

Judgement Post-stratification with Multiple Rankers

Lynne STOKES, Xinlei WANG and Min CHEN*

February 2006

Abstract

Judgement post-stratification is a method of data collection in which the members of a random sample are stratified after selection by ranking each one among its own randomly chosen comparison sample. The original random sample units are measured, whereas those in the comparison sample are not. An estimator of the mean that is similar to that from a ranked set sample can be constructed from this sample, and it has similar properties. That is, if ranking is reasonably accurate and measurement is expensive compared to ranking, this estimation procedure improves efficiency in estimation of the mean. In this paper, we develop several estimators of the mean that make use of judgement ranks from more than one ranker from a judgement post-stratified sample. We compare their performance through simulation. We also provide insights about when extra rankers are useful.

Keywords: Ranked Set Sampling; Best Linear Unbiased Estimator; Raking; Bootstrapping.

*Lynne Stokes is Professor, and Xinlei Wang is Assistant Professor, Department of Statistical Science, Southern Methodist University, 3225 Daniel Avenue, P O Box 750332, Dallas, Texas 75275-0332, slstokes@mail.smu.edu and swang@mail.smu.edu. Min Chen is Ph.D candidate, McCombs School of Business, University of Texas at Austin, minchen@mail.utexas.edu.

1 Introduction

In this paper, we examine a method of data collection studied by MacEachern, Stasny and Wolfe (2004). In their procedure, each of a set of measured observations is ranked by eye or some other relatively inexpensive method among its own set of unmeasured observations. The assigned ranks provide auxiliary information about the measured units. Sampling theory suggests a variety of ways that this information could be used, but when it is used to form post-strata, MSW refer to the procedure as judgement post-stratification (JP-S). They illustrate that when rankers are allowed to express uncertainty about ranks, which they call imprecise ranking, rather than being forced into stating an exact ordering, then this information can be used to construct estimators that perform better. They also show one example simulation in which use of multiple rankers can provide better estimation than a single ranker.

Judgement post-stratification proceeds as follows. A sample of size n is selected at random, and the characteristic of interest is measured for each. Then a random sample of $H - 1$ additional units is selected and compared with the first measured observation, and a rank (or ranks, if there is more than one ranker) assigned to it. A second random sample of $H - 1$ observations is selected and compared with the second measured observation, and a rank is assigned to it. The procedure continues until all n observations have been ranked among its own set of $H - 1$ randomly chosen units.

This method is similar, both in practical implementation and in theoretical development, to ranked set sampling (RSS). In both cases, an independent sample of order statistics, or judgement order statistics if ranking is imperfect, is available for analysis, along with information about the (judgement) rank of each. Judgement post-stratification differs in that the number of measured judgement order statistics of each rank is random, while for ranked set sampling, it is typically fixed in advance. The earliest implementations of RSS were designed for estimating the mean (McIntyre 1952, Takahashi and Wakimoto 1968). Both RSS and JP-S provide gains in efficiency for estimating the mean when measurement

is much more expensive than ranking. Applications in agriculture (e.g., Cobby et al. 1985), forestry (Dell and Clutter 1972), and environmental assessment (e.g., Kvam 2003) have been most frequently reported.

MSW (2004) point out that an advantage of JP-S over RSS is that if one ignores the ranking information, the measured observations can be analyzed using conventional statistical methods, as they are a standard random sample. Ranked set samples are not, since they are composed of independent (judgement) order statistics. For those fearing that subject matter journals will discourage nonstandard statistical analyses, or who anticipate using some advanced data analysis methods not yet developed for ranked set samples, an underlying random sample is attractive. The second advantage of JP-S is, that it is possible to allow more than one ranker to provide ranking information on the same measured unit, while it is impossible in RSS (unless the rankers agree), since the unit to be measured is determined by the specified rank. We will see that multiple rankers, when they have some ranking skill and are not identical, can provide information that allows better estimation of the mean than that of a single ranker. Another advantage is that JP-S might allow for a large number of ranking classes (i.e., H) in some applications, since we only need the rank of each fully measured unit among its comparison group. In RSS, by contrast, we need to determine, within each set which is the one with a given rank, which is more difficult when H is large and the rank is close to $H/2$.

The purpose of this paper is to examine how the information from multiple rankers can be used, and to determine when it is worthwhile to use them. In Section 2, we discuss three estimators of the mean that can be computed from a sample that is judgement post-stratified by m rankers. In Section 3, we use normal examples assuming large samples to provide insights about how much and under what circumstances an advantage from additional rankers can be expected. Section 4 reports simulation results for comparing those estimators based on simulated data from a parametric model with different types of distributions. The estimators are also compared on a real data example. A discussion follows in Section 5.

2 Estimators

Let $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)^\top$ be an i.i.d random sample from a population of interest with mean μ and variance σ^2 . Let $\mathbf{Y} = (y_1, y_2, \dots, y_n)^\top$, a realization of \mathbf{Y} , denote values of those fully measured units in a judgement post-stratified sample. Suppose there are m rankers; each of them can be either perfect or imperfect in judgement ranking. Define $I_{ih}^{(j)} = 1$ if Ranker j assigns the rank h to y_i among its $H-1$ comparison units, otherwise $I_{ih}^{(j)} = 0$, for $i = 1, \dots, n$, $j = 1, \dots, m$ and $h = 1, \dots, H$; the vector of ranks is denoted by $\mathbf{R}_i = (R_{i1}, \dots, R_{im})$, where R_{ij} is the rank assigned to y_i by Ranker j . There are thus H^m post-strata jointly grouped by the ranks $\mathbf{R} = (\mathbf{R}_i)_{i=1}^n$. Let $\text{PS}_{\mathbf{r}}$ denote the post-stratum in which $\mathbf{R}_i = \mathbf{r}$, and $\pi_{\mathbf{r}}$, $\mu_{\mathbf{r}}$, $n_{\mathbf{r}}$, and $\bar{Y}_{[\mathbf{r}]}$ denote the probability, mean, number and sample mean of observations falling in $\text{PS}_{\mathbf{r}}$.

In what follows, we discuss three methods to estimate the mean μ , using information from multiple rankers.

2.1 The MSW method

When assessments of ranks are available from m rankers for each y_i , MacEachern et al. (2004) proposed the estimator

$$\hat{\mu}_M^{(m)} = \frac{1}{H} \sum_{h=1}^H \frac{\sum_{i=1}^n y_i \hat{p}_{ih}}{\sum_{i=1}^n \hat{p}_{ih}}, \quad (1)$$

where $\hat{p}_{ih} = \sum_{j=1}^m I_{ih}^{(j)} / m$ is the proportion of rankers who classify y_i as having rank h . When there is only one ranker (i.e., $m = 1$), (1) becomes the JP-S estimator studied in MacEachern et al. (2004):

$$\hat{\mu} = \frac{1}{H} \sum_{h=1}^H \frac{\sum_{i=1}^n y_i I_{ih}}{\sum_{i=1}^n I_{ih}}. \quad (2)$$

The estimator in (1) is easy to compute. However, it is difficult to obtain its analytical properties in the general case. For simplicity, we restrict attention to the case of $m = 2$

rankers and investigate the question of whether using an extra ranker improves estimation of the mean. We hope this can shed light on cases with more than two rankers.

Let $PS_{s,t}$ denote the post-stratum in which Ranker 1 assigns rank s ($R_{i1} = s$) and Ranker 2 assigns rank t ($R_{i2} = t$). The sample mean of Y for $PS_{s,t}$ is

$$\bar{Y}_{[s,t]} = \frac{\sum_{i=1}^n I_{is}^{(1)} I_{it}^{(2)} y_i}{n_{s,t}}, \quad 1 \leq s, t \leq H.$$

By rearranging the terms of (1), one can write $\hat{\mu}_M^{(2)}$ as a weighted average of the sample means of the post-strata, namely,

$$\hat{\mu}_M^{(2)} = \sum_{s=1}^H \sum_{t=1}^H \hat{w}_{s,t} \bar{Y}_{[s,t]},$$

where

$$\hat{w}_{s,t} = \frac{1}{H} \left(\frac{n_{s,t}}{n_{s\cdot} + n_{\cdot s}} + \frac{n_{s,t}}{n_{t\cdot} + n_{\cdot t}} \right),$$

and $n_{s\cdot}$ ($n_{\cdot t}$) denotes the number of sample units for which $R_{i1} = s$ ($R_{i2} = t$).

First note that $\hat{\mu}_M^{(2)}$ is a consistent estimator of μ . This follows from the observations that $n_{s,t} \sim \text{Binomial}(n, \pi_{s,t})$, $n_{s\cdot}$ ($n_{\cdot t}$) $\sim \text{Binomial}(n, 1/H)$, and $E(\bar{Y}_{[s,t]}) = \mu_{[s,t]}$. Then $\hat{w}_{[s,t]} \rightarrow \pi_{s,t}$ and $\hat{\mu}_M^{(2)} \rightarrow \sum_{s=1}^H \sum_{t=1}^H \pi_{s,t} \mu_{[s,t]} = \mu$ as $n \rightarrow +\infty$.

We assess the variance of $\hat{\mu}_M^{(2)}$ by conditioning on the realized post-stratum sample sizes. This yields

$$\text{Var}(\hat{\mu}_M^{(2)} | n_{s,t} > 0; s = 1, \dots, H; t = 1, \dots, H) = \sum_{s=1}^H \sum_{t=1}^H \hat{w}_{s,t}^2 \frac{\sigma_{[s,t]}^2}{n_{s,t}}.$$

To measure the marginal value of the second ranker, we need to compare $\hat{\mu}_M^{(2)}$ to $\hat{\mu}$ ranking

with only Ranker 1, as in (2) . The (conditional) variance of the latter is

$$\text{Var}(\hat{\mu} | n_{s\cdot} > 0; s = 1, \dots, H) = \frac{1}{H^2} \sum_{s=1}^H \frac{\sigma_{[s,\cdot]}^2}{n_{s\cdot}},$$

where $\sigma_{[s,\cdot]}^2 = \text{Var}(Y_i | R_{i1} = s)$. The ratio of the variances of $\hat{\mu}$ and $\hat{\mu}_M^{(2)}$ is denoted by RE .

Then

$$ARE = \lim_{n \rightarrow \infty} RE = \frac{\sum_{s=1}^H \sigma_{[s,\cdot]}^2 / H}{\sum_{s=1}^H \sum_{r=1}^H \pi_{s,t} \sigma_{[s,t]}^2}. \quad (3)$$

$ARE \geq 1$ since

$$\begin{aligned} \sigma_{[s,\cdot]}^2 &= E_S [\text{Var}(Y_i | R_{i1} = s, R_{i2} = t)] + \text{Var}_S [E(Y_i | R_{i1} = s, R_{i2} = t)] \\ &= H \sum_{t=1}^H \pi_{s,t} \sigma_{[s,t]}^2 + \left[H \sum_{t=1}^H \pi_{s,t} \mu_{[s,t]}^2 - H^2 \left(\sum_{t=1}^H \pi_{s,t} \mu_{[s,t]} \right)^2 \right] \\ &\geq H \sum_{t=1}^H \pi_{s,t} \sigma_{[s,t]}^2. \end{aligned}$$

Thus, using an extra ranker with the estimator (1) is beneficial at least in an asymptotic sense.

2.2 The BLUE of rankers' JP-S estimators

For a JP-S sample with multiple rankers, each ranker can have his own estimator of μ in the form of (2) based on the set of ranks he assigns to \mathbf{Y} . We propose an estimator which linearly combines the estimators from the m rankers to form a new one having minimum mean squared error (MSE).

Let $\hat{\mu}_j$ denote Ranker j 's estimator of the form (2) and w_j denote the weight associated with Ranker j , for $j = 1, \dots, m$. We begin with

$$\tilde{\mu}^{(m)} = \mathbf{w}^T \hat{\boldsymbol{\mu}} \quad (4)$$

where $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \dots, \hat{\mu}_m)^\top$, and $\mathbf{w} = (w_1, \dots, w_m)^\top$. Since each $\hat{\mu}_j$ is unbiased, $\tilde{\mu}^{(m)}$ is unbiased if $\sum_{j=1}^m w_j = 1$. Let $\boldsymbol{\Sigma}$ be the covariance matrix of $\hat{\boldsymbol{\mu}}$ so $Var(\tilde{\mu}^{(m)}) = \mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w}$. Our objective is to

$$\min_{\mathbf{w}} Var(\tilde{\mu}^{(m)}) \quad \text{s.t.} \quad \mathbf{w}^\top \cdot \mathbf{1} = 1$$

where $\mathbf{1}$ is a vector of all 1's. The solution to this optimization problem is given by

$$\mathbf{w}_0 = \frac{\boldsymbol{\Sigma}^{-1} \mathbf{1}}{\mathbf{1}^\top \boldsymbol{\Sigma}^{-1} \mathbf{1}} \quad (5)$$

and the minimized variance is

$$Var(\tilde{\mu}_B^{(m)}) = \frac{1}{\mathbf{1}^\top \boldsymbol{\Sigma}^{-1} \mathbf{1}} \quad (6)$$

where $\tilde{\mu}_B^{(m)}$ denotes the best linear unbiased estimator (BLUE), i.e., $\tilde{\mu}^{(m)}$ with the optimal weights \mathbf{w}_0 . Based on (5) and (6), the optimal weight associated with Ranker j is the sum of elements in the j th row of $\boldsymbol{\Sigma}^{-1}$, divided by the sum of all elements of $\boldsymbol{\Sigma}^{-1}$; the minimum variance is the inverse of the sum of all elements of $\boldsymbol{\Sigma}^{-1}$. Obviously, $Var(\tilde{\mu}_B^{(m)}) \leq Var(\hat{\mu}_j)$ for any j , so that $\tilde{\mu}_B^{(m)}$ improves the JP-S estimator from any single ranker. Also, adding an extra ranker is beneficial because $Var(\tilde{\mu}_B^{(m+1)}) \leq Var(\tilde{\mu}_B^{(m)})$ since in $\tilde{\mu}_B^{(m)}$, we can set $w_{m+1} = 0$ to get a $\tilde{\mu}^{(m+1)}$ that is not optimal.

Calculating the BLUE $\tilde{\mu}_B^{(m)}$ requires the covariance matrix $\boldsymbol{\Sigma}$ of $\hat{\boldsymbol{\mu}}$. It can be verified using the delta method that

$$\text{var}(\hat{\mu}_j) \approx \frac{1}{n} \left[\sigma^2 - \frac{1}{H} \sum_{h=1}^H (\mu_{[h]}^j - \mu)^2 \right] \left[1 + \frac{H-1}{n} \right] \quad (7)$$

$$\text{cov}(\hat{\mu}_k, \hat{\mu}_l) \approx \frac{1}{n} \sum_{s=1}^H \sum_{t=1}^H p_{[s,t]}^{k,l} \left\{ \left[\mu_{[s]}^k - \mu_{[s,t]}^{k,l} \right] \left[\mu_{[t]}^l - \mu_{[s,t]}^{k,l} \right] + \left(\sigma_{[s,t]}^{k,l} \right)^2 \right\} \quad (8)$$

where $\mu_{[h]}^j = E(Y_i | I_{ih}^{(j)})$, $p_{[s,t]}^{k,l} = E(I_{is}^{(k)} = 1, I_{it}^{(l)} = 1)$, $\mu_{[s,t]}^{k,l} = E(Y_i | I_{is}^{(k)} = 1, I_{it}^{(l)} = 1)$ and $\left(\sigma_{[s,t]}^{k,l} \right)^2 = Var(Y_i | I_{is}^{(k)} = 1, I_{it}^{(l)} = 1)$, for $h, s, t = 1, \dots, H$ and $j, k, l = 1, \dots, m$. The proof is lengthy and omitted for brevity. These approximations work well for large n .

Formulas (7) and (8) can be computed under certain distributional assumptions with known parameters. This is quite restrictive in practice. A nonparametric approach is to substitute sample proportions, means, and variances in the post-strata as surrogates of those $p_{[s,t]}^{k,l}$, $\mu_{[h]}^j$, $\mu_{[s,t]}^{k,l}$ and $\left(\sigma_{[s,t]}^{k,l}\right)^2$. However, we have found that the resulting weights are unstable, and the estimator performs poorly. To mitigate this difficulty, we propose a bootstrapping procedure to compute Σ , which is outlined by the following steps .

1. Take a sample \mathbf{D}' of size q ($q < n$) with replacement from the data $\mathbf{D} = (y_i, \mathbf{R}_i)_{i=1}^n$ collected by judgement post-stratification.
2. For each ranker, calculate the JP-S estimator $\hat{\mu}_j$ based on \mathbf{D}' . Let $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \dots, \hat{\mu}_m)^T$.
3. Repeat the above steps A times.
4. Calculate the sample covariance $\hat{\Sigma}$ matrix for $\hat{\boldsymbol{\mu}}^{(1)}, \dots, \hat{\boldsymbol{\mu}}^{(A)}$. This will be used to approximate Σ .

This procedure requires no distributional assumptions, and is easy to implement. We have found that for large n , (6) provides a good approximation to the variance of $\tilde{\mu}_B^{(m)}$ with the bootstrapping procedure.

2.3 The raking method

From sampling theory, we know that post-stratification can improve estimation of the mean when the proportion of units in each post-stratum is known from some source outside the sample. Commonly, post-strata are defined by cross-classifying units by several variables. If only the marginal proportions are known, a method known as raking (Deming and Stephan 1940) can be used for estimating the proportions in the cross-classified cells. Raking involves iterative proportional fitting of the cell proportions to successively match the known one-dimensional marginal probabilities.

In judgement post-stratification, we consider the rank assignment of each ranker as one post-stratifying variable, and their joint ranks, denoted by \mathbf{r} , to define the post-strata. We do not know the probability of a randomly selected unit falling in any post-stratum unless assumptions are made about both the distribution of Y and the ranking process. But we do know the probability that each ranker classifies a unit into each category; i.e., $Pr[R_{ir} = s] = 1/H$ for all i, r , and s . So we propose to use raking for estimating the cell probabilities non-parametrically, resulting in an estimator of the form

$$\hat{\mu}_R^{(m)} = \sum_{\mathbf{r}} \hat{\pi}_{\mathbf{r}}(\mathbf{n}) \bar{Y}_{[\mathbf{r}]}$$

where the summation is over all H^m realizations of the rank vector, \mathbf{n} is the random vector containing the counts of Y in the H^m post-strata; and $\hat{\pi}_{\mathbf{r}}(\cdot)$ is the estimate of the cell probability based on raking.

Both $\hat{\mu}_R^{(m)}$ and $\hat{\mu}_M^{(m)}$ can be thought of as weighted averages of estimated post-stratum cell means, where the weights are nonparametric estimates of cell proportions. They differ in the way these cell proportions are estimated. Another way the two estimators differ is that $\hat{\mu}_R^{(m)}$ cannot be calculated when there exist one or more empty ranking classes for some ranker, as raking is then not possible. So $\hat{\mu}_R^{(m)}$ is a feasible estimator only when H is small relative to n so that empty ranking classes are unlikely to occur.

3 When are extra rankers helpful?

Here, we are interested in the question when extra rankers are helpful. To aid intuition, we use examples that assume large samples from a multivariate normal distribution. Our discussion is based on the estimator $\hat{\mu}_M^{(m)}$. For the other two estimators, the patterns appear similar, so are not reported here.

We first consider the effect of ranker similarity and quality. Suppose ranking is done via concomitant variables X_1 and X_2 . Let Y, X_1, X_2 follow a multivariate normal distribution,

Figure 1: Asymptotic Efficiency of $\hat{\mu}_M$ (2 rankers, equally effective) to $\hat{\mu}$ (1 ranker)

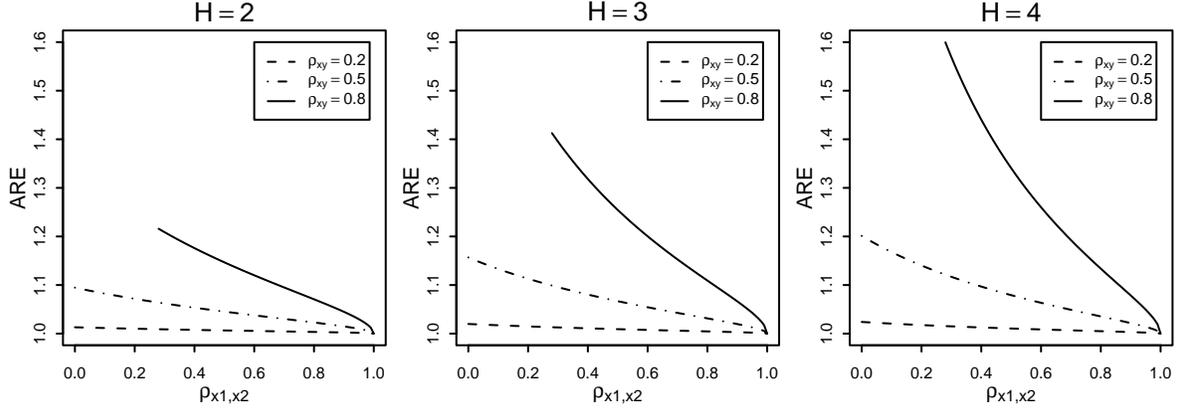
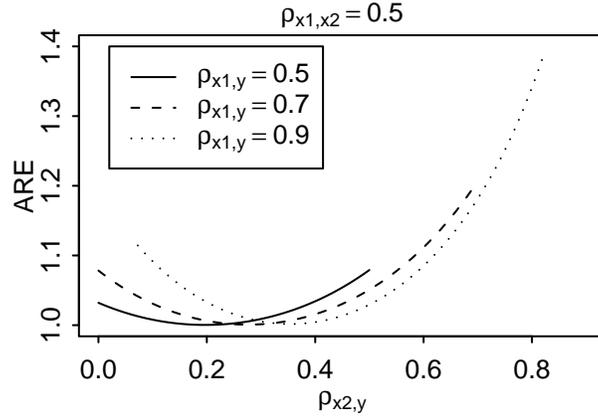


Figure 2: Asymptotic Efficiency of $\hat{\mu}_M$ (2 rankers, ranker 2 worse than ranker 1) to $\hat{\mu}$ (best ranker), $H = 4$



so that $\rho_{X_1,Y}$ and $\rho_{X_2,Y}$ measure the effectiveness of ranker 1 and 2, respectively, and ρ_{X_1,X_2} measures the similarity of rankers 1 and 2. We shall compare the performance of $\hat{\mu}_M^{(2)}$ based on the large-sample property in (3) for various values of ρ 's, where variances were calculated using numerical quadrature (see Wang and Stokes 2005).

Figure 1 shows the *ARE* for two equally effective rankers (i.e, $\rho_{X_1,Y} = \rho_{X_2,Y} \equiv \rho_{X,Y}$) and the three values each of $\rho_{X,Y}$ and H . It shows that the gain from the second ranker can be substantial. It increases as the ranking quality or the number of ranking classes increases, and decreases as the two rankers become more similar. Note that the patterns of gain are similar for different H . This is true for all the other examples in this section, so we only consider $H = 4$ below.

Figure 3: Asymptotic Efficiency of $\hat{\mu}_M$ (2 rankers with one perfect) to $\hat{\mu}$ (perfect ranker), $H = 4$

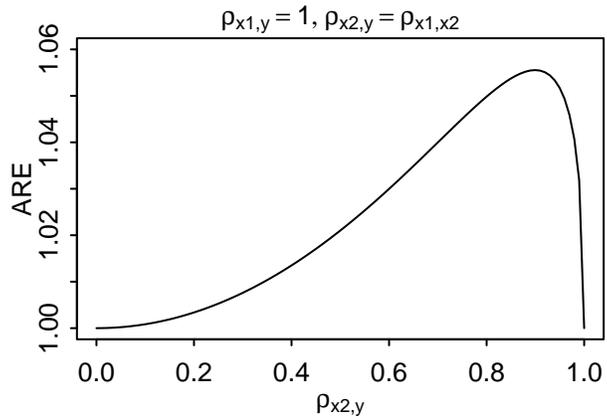


Figure 4: Asymptotic Efficiency of $\hat{\mu}_M$ (2 rankers with one independent of Y) to $\hat{\mu}$ (best ranker), $H = 4$

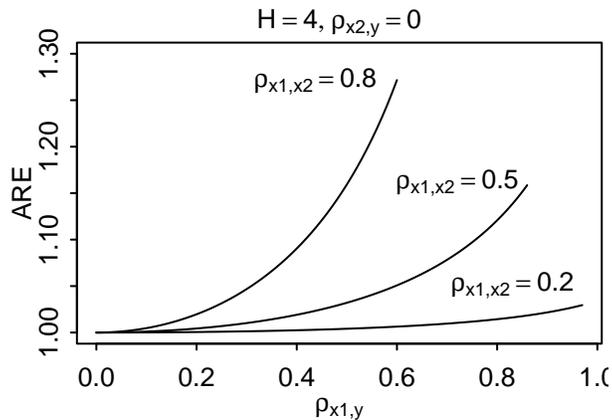


Figure 5: Simulated Efficiency of $\hat{\mu}_M$ to $\hat{\mu}$ (best ranker), ranking by concomitants, similar rankers, $H = 4$, normal model, $n = 100$

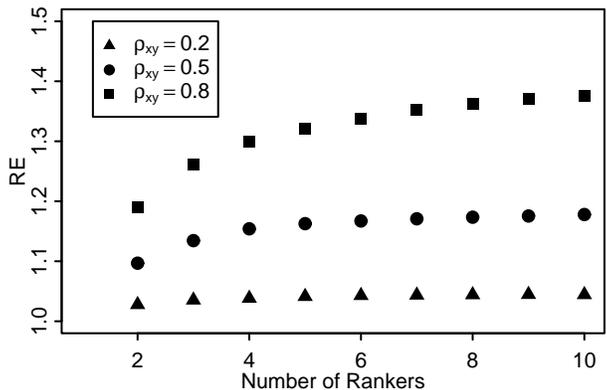


Figure 2 shows that a “bad” ranker can help a better one. Surprisingly, the gain is not monotonically increasing as the quality of the “bad” ranker increases; the advantage is lowest when she has a correlation of around .20–0.30 with Y .

The second ranker can be useful even when it might not be expected to. Figure 3 shows that if perfect ranking on Y is available through Ranker 1 (i.e., $\rho_{X_1,Y} = 1$ and $\rho_{X_2,Y} = \rho_{X_1,X_2}$), using the information from Ranker 2 is helpful in estimating μ except when she is too poor or too good in ranking. In the latter case she will be too similar to the better ranker and thus provides little additional information. Figure 4 shows that there can be benefit from Ranker 2 even when she is independent of Y (i.e., $\rho_{X_2,Y} = 0$). In this case, the advantage comes from the information Ranker 2 contains about Ranker 1. The figure also confirms that the benefit increases with ρ_{X_1,X_2} .

Figure 5 shows simulated efficiency for different numbers of equally effective rankers and the three values of $\rho_{X,Y}$. Here, we fix the sample size at $n = 100$, and let $\rho_{X_i,X_j} = \rho_{X_i,Y} \cdot \rho_{X_j,Y}$, which implies that $\rho_{X_i,X_j|Y} = 0$, for any i and j ; efficiency is defined as MSE of $\hat{\mu}_M^{(m)}$ to $\hat{\mu}$, where MSE is estimated from 10,000 replicates. In this case, using more rankers helps but not much after the number of rankers reaches about 4.

4 Comparison of the estimators

4.1 A simulation study

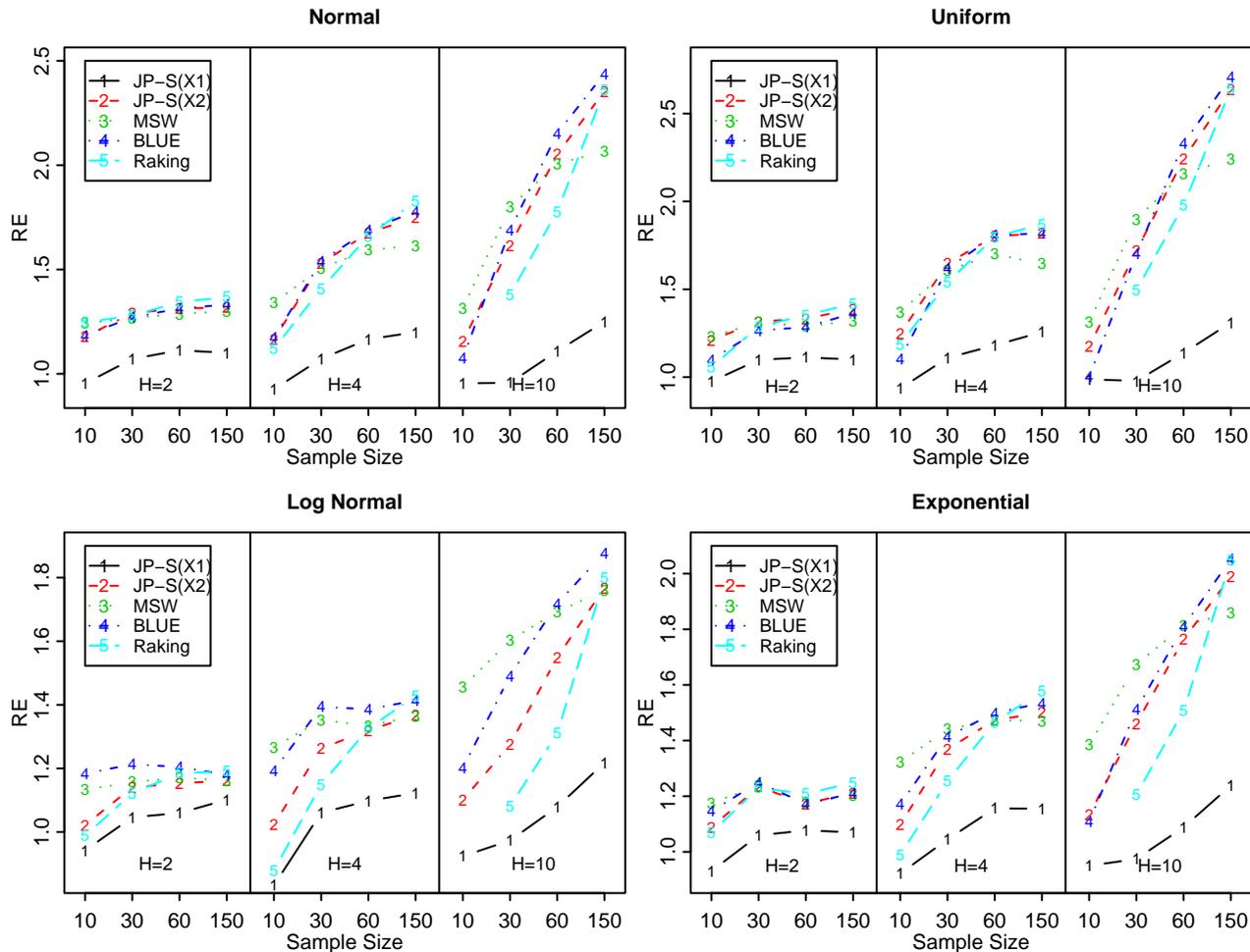
In order to study the behavior of the three estimators discussed in Section 2, we model the imperfect ranking process by regarding Y as the concomitant of a vector of ranking variables that can be accurately and easily measured. That is, we assume that Ranker j , $j = 1, \dots, m$, behaves as if he assesses the rank of Y_i by assigning it the true rank that some ranking variable X_{ij} has among its comparison group of size H . Further, assume $X_{ij} = Y_i + e_{ij}$, where $e_{ij} \sim N(0, \sigma_j^2)$ and e_{ij} is independent of Y_i for each i and j . We restrict attention to the case of $m = 2$ rankers in our numerical comparison for simplicity. So in the discussion

below, we omit the superscripts of the estimators $\hat{\mu}_M^{(m)}$, $\tilde{\mu}_B^{(m)}$ and $\hat{\mu}_R^{(m)}$.

We implemented two sets of simulations. In the first set we study two nonequally effective rankers, where we chose $Var(X_1) = 3.3$, $Var(X_2) = 0.5$, the correlations $\rho(X_1, Y) \approx 0.58$, $\rho(X_2, Y) \approx 0.88$, and $\rho(X_1, X_2) \approx 0.50$. Here, Ranker 2 is more effective than Ranker 1. In the second set we study two equally effective rankers, where we chose $Var(X_1) = Var(X_2) = 1$, the correlations $\rho(X_1, Y) = \rho(X_2, Y) \approx 0.71$ and $\rho(X_1, X_2) \approx 0.50$. For each case, we set H to be 2, 4, 10 and n to be 10, 30, 60, 150. We simulated Y from four types of distributions: normal, uniform, lognormal and exponential. We chose parameters their parameters to achieve the specified correlations and variances. When calculating $\tilde{\mu}_B$ from each sample, we used the bootstrapping method with $q = n/2$ and $A = 200$. Figure 6 and 7 report the simulated relative efficiency of the five estimators, $\hat{\mu}_M$, $\tilde{\mu}_B$, $\hat{\mu}_R$, $\hat{\mu}$ ranking with only Ranker 1, and $\hat{\mu}$ ranking with only Ranker 2 to the SRS estimator \bar{Y} for each setting. Here, efficiency is defined as the ratio of the variance of \bar{Y} to MSE of each estimator, where MSE is estimated from 10,000 replicates. We denote these five estimators MSW, BLUE, Raking, JP-S(X_1) and JP-S(X_2) respectively in the figures.

The results in Figure 6 show that in all cases, using two rankers is much better than using the bad one; and the performance with two rankers is better or comparable to that with the better one. This indicates that if the quality of rankers is unknown, combining is beneficial on average. For the lognormal distribution, the bad ranker can help the better one if using with $\hat{\mu}_M$ or $\tilde{\mu}_B$; furthermore, it appears that $\tilde{\mu}_B$ is better than $\hat{\mu}_M$ except for the cases with large H and small n . For the other three distributions, the bad ranker often provides little or slight help to the better one, except for using $\hat{\mu}_M$ with small n . This appears to contradict the results for normal data in Figure 2. But it is not since Figure 2 is based on asymptotic properties while Figure 6 is based on finite samples. Further, Figure 6 suggests that the efficiency of $\hat{\mu}_M$ over $\hat{\mu}$ is not a monotonically increasing function of n . So in practical situations when n is not very large, if we know which ranker is better, use $\hat{\mu}_M$ for small n and simply use the better ranker otherwise for normal, uniform and exponential

Figure 6: Simulated efficiency of the estimators, ranking by two nonequally effective rankers.

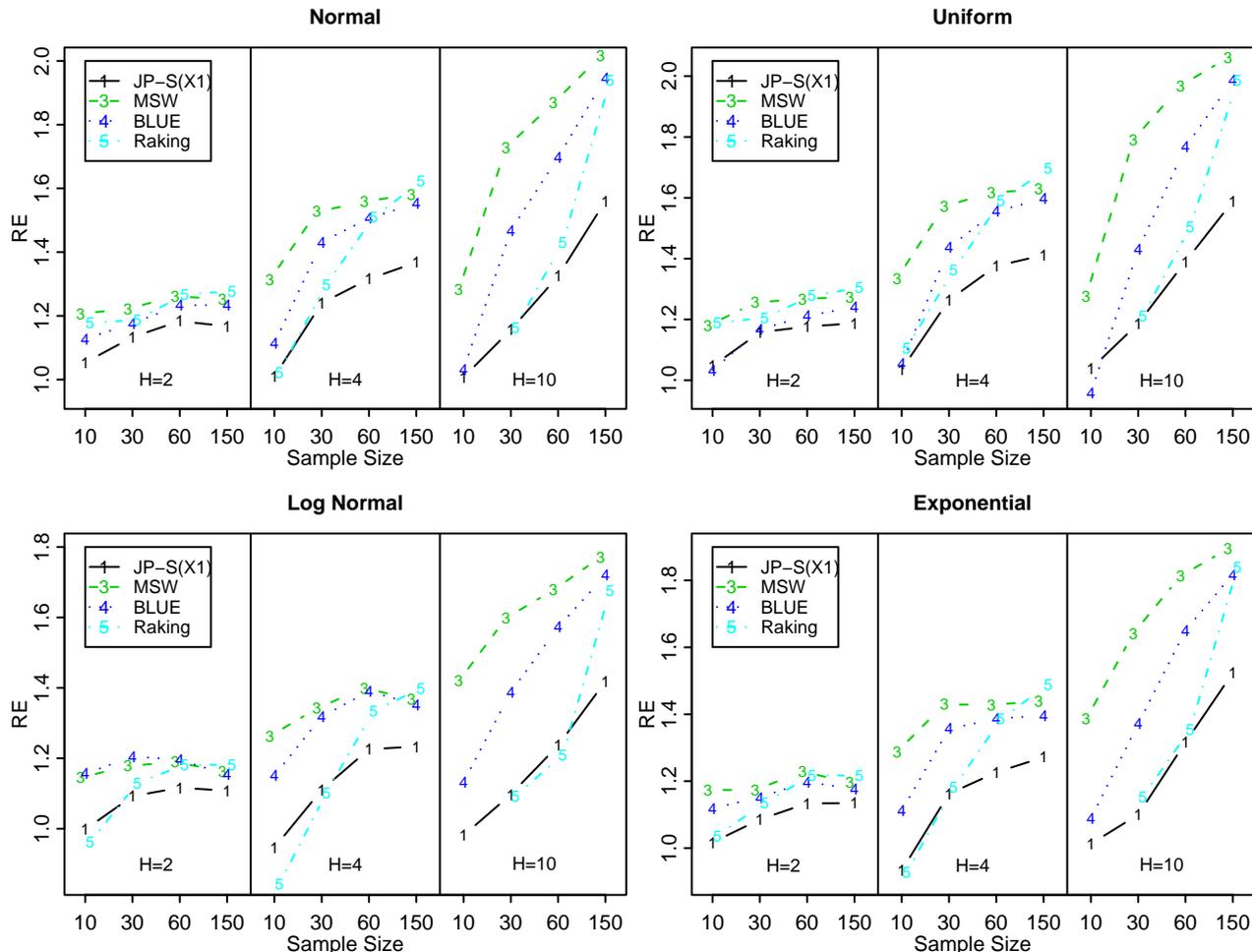


data.

Figure 7 shows for two equally effective rankers, $\hat{\mu}_M$ appears to be the best nearly in all the cases. This is not surprising since the MSW method treats each ranker equally. It prorates a measured value among the ranking classes receiving any “votes” from a ranker. So its best scenario is that all the rankers are of the same quality.

Figure 6 and 7 also show that overall the raking method does not work well. It cannot be applied when empty ranking classes occur, including the cases of $H = 10$ and $n = 10$. Although it appears to be the best for large n and small H , the improvement over the second best estimator is often small.

Figure 7: Simulated efficiency of the estimators, ranking by two equally effective rankers



4.2 An empirical comparison: adjusted brain weights of mammals

Following Section 5 in MacEachern et al. (2004), we use a data set that consists of allometric measurements for 62 species of mammals, and set our goal as estimation of the mean of Y , the log of adjusted brain weight, defined as $Y = \log\{\text{brain weight}/(\text{body weight})^{2/3}\}$.

The data were randomly grouped into 20 sets of 3 species (2 species were randomly selected and discarded for this purpose). We assume that each of the 20 sets represent three independent draws from a large population of species. Each set was presented to two rankers that assigned ranks within the set independently. The rankers made judgements based on the conjecture that a “clever” species tends to have a large adjusted brain weight. The data

generated are described by Table 1.

Table 1: A judgement post-stratified sample (2 rankers, $H = 3$) based on the mammals data

mammals	set	Ranker1	Ranker2	y	mammals	set	Ranker1	Ranker2	y
Genet	1	2	1	2.63	Cat	11	2	2	2.45
Rat	1	3	2	1.49	Human	11	1	1	4.43
Cow	1	1	3	1.95	Rabbit	11	3	3	1.88
African giant pouched rat	2	3	2	1.89	Artic fox	12	1	1	2.98
Kangaroo	2	2	3	1.66	Nine-banded armadillo	12	3	3	1.54
Red fox	2	1	1	2.96	Brazilian tapir	12	2	2	1.75
Lesser short-tailed shrew	3	3	3	1.57	Tree hyrax	13	3	3	2.05
Jaguar	3	1	1	1.99	Pig	13	1	2	1.69
Rock hyrax-a	3	2	2	2.70	Guinea pig	13	2	1	1.68
Baboon	4	1	1	3.62	Water opossum	14	3	3	0.53
Phalanger	4	2	2	2.11	Rhesus monkey	14	1	1	3.91
Sheep	4	3	3	2.49	African elephant	14	2	2	2.78
Gorilla	5	1	1	2.45	N.A. opossum	15	3	3	1.49
Giant armadillo	5	3	3	1.66	Roe deer	15	1	2	2.79
Yellow-bellied marmot	5	2	2	1.90	Okapi	15	2	1	2.51
Echidna	6	3	2	2.49	Donkey	16	1	2	2.55
Owl monkey	6	1	1	3.23	Mountian beaver	16	3	3	1.89
Giraffe	6	2	3	2.34	Horse	16	2	1	2.31
Grey wolf	7	1	1	2.39	Tree shrew	17	3	3	2.43
Big brown bat	7	3	3	1.31	Galago	17	1	1	2.68
Goat	7	2	2	2.53	Golden hamster	17	2	2	1.41
Little brown bat	8	1	3	1.68	European hedgehog	18	3	3	1.41
Tenrec	8	2	1	1.03	Ground squirrel	18	1	1	2.91
Desert hedgehog	8	3	2	1.27	Rock hyrax-b	18	2	2	2.19
Asian elephant	9	3	1	3.21	Raccoon	19	1	2	2.70
E. American mole	9	2	2	1.91	Mouse	19	2	1	1.60
Verbet	9	1	3	3.11	Musk shrew	19	3	3	0.92
Chinchilla	10	2	2	2.43	Star-nosed mole	20	3	3	1.88
Grey seal	10	1	3	2.82	Slow loris	20	2	2	2.30
Mole rat	10	3	1	2.50	Patras monkey	20	1	1	3.21

To compare the estimators with one or two rankers, we conducted a simulation. In each iteration, a sample of $n = 20$ species was selected, with one species from each set. The following table summarizes the results based on 10,000 iterations. Again, the BLUE was calculated using bootstrapping with $q = 10$.

Table 2: Comparing simulated relative efficiency of the estimators to the SRS estimator \bar{Y}

Estimator	JPS(Ranker 1)	JPS(Ranker 2)	MSW	BLUE	Raking
RE	1.30	1.06	1.54	1.44	1.17

Table 2 shows that combining the two rankers using either the MSW or BLUE method definitely has a value in improving estimation of the mean. In this example, Ranker 1 is much better than Ranker 2; the distribution of Y is roughly symmetric. It is interesting to

observe the MSW method performed better than the BLUE method here. Although ranking in this example was not done through concomitants, this result is consistent with what we find in Section 4.1; that is, for two nonequally effective rankers, MSW works best for small n , except for the case with lognormal data.

5 Discussion

In this paper, we have discussed three methods for combining information from multiple rankers, for use with judgement post-stratified samples. Through examples and simulation, we have provided insights about when it is worthwhile to use extra rankers and which method to use. We show that when rankers are not identical, there can be considerable benefit in having more than one. Especially in applications where the quality of rankers is hard to assess, combining can help avoid getting the worst estimation and achieve similar or better performance than the best ranker. Among the three estimators, the MSW method was the best when rankers are similarly effective. The raking method generally performed poorly. The BLUE method was found useful for lognormal data.

Finally, we should mention several directions for future investigation. First, we have shown that multiple rankers are useful for estimating the mean. They might lead to better estimation of other parameters, too. Second, using multiple rankers provides optimization opportunities that take into account quality of rankers, number of rankers, sample size and number of ranking classes, etc. Last, the same idea can be applied to ranked set sampling, with one primary ranker to specify units to measure and others to provide auxiliary information. This could be still beneficial.

References

Cobby, J., Ridout, M., Bassett, P., and Large, R. (1985). An investigation into the use of ranked set sampling on grass and grass-clover swards. *Grass and Forage Science*, 40:257–

- Dell, T. and Clutter, J. L. (1972). Ranked set sampling theory with order statistics background. *Biometrics*, 28(2):545–555.
- Deming, W. E. and Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4):427–444.
- Kvam, P. H. (2003). Ranked set sampling based on binary water quality data with covariates. *Journal of Agricultural, Biological and Environmental Science*, 8(3):271–279.
- MacEachern, S. N., Stasny, E. A., and Wolfe, D. A. (2004). Judgement post-stratification with imprecise rankings. *Biometrics*, 60:207–215.
- McIntyre, G. A. (1952). A method for unbiased selective sampling, using, ranked sets. *Australian Journal of Agricultural Research*, 3(4):385 – 390.
- Takahashi, K. and Wakimoto, K. (1968). On the unbiased estimates of the population mean based on the sample stratified by means of ordering. *Annals of Institute of Statistical Mathematics*, 20:1–31.
- Wang, X. and Stokes, S. L. (2005). Moments of bivariate order statistics for the normal distribution. Technical report, Southern Methodist University, Dallas, Texas.