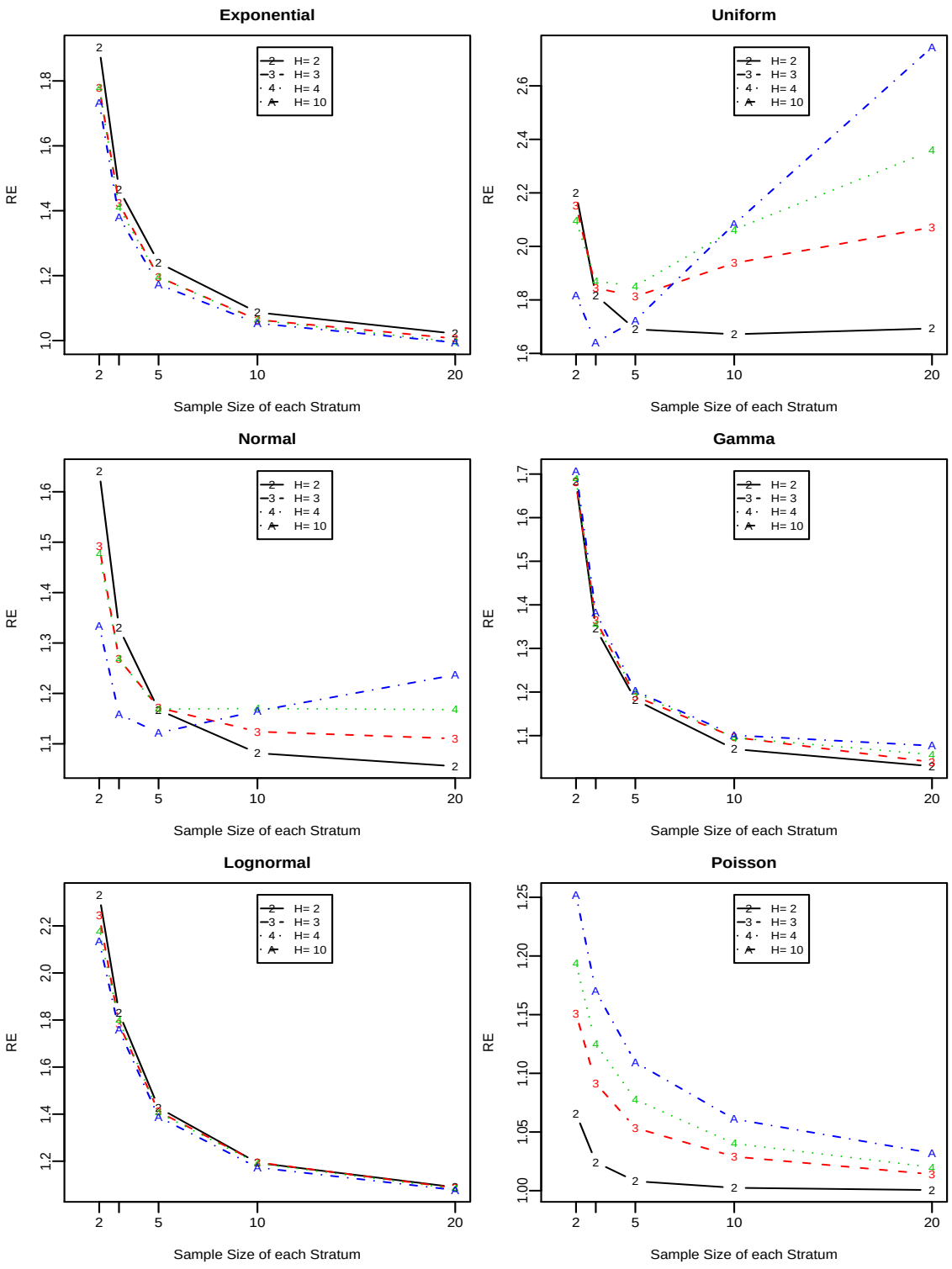

Figure 6: Simulated relative efficiency of $\widehat{V}_{\mu}^{\mathrm{PI}}$ to $\widehat{V}_{\mu}^{\mathrm{E}}$ for estimating $\mathrm{var}(\widehat{\mu})$

estimator $\hat{p}$ is the RSS mean estimator when the variable of interest X is binary. The unknown true variance of $\hat{p}$ is then

$$\mathrm{var}(\widehat{p}) = \frac{1}{H^2} \sum_r \frac{p_{(r)} \left[1 - p_{(r)}\right]}{n_r},$$

where $p_{(r)}$ is the probability of 1 for the rth order statistic $X_{(r)i}$.

Two estimators of the above SE of $\hat{p}$ were proposed in the literature (Chen et al., 2003). One is based on the variance of binary observations; that is

$$\widehat{V}_p^{\mathrm{B}} = \frac{1}{H^2} \sum_{r=1}^{H} \frac{\widehat{p}_{(r)}[1 - \widehat{p}_{(r)}]}{n_r},$$

where $\widehat{p}_{(r)}$ is an estimator of $p_{(r)}$. The other is based on the sample variance of each stratum; that is

$$\widehat{V}_p^{\mathrm{E}} = \frac{1}{H^2} \sum_{r=1}^{H} \frac{s_{(r)}^2}{n_r}.$$

The estimators $\widehat{V}_p^{\mathrm{B}}$ and $\widehat{V}_p^{\mathrm{E}}$ have larger variances when n_rs are small for some r. Further, they are not well defined when $n_r = 1$ for some r.

We extend the results in Section 2 to binary observations and propose a plug-in estimator of the SE for the RSS population estimator. Here, we propose to estimate $p_{(r)}$ using (1). Noting that the estimate of the population CDF $\widehat{F}(x)$ has a support on $x = 0, 1$, (1) results in $\widehat{p}_{(r)}^{\mathrm{PI}} = \mathrm{I}_{\widehat{p}}\big(r, H - r + 1\big)$, where $\widehat{p} = \widehat{F}(1)$. Thus, the proposed plug-in estimator of $\mathrm{var}(\hat{p})$ is

$$\widehat{V}_p^{\mathrm{PI}} = \frac{1}{H^2} \sum_r \frac{\widehat{p}_{(r)}^{\mathrm{PI}} \left[1 - \widehat{p}_{(r)}^{\mathrm{PI}}\right]}{n_r}.$$

We again numerically compare the MSE of the plug-in estimator $\widehat{V}_p^{\mathrm{PI}}$ with the two existing estimators $\widehat{V}_p^{\mathrm{E}}$ and $\widehat{V}_p^{\mathrm{B}}$. We consider the cases with $H = 5$, 10 and 20. For each H, the n_r are chosen as (i) all are 1 (ALL1), (ii) all are 3 (ALL3), and (iii) the first half of n_rs are 3 and the other half are 10 (MIX). In the simulation, the true population proportion is fixed as $p = 0.1, 0.3$, or 0.5. We generate $10,000$ data sets for each case and compute the RE of the plug-in estimator to each existing estimator. The REs are reported in Table 1, which clearly shows the plug-in estimator significantly outperforms the two existing estimators.

		p=0.1		p=0.3		p=0.5	
H	Case	B/PI	E/PI	B/PI	E/PI	B/PI	E/PI
5	ALL1	(0.01034)		(0.00657)		(0.00285)	
	ALL3	1.14	1.36	4.14	4.56	14.37	14.76
	MIX	3.07	2.94	14.33	13.60	3.24	4.23
10	ALL1	(0.00256)		(0.00084)		(0.00030)	
	ALL3	1.94	2.12	7.89	7.77	43.55	41.61
	MIX	6.24	6.00	47.85	45.02	3.41	3.60
20	ALL1	(0.0004)		(0.00015)		(0.00003)	
	ALL3	3.83	3.83	13.07	11.54	99.83	88.02
	MIX	11.35	10.89	66.39	63.19	3.45	3.50

Table 1: Simulated relative efficiency of the propose plug-in estimator to each of the two existing methods. "B" represents $\widehat{V}_p^{\mathrm{B}}$, "E" represents $\widehat{V}_p^{\mathrm{E}}$, and "PI" represents $\widehat{V}_p^{\mathrm{PI}}$. The numbers in parenthesis is the MSE of the plug-in estimator since the existing methods are not applicable to ALL1.

4 Empirical Studies

In this section, we apply the proposed plug-in estimator to analyze tree data from Platt et al. (1988) and Chen et al. (2003). We treat the reported tree data as the true population. We apply balanced RSS to the entire height in feet (X) and estimate the average height using the RSS.

First, we numerically compute the REs of the proposed plug-in estimator to the empirical estimator in estimating the SE of the RSS mean estimator $\hat{\mu}$. We draw RSS samples from the entire tree data set with the number of strata $H = 2, 3, 4$ and number of observations in each stratum $m = 2, 3, 5, 10$. We generate $10,000$ RSS data sets and, in each RSS data set, we compute the estimates of variances of strata r, $r = 1, 2, \ldots, H$. The true variance of each stratum is computed by treating the tree data as the population. The REs for estimating the SE of $\hat{\mu}$ are plotted in Figure 7. The proposed plug-in estimator $\widehat{V}_\mu^{\mathrm{PI}}$ is much more efficient than the empirical estimator $\widehat{V}_\mu^{\mathrm{E}}$ for all Hs when m is small, and the RE decreases to 1 as m increases.

Second, to show the practical importance of the proposed estimator, we study the cover-

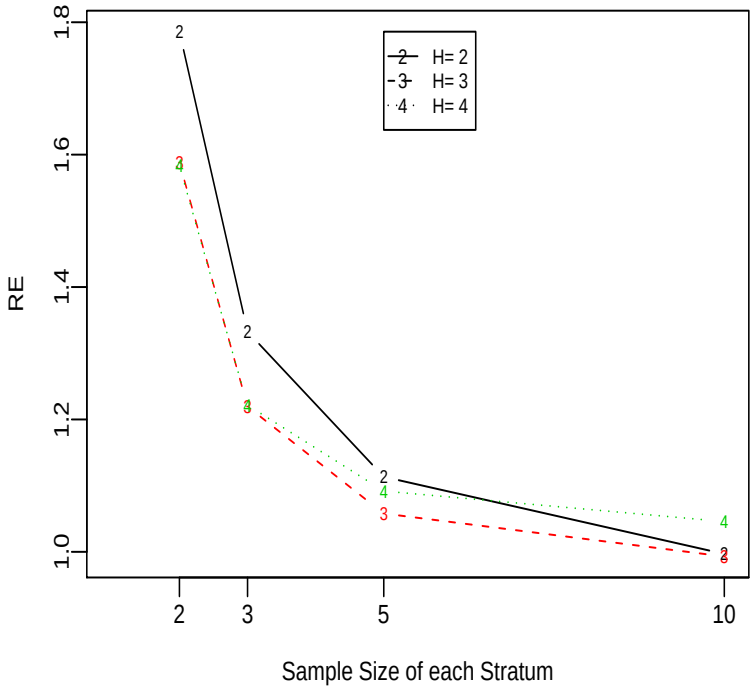

Figure 7: Simulated relative efficiency of $\widehat{V}_{\mu}^{\mathrm{PI}}$ to $\widehat{V}_{\mu}^{\mathrm{E}}$ for estimating var($\widehat{\mu}$) in tree data

age probability of the interval estimator of the population mean, which is closely related with the performance of the SE of the RSS mean estimator. In each RSS data set, we construct $100(1-\alpha)\%$ confidence interval of the population mean as

$$\hat{\mu} \pm z_{\alpha/2}\widehat{V}_{\mu}^{1/2},$$

where $z_{\alpha/2}$ is $100\alpha-$th upper percentile of the standard normal distribution. The computed coverage probability is given by the proportion of data sets whose confidence intervals contain the true average height μ. The results are reported in Table 2. Overall, $\widehat{V}_{\mu}^{\mathrm{PI}}$ provides more accurate coverage probabilities than $\widehat{V}_{\mu}^{\mathrm{E}}$. The improvement is largest for small m.

		90% CP		95% CP	
H	m	E	PI	E	PI
2	2	0.770	0.804	0.833	0.852
2	3	0.815	0.827	0.867	0.873
2	5	0.843	0.848	0.890	0.891
2	10	0.885	0.886	0.927	0.927
3	2	0.789	0.862	0.849	0.906
3	3	0.828	0.860	0.880	0.903
3	5	0.869	0.883	0.912	0.922
3	10	0.905	0.910	0.947	0.949
4	2	0.793	0.882	0.856	0.922
4	3	0.844	0.884	0.891	0.922
4	5	0.886	0.905	0.928	0.940
4	10	0.921	0.928	0.959	0.963

Table 2: Coverage probabilities based on $\widehat{V}_{\mu}^{\mathrm{PI}}$ and $\widehat{V}_{\mu}^{\mathrm{E}}$ for interval estimation in tree data. "E" and "PI" stand for $\widehat{V}_{\mu}^{\mathrm{PI}}$ and $\widehat{V}_{\mu}^{\mathrm{E}}$, respectively.

5 Discussion

In this paper, we are motivated by the identity between the CDFs $F_{(r)}(\cdot)$ and $F(\cdot)$, and propose a new plug-in estimator for variances of strata in RSS. We analytically derive its asymptotic distribution and consistency. We numerically show that the proposed estimator is more efficient than the empirical estimator. We further apply it to estimate the SE of the RSS mean estimator and show it outperforms existing estimators through simulation. The analysis of tree data shows that our plug-in estimator of the SE provides better interval estimation of the population mean than the empirical estimator.

In addition, for symmetric distributions, we propose a modified plug-in estimator $\tilde{V}_{(r)}^{\mathrm{PI}}$ of $\sigma_{(r)}^2$ to take into account the symmetry, which improves the original plug-in estimator $\widehat{V}_{(r)}^{\mathrm{PI}}$. It should be mentioned that $\tilde{V}_{(r)}^{\mathrm{PI}}$ does not necessarily improve the estimation of the SE. For example, if we consider a balanced RSS study, both estimators $\tilde{V}_{(r)}^{\mathrm{PI}}$ and $\widehat{V}_{(r)}^{\mathrm{PI}}$ gives the same SE estimator, namely

$$\widehat{V}_{\mu}^{\mathrm{PI}} = \frac{1}{mH^2} \sum_{r=1}^{H} \widehat{V}_{(r)}^{\mathrm{PI}}.$$

We conclude this paper by pointing out two merits of our proposed methods. First, the plug-in estimators are well defined with any population CDF estimate including those from SRS or reported in the previous literature. Thus, it does not require a pilot RSS study. They can also be applied directly to judgment post-stratification (e.g.,MacEachern et al. 2004, Du and MacEachern 2008, Wang et al. 2006), which is an alternative to RSS.

References

Chen, H., Stasny, E., and Wolfe, D.A. (2005). Ranked set sampling for efficient estimation of a population proportion. *Statistics in Medicine*, **24**, 3319-3329.

Chen, H., Stasny, E., and Wolfe, D.A. (2006). Unbalanced ranked set sampling for estimating a population proportion. *Biometrics*, **62**, 150-158.

Chen, Z. (1999). Density estimation using ranked set sampling data. *Environmental and Ecological Statistics* **6**, 135-146.

Chen, Z., Bai, Z., and Sinha, B. K. (2003). *Ranked Set Sampling: theory and applications.* Springer. New York.

David, H.A. and Nagaraja, H.N. (2003). *Order Statistics. 3rd edition.* John Wiley and Sons. New York.

Du, J. and MacEachern, S.N. (2008). Judgement Post-Stratification for Designed Experiments. *Biometrics* **64**, 345-354.

Kvam, P.H. and Samaniego, F.J. (1994). Nonparametric maximum likelihood estimation based on ranked set samples. *Journal of the American Statistical Association* **89**, 526-537.

Lacayo, H., Neerchal, N. K., and Sinha, B. K. (2002). Ranked set sampling from a dichotomous population. *Journal of Applied Statistical Science*, **11**, 83-90.

MacEachern, S.N., Óztúrk, O.,Wolfe, D.A., and Stark, G.V. (2002). A new ranked set sample estimator of variance. *Journal of the Royal Statistical Society-Series B*, **64**, 177-188.

MacEachern, S. N., Stasny, E. A., and Wolfe, D. A. (2004). Judgment poststratification with imprecise rankings. *Biometrics*, **60**, 207-215.

Mode, N. A., Conquest, L. L., and Marker, D. A. (1999). Ranked set sampling for ecological research: Accounting for the total costs of sampling. *Environmetrics*, **10**, 179-194.

Óztúrk, O. (2007). Statistical inference under a stochastic ordering constraint in ranked set sampling. *Journal of Nonparametric Statistics*, **19**, 131-144.

Óztúrk, O. (2008). Statistical inference in the presence of ranking error in ranked set sampling. *The Canadian Journal of Statistics*, 35, 577-594

Platt, W.J., Evans, G.M., and Rathbun, S.L. (1988). The population dynamics of long-lived conifer (Pinus palustris). *American Natrualist*, **131**, 491-525.

Ross, N. P. and Stokes, L. (1999). Editorial: Special issue on statistical design and analysis with ranked set sampling. *Environmental and Ecological Statistics*, **6**, 5-9.

Stokes, S.L. and Sager, T.W. (1988). Characterization of a ranked set sample with application to estimating distribution functions. *Journal of the American Statistical Association,* **83**, 374–381.

Wang, X., Stokes, L., Lim, J., and M. Chen (2006). Concomitants of Multivariate Order Statistics With Application to Judgment Poststratification. *Journal of the American Statistical Association*, **101**, 1693-1704.

Wang, X., Lim, J., and Stokes, L. (2008). A nonparametric mean estimator for judgment poststratified data. *Biometrics*, **64**, 355-363.