Statistica Sinica **25** (2015), 1691-1712
doi:http://dx.doi.org/10.5705/ss.2014.163

DISTRIBUTION FREE TWO-SAMPLE METHODS FOR JUDGMENT POST-STRATIFIED DATA

Omer Ozturk

The Ohio State University

Abstract: A distribution-free procedure is developed to test a stochastic order relation between two distributions based on judgment post-stratified (JPS) data. The proposed inference relies on Mann-Whitney rank sum statistics. A first class of tests constructs test statistics by comparing all units in both samples, while a second class first stratifies the data into judgment classes and then constructs a rank sum statistic in each stratum, with the final test statistic constructed from a linear combination of these within-judgment class rank sum statistics. Distributional properties of the testing procedures are investigated. The null distributions of the test statistics in the first class depend on the quality of ranking information while the null distributions of the test statistics in the second class are distribution-free for any sample sizes, regardless of the quality of ranking information. Both tests have higher efficiencies than corresponding tests based on a simple random sample rank sum statistic. For large samples, testing procedures in the first and second classes are equivalent, respectively, to Bohn-Wolfe and Fligner-MacEachern testing procures in a ranked set sampling design.

Key words and phrases: Calibration, imperfect ranking, Mann-Whitney, ranked-set sampling, rank sum test, stochastic order.

1. Introduction

There are many experimental settings where full, precise measurement of a unit is much more expensive than obtaining an informal rough measurement through a relative ranking in a small set. This rough measurement can be used to create a structure among fully measured units to increase the information content of the data. Ranked-set sampling (RSS) provides a set of rules to use these informal and rough measurements to create relative ranks of units in a set.

For the construction of a balanced RSS, one first specifies a set size H and a number of cycles c. One then selects $n \equiv Hc$ independent simple random samples (sets) of size H from a distribution F. These n sets contain a total of cH^2 experimental units. The H units in each of these n sets are independently ranked from smallest to largest without making any measurements. Since ranks are assigned without measurement, they are called judgment ranks here. After ranking the units in each set, we select the unit with judgment rank 1 from each

of the first c sets for measurement. We select the unit with judgment rank 2 from each of the next c sets for measurement, and so on until we select the units with judgment rank H from the last c sets for measurement. This process yields $n = cH$ measurements. A general form of RSS allows the number of measured values to vary from one rank to another. In this case, instead of using the same c in each rank, we define a sample size vector $\boldsymbol{n} = (n_1, \ldots, n_H)$ such that n_h specifies the number of units with rank h to be selected for measurement. Unbalanced ranked-set sample then consists of $n = \sum_{h=1}^{H} n_h$ independent judgment order statistics. For a general unbalanced RSS design, let $X_{[r_j]j}$; $1 \leq r_j \leq H$, be the measured observation in set j for the unit judged to be the r_j-th smallest in the set. The observations $X_{[r_j]j}$; $j = 1, \ldots, n$, with $n_h = \sum_{j=1}^{n} I(r_j = h)$, constitute a ranked set sample of size n from distribution F, where $I()$ is an indicator function. If $c = n_h = n/H$; $h = 1, \ldots, H$, RSS is a balanced design. In this paper, square brackets are used to denote that the ranking process may be in error. If the ranking process does not involve any error, the square brackets are replaced with round parentheses. Let $F_{[r_j]j}$ be the cumulative distribution function (cdf) of $X_{[r_j]j}$. If the ranking process is perfect, then $F_{[r_j]}(x) = F_{(r_j)}(x)$ is just the cdf of the r_j-th order statistic from a random sample of size H. If the ranking is made totally at random, then $F_{[r_j]}(x) = F(x)$.

In a two-sample setting, a second ranked set sample, $Y_{[s_j]j}$, $1 \leq s_j \leq Q$; $j = 1, \ldots, m$, of size m with set size Q can be constructed from a distribution $G(y) = F(y - \Delta)$. In a balanced ranked set sample, the cycle size for the Y-sample data is denoted by $z = m/Q$. Further details of RSS sampling designs in a two sample problem can be found in Fligner and MacEachern (2006) and Bohn and Wolfe (1992).

In an RSS sample, the ranking information is used prior to measurement to determine which ranked unit should be selected for measurement in each set. Hence the judgment rank r_j and measured observation $X_{[r_j]j}$ in set j cannot be separated. This could be a problem in certain settings where the data set is collected for a general purpose analysis and inferential procedures for a ranked set sample analysis have not been developed yet. To address this concern MacEachern, Stasny, and Wolfe (2004) introduced judgment post-stratified sampling. To construct a JPS sample from a single population, say F, the experimenter first selects a simple random sample of size n and measures all of them. For each measured unit X_i; $i = 1, \ldots, n$, an additional $H - 1$ units are selected to form a set of size H. Units in this set are ranked from smallest to largest without measurement and the rank (R_i) of the measured unit X_i is recorded. The JPS sample then contains n fully measured units $(X_i; i = 1, \ldots, n)$ and n ranks $(R_i; i = 1, \ldots, n)$ associated with these fully measured units. To avoid possible bias, a ranker is blinded to the unit on which the measurement is made. Further details in JPS

sampling design can be found in MacEachern, Stasny, and Wolfe (2004), Wang, Lim, and Stokes (2008), Stokes, Wang, and Chen (2007), Wang et al. (2006), Frey and Ozturk (2011), Ozturk (2012), Wang, Wang, and Lim (2012), Frey and Feeman (2012, 2013), and Ozturk (2013, 2014a,b).

In a similar fashion, a JPS sample (Y_j, W_j); $j = 1, \ldots, m$, with set size Q, is obtained from a second population with continuous cdf $G(y) = F(y - \Delta)$, where $W_1, \ldots, W_m$ are the ranks of the measured observations Y_j, $j = 1, \ldots, m$. The sample sizes (m and n) and set sizes (H and Q) in the Y- and X-sample data can be different.

One of the major difference between RSS and JPS samples is the nature of the sample sizes in the judgment classes. Let $\boldsymbol{N} = (N_1, \ldots, N_H)$ and $\boldsymbol{M} = (M_1, \ldots, M_Q)$ be the vectors of sample sizes of judgment classes in the X- and Y-samples, respectively. In JPS samples, both $\boldsymbol{N}$ and $\boldsymbol{M}$ are random vectors, $\boldsymbol{N}$ has a multinomial distribution with parameters n and $(1/H, \ldots, 1/H)$ and $\boldsymbol{M}$ has a multinomial distribution with parameters m and $(1/Q, \ldots, 1/Q)$. Since $\boldsymbol{N}$ and $\boldsymbol{M}$ are random vectors, it is highly possible that some N_h and M_q may be zero, especially for small sample sizes and statistical inference needs to account for this possibility. In ranked-set sampling, the sample size vectors $\boldsymbol{n} = (n_1, \ldots, n_H)$ and $\boldsymbol{m} = (m_1, \ldots, m_Q)$ are non-random, pre-determined constant vectors. Zero values of n_h or m_q could, of course, result from design choices to increase the efficiency of the inference; see, for example, Ozturk and Wolfe (2000a), Chen (2001), Kaur et al. (2002).

Two-sample distribution-free inference in RSS has been studied extensively in the literature. Bohn and Wolfe (1992) proposed a rank sum statistic to test the null hypothesis $H_0 : \Delta = 0$ against the alternative hypothesis $H_A : \Delta \neq 0$ under a perfect ranking assumption. The proposed test rejects the null hypothesis for extreme values of

$$BW = \sum_{h=1}^{H} \sum_{i=1}^{c} \sum_{k=1}^{Q} \sum_{j=1}^{z} \psi(X_{[h]i} - Y_{[k]j}) = \sum_{h=1}^{H} \sum_{k=1}^{Q} BW_{h,k},$$

where $BW_{h,k} = \sum_{i=1}^{c} \sum_{j=1}^{z} \psi(X_{[h]i} - Y_{[k]j})$ and $\psi(a) = 1$ if $a > 0$ and $\psi(a) = 0$ if $a \leq 0$. The null distribution of BW is distribution-free under either perfect or random ranking. Under random ranking, it has the same null distribution as the Mann-Whitney-Wilcoxon statistic based on simple random samples

$$MW = \sum_{i=1}^{n} \sum_{j=1}^{m} \psi(X_i - Y_j),$$

where $X_i; i = 1, \ldots, n$ and $Y_j; j = 1, \ldots, m$ are simple random samples from X- and Y-sample distributions F and G, respectively.

Fligner and MacEachern (2006) proposed a new distribution-free statistic to test the null hypothesis H_0 against the alternative hypothesis H_A. With set sizes in the X- and Y-samples equal ($H = Q$), their test statistic is

$$FM = \sum_{h=1}^{H} BW_{h,h}.$$

Under the null hypothesis $F_{[h]}(x) \equiv G_{[h]}(x)$, the null distribution of $BW_{h,h}$ is the same as the MW statistic based on simple random samples of sizes c and z. The null distribution for each $BW_{h,h}$ remains the same regardless the presence of ranking error as long as the same ranking mechanism is used in both populations. Since the $BW_{h,h}$ are mutually independent, the null distribution of FM is the convolution of H independently distributed MW statistics. The articles by Bohn (1998), Ozturk (2002), Ozturk and Wolfe (2000b,c), and the references therein provide a review of the other relevant research in rank-based inference in ranked-set sample. Readers are referred to Hollander, Wolfe, and Chicken (2014, Chap. 15) and Wolfe (2012) for more recent comprehensive reviews of ranked set sampling designs.

Two-sample distribution-free inference in JPS samples has not been studied in the literature. We develop distribution-free inference for the stochastic ordering between two distributions F and G based on judgement post-stratified samples. Section 2 introduces a class of rank sum tests for JPS data, develops their distributional properties, and compares them with the traditional MW statistic in SRS and the BW statistic in RSS data. Section 3 introduces another class of rank sum statistics. It is shown that statistics in this class are distribution-free regardless of the quality of the ranking information. Section 4 provides empirical evidence for the performance of the proposed tests. Section 5 applies the proposed methods to an experiment on spray deposits. Section 6 provides a concluding remark. Proofs are in the Supplementary Material.

2. Rank Sum Tests for Judgment Post-Stratified Data

In this section, we introduce a rank sum statistic based on JPS samples to test the stochastic order relation between two distributions F and G. Let $I_{hx} = 1$ if $N_h > 0$ and $I_{hx} = 0$ if $N_h = 0$. In similar fashion we use the indicator function I_{qy} to denote that judgment class q in the Y-sample data is nonempty. Let $d_n = \sum_{h=1}^{H} I_{hx}$ and $d_m = \sum_{q=1}^{Q} I_{qy}$ be the number of non-empty judgement classes in the X- and Y-sample data, respectively. We set

$$J_{hx}^{a} = \begin{cases} 0 & N_h = 0, \\ \frac{1}{N_h^a} & N_h > 0, \end{cases} \qquad J_{qy}^{b} = \begin{cases} 0 & M_q = 0, \\ \frac{1}{M_q^b} & M_q > 0, \end{cases}$$

and for J^1_{hx} and J^1_{qy}, we simply write J_{hx} and J_{qy}. Our proposed test rejects the null hypothesis $H_0 : F \overset{s}{=} G$ against $H_A : F \overset{s}{\neq} G$ for large values of

$$T = \frac{1}{d_n d_m} \sum_{h=1}^{H} \sum_{q=1}^{Q} I_{hx} J_{hx} I_{qy} J_{qy} T_{hq},\ T_{hq} = \sum_{i=1}^{n} \sum_{j=1}^{m} \psi(Y_j - X_i) I(R_i = h) I(W_j = q),$$

where $I(R_i = h) = 1$ if $R_i = h$ and zero otherwise. We use the notation "$\overset{s}{=}$" to indicate the stochastic equality between the distribution F and G. The proposed statistic can be interpreted as a weighted version of the BW statistic, where weights are introduced to minimize the effect of empty judgment classes. We let

$$\tau_{[hq]}(F,G) = \int F_{[h]} dG_{[q]}(y), \qquad \tau_{[.q]}(F,G) = \sum_{h=1}^{H} \tau_{[hq]}(F,G),$$

$$\tau_{[h.]}(F,G) = \sum_{q=1}^{Q} \tau_{[hq]}(F,G), \qquad \tau_{[..]}(F,G) = \sum_{h=1}^{H} \sum_{q=1}^{Q} \tau_{[hq]}(F,G),$$

$$\gamma_{[HQ]}(F,G) = \sum_{h=1}^{H} \sum_{q=1}^{Q} \tau^2_{[hq]}(F,G), \qquad \gamma_{[.Q]}(F,G) = \sum_{q=1}^{Q} \tau^2_{[.q]}(F,G),$$

$$\gamma_{[H.]}(F,G) = \sum_{h=1}^{H} \tau^2_{[h.]}(F,G), \qquad \theta(F,G) = \int (1 - F(y))^2 dG(y),$$

$$\eta_{[H]}(F,G) = Q \sum_{h=1}^{H} \int F^2_{[h]}(y) dG(y), \qquad \eta_{[Q]}(G,F) = H \sum_{q=1}^{Q} \int G^2_{[q]}(y) dF(y),$$

$$\xi_{[HQ]}(F,G) = \sum_{h=1}^{H} \sum_{q=1}^{Q} \int \{F_{[h]}(y) - \tau_{[hq]}(F,G)\}^2 dG_{[q]}(y) = \eta_{[H]}(F,G) - \gamma_{[HQ]}(F,G),$$

$$\xi_{[QH]}(G,F) = \sum_{q=1}^{Q} \sum_{h=1}^{H} \int \{G_{[q]}(y) - \tau_{[qh]}(G,F)\}^2 dF_{[h]}(y) = \eta_{[Q]}(G,F) - \gamma_{[QH]}(G,F),$$

where square brackets are used to indicate that the quantities are computed under the assumption of imperfect ranking. When we use the perfect ranking assumption, we replace the square brackets with parentheses.

The distributions of the indicator functions I_{hx}; $h = 1, \ldots, H$, I_{qy}; $q = 1, \ldots, Q$, and the sample size vectors $\mathbf{N}$ and $\mathbf{M}$ do not depend on the ranking mechanism. The following Lemma is from Ozturk (2014b); the proof is omitted here.

Lemma 1. *In a JPS sample $(X_j, R_j); j = 1, \ldots, n$, let I_{hx} be the indicator function that the judgment class h is not empty and d_n be the number of non-empty judgment classes, $d_n = \sum_{h=1}^{H} I_{hx}$. For the distribution of I_{hx}/d_n, we have*

$$\text{(i)}\quad P(\tfrac{I_{hx}}{d_n}=u)=\begin{cases}\left\{\frac{H-1}{H}\right\}^n & u=0,\\ \frac{1}{H^n}\binom{H-1}{k-1}\sum_{j=1}^{k}(-1)^{j-1}\binom{k}{j-1}(k-j+1)^n & u=\frac{1}{k};k=1,\ldots,H;\end{cases}$$

$$\text{(ii)}\quad E(\tfrac{I_{hx}}{d_n})=\tfrac{1}{H};$$

$$\text{(iii)}\quad E(\tfrac{I_{hx}^2}{d_n^2})=\tfrac{1}{H^2}\sum_{k=1}^{H}\left(\tfrac{k}{H}\right)^{n-1};$$

$$\text{(iv)}\quad E(\tfrac{I_{hx}I_{h'x}}{d_n^2})=\tfrac{1}{H(H-1)}\left(1-\tfrac{1}{H}\sum_{k=1}^{H}\left(\tfrac{k}{H}\right)^{n-1}\right),\quad h\neq h';$$

$$\text{(v)}\quad E\left(\tfrac{I_{hx}^2J_{hx}}{d_n^2}\right)=\tfrac{1}{H^n}\left\{\tfrac{1}{n}+\sum_{k=2}^{H}\sum_{j=1}^{k-1}\sum_{n_h=1}^{n-k+1}\tfrac{(-1)^{j-1}}{k^2n_h}\binom{H-1}{k-1}\binom{k-1}{j-1}\binom{n}{n_h}(k-j)^{n-n_h}\right\}.$$

Expressions similar to the ones given in Lemma 1 can also be written for the indicator function I_{yq} and the judgment class sample sizes M_q; $q=1,\ldots,Q$. Using Lemma 1, the following expressions can be evaluated analytically.

$$a_1(n,m,H,Q)=\left\{E(\frac{I_{1y}^2}{d_m^2})-\frac{1}{H^2Q^2}\right\},$$

$$a_2(n,m,H,Q)=\left\{E(\frac{I_{1x}^2}{d_n^2})E(\frac{I_{1y}I_{2y}}{d_m^2})-\frac{1}{H^2Q^2}\right\},$$

$$a_3(n,m,H,Q)=\left\{E(\frac{I_{1y}^2}{d_m^2})E(\frac{I_{1x}I_{2x}}{d_n^2})-\frac{1}{H^2Q^2}\right\},$$

$$a_4(n,m,H,Q)=\left\{E(\frac{I_{1x}I_{2x}}{d_n^2})E(\frac{I_{1y}I_{2y}}{d_m^2})-\frac{1}{H^2Q^2}\right\},$$

$$b_1(n,m,H,Q)=\left\{E(\frac{I_{1x}^2J_{1x}}{d_n^2})E(\frac{I_{1y}^2}{d_m^2})-E(\frac{I_{1x}^2J_{1x}}{d_n^2})E(\frac{I_{1y}^2J_{1y}}{d_m^2})-E(\frac{I_{1x}^2J_{1x}}{d_n^2})E(\frac{I_{1y}I_{2y}}{d_m^2})\right\},$$

$$b_2(n,m,H,Q)=\left\{E(\frac{I_{1y}^2J_{1y}}{d_m^2})E(\frac{I_{1x}^2}{d_n^2})-E(\frac{I_{1x}^2J_{1x}}{d_n^2})E(\frac{I_{1y}^2J_{1y}}{d_m^2})-E(\frac{I_{1y}^2J_{1y}}{d_m^2})E(\frac{I_{1x}I_{2x}}{d_n^2})\right\},$$

$$b_3(n,m,H,Q)=E(\frac{I_{1x}^2J_{1x}}{d_n^2})E(\frac{I_{1y}^2J_{1y}}{d_m^2}),\quad b_4(n,m,H,Q)=E(\frac{I_{1x}I_{2x}}{d_n^2})E(\frac{I_{1y}^2J_{1y}}{d_m^2}),$$

$$b_5(n,m,H,Q)=E(\frac{I_{1y}I_{2y}}{d_m^2})E(\frac{I_{1x}^2J_{1x}}{d_n^2}).$$

We say that ranking scheme is consistent provided

$$F(x)=\frac{1}{H}\sum_{h=1}^{H}F_{[h]}(x)\text{ for all }x.$$

Under a consistent ranking scheme, we can find the mean and variance of the statistic T for any sample sizes $n\geq 1$ and $m\geq 1$, and set sizes $H\geq 1$ and $Q\geq 1$.

Lemma 2. *Let (X_i, R_i); $i = 1, \ldots, n$, and (Y_j, W_j); $j = 1, \ldots, m$, be two JPS samples with a consistent ranking scheme from the distributions F and G, respectively. Then for any $n \geq 1$ and $m \geq 1$,*

$$E(T) = \frac{\tau_{[..]}(F,G)}{HQ} = \int \{1 - G(y)\} dF(y)$$

and the variance of $\sqrt{n+m}T$ is

$$\sigma^2_{n,m,[H,Q]}(F,G) = (n+m)A_{n,m,[H,Q]}(F,G) + (n+m)B_{n,m,[H,Q]}(F,G),$$

where

$$\begin{aligned}
&A_{n,m,[H,Q]}(F,G)\\
&\quad = \gamma_{[HQ]}(F,G)a_1(n,m,H,Q) + \left\{\gamma_{[H.]}(F,G) - \gamma_{[HQ]}(F,G)\right\} a_2(n,m,H,Q)\\
&\qquad + \left\{\gamma_{[.Q]}(F,G) - \gamma_{[HQ]}(F,G)\right\} a_3(n,m,H,Q)\\
&\qquad + \left\{\tau^2_{[..]}(F,G) - \gamma_{[.Q]}(F,G) - \gamma_{[H.]}(F,G) + \gamma_{[HQ]}(F,G)\right\} a_4(n,m,H,Q),\\
&B_{n,m,[H,Q]}(F,G)\\
&\quad = \xi_{[QH]}(G,F))b_1(n,m,H,Q) + \xi_{[HQ]}(F,G)b_2(n,m,H,Q)\\
&\qquad + \left\{\tau_{[..]}(F,G) - \gamma_{HQ}(F,G)\right\} b_3(n,m,H,Q)\\
&\qquad + \left\{H^2Q\theta(F,G) - \gamma_{[.Q]}(F,G)\right\} b_4(n,m,H,Q)\\
&\qquad + \left\{HQ^2\theta(G,F) - \gamma_{[H.]}(F,G)\right\} b_5(n,m,H,Q).
\end{aligned}$$

In Lemma 2, if $G = F$, then $E(T) = 1/2$, but the variance of T still depends on F through the ranking process. In addition to the equality of $G = F$, if the ranking process is perfect, the variance of T is distribution free.

Corollary 1. *If the ranking process is perfect and $F = G$ then for any $n \geq 1$, $m \geq 1$, $H \geq 1$ and $Q \geq 1$,*

$$\gamma_{(HQ)}(F,F) = \sum_{h=1}^{H}\sum_{q=1}^{Q}\left\{\sum_{i=h}^{H}\frac{Q\binom{H}{i}\binom{Q-1}{q-1}}{(H+Q)\binom{H+Q-1}{q+i-1}}\right\}^2,\ \tau_{[..]}(F,F) = \frac{HQ}{2},$$

$$\theta(F,F) = \frac{1}{3},\quad \gamma_{(.Q)}(F,F) = \frac{H^2Q(2Q+1)}{6(Q+1)},\quad \gamma_{(H.)}(F,F) = \frac{Q^2H(2H+1)}{6(H+1)},$$

$$\eta_{(H)}(F,F) = \sum_{h=1}^{H}\sum_{i=h}^{H}\sum_{j=h}^{H}\frac{Q\binom{H}{i}\binom{H}{j}}{(2H+1)\binom{2H}{i+j}},\ \eta_{(Q)}(F,F) = \sum_{q=1}^{Q}\sum_{i=q}^{Q}\sum_{j=q}^{Q}\frac{H\binom{Q}{i}\binom{Q}{j}}{(2Q+1)\binom{2Q}{i+j}},$$

$$\sigma^2_{n,m,(H,Q)}(F,F) = \sqrt{n+m}A_{n,m,(H,Q)}(F,F) + \sqrt{n+m}B_{n,m,(H,Q)}(F,F),$$

where the round brackets in subscripts indicate that within-set ranking process is perfect.

The result has that, under a perfect ranking assumption, the variance of the rank sum statistic under H_0 in JPS setting is distribution-free as is the Bohn-Wolfe statistic in RSS setting, but it has additional terms due to the random nature of the sample size vectors $\mathbf{N}$ and $\mathbf{M}$. For large sample sizes, the results of Lemma 2 can be simplified. Let n_0 be the minimum of n and m and $\lambda = \lim_{n_0 \to \infty} n/(n+m)$.

Corollary 2. *As n_0 goes to infinity, $\sigma^2_{n,m,[H,Q]}(F,G)$ reduces to*

$$\sigma^2_{\lambda,[H,Q]}(F,G) = \left\{H^2 Q\theta(F,G) - \gamma_{[.Q]}(F,G)\right\} \frac{1}{(1-\lambda)H^2Q} + \left\{HQ^2\theta(G,F) - \gamma_{[H.]}(F,G)\right\} \frac{1}{\lambda Q^2 H}$$

under any consistent ranking scheme, and to

$$\sigma^2_{\lambda,(H,Q)}(F,F) = \left\{\frac{1}{3} - \frac{2Q+1}{6(Q+1)}\right\} \frac{1}{1-\lambda} + \left\{\frac{1}{3} - \frac{2H+1}{6(H+1)}\right\} \frac{1}{\lambda}$$

under a perfect ranking scheme.

Here the asymptotic variance of the rank sum statistics in a JPS sample is the same as the variance of the rank sum statistics in a ranked set sample in Bohn and Wolfe (1992), but equivalence may require large sample sizes. To inspect the rate of convergence, we plotted $A_{n,m,(H,Q)}(F,F)$ and $B_{n,m,(H,Q)}(F,F)$ against $n+m$ in Figure 1. The dashed and solid horizontal lines are the limiting values of $B_{n,m,(H,Q)}(F,F)$ and $A_{n,m,(H,Q)}(F,F)$, respectively. In Figure 1, $A_{n,m,(H,Q)}(F,F)$ converges to zero very rapidly, while the convergence of $B_{n,m,(H,Q)}(F,F)$ to its limit is not as fast. Equivalence between JPS and RSS rank sum statistics may require relatively large sample sizes.

We now look at the limiting distribution of T when, in the X- and Y-samples, ranks and sample size vectors are random variables. In the construction of the limiting distribution, we then need to pay attention to random behavior of the sample sizes.

Lemma 3. *Let (X_i, R_i); $i = 1, \ldots, n$, and (Y_j, W_j); $j = 1, \ldots, m$, be JPS samples from the X- and Y-sample distributions, respectively. Assume that $F = G$. As n and m grow large,*

$$\sqrt{n+m}(T - \frac{1}{2}) = \sqrt{n+m}\left\{\sum_{h=1}^{H}\sum_{i=1}^{n} \frac{I_{hx} I(R_i = h)}{d_n N_h}(1 - F(X_i) - \bar{\tau}_{[h.]}(F,F))\right\} + \sqrt{n+m}\left\{\sum_{q=1}^{Q}\sum_{i=1}^{n} \frac{I_{qy} I(W_j = q)}{d_m M_q}(F(Y_j) - \bar{\tau}_{[.q]})\right\} + o_p(1),$$

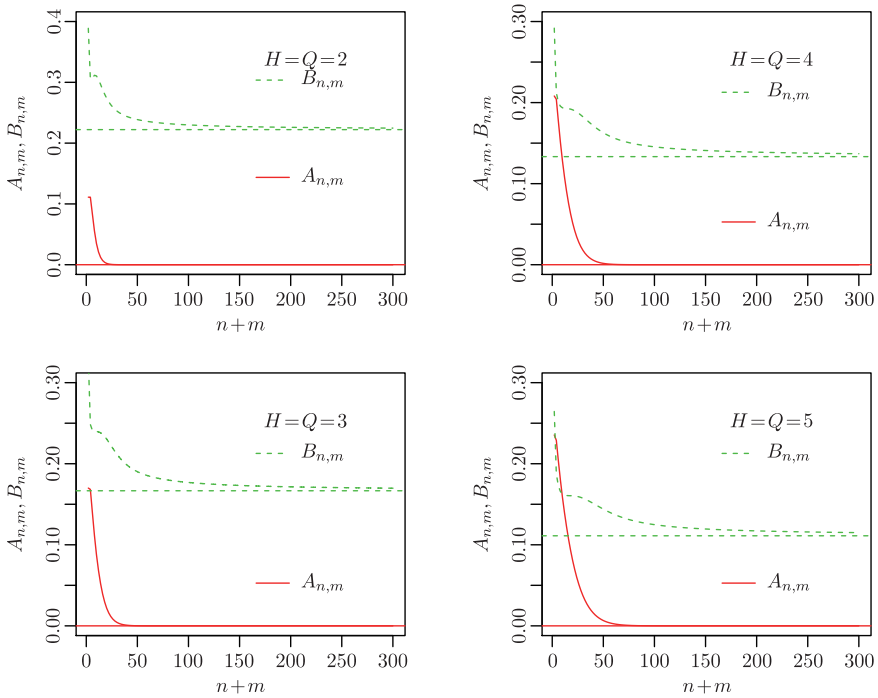

Figure 1. The plots of $A_{n,m} = A_{n,m,(H,Q)}$ and $B_{n,H} = B_{n,m,(H,Q)}$ with respect to sum of the sample sizes (n+m).

and $\sqrt{n+m}(T-1/2)$ *converges to a normal distribution with mean zero and variance*

$$\sigma^2_{\lambda,[H,Q]}(F,F) = \left\{\frac{1}{3} - \frac{1}{H}\sum_{h=1}^{H}\left(\int F(y)dF_{[h]}(y)\right)^2\right\}\frac{1}{\lambda} + \left\{\frac{1}{3} - \frac{1}{Q}\sum_{q=1}^{Q}\left(\int F(y)dF_{[q]}(y)\right)^2\right\}\frac{1}{(1-\lambda)},$$

$\bar{\tau}_{[h.]}(F,F) = \sum_{q=1}^{Q}\tau_{[hq]}(F,F)/Q$ *and* $\bar{\tau}_{[.q]}(F,F) = \sum_{h=1}^{H}\tau_{[hq]}(F,F)/H$.

Here $\sigma^2_{\lambda,[H,Q]}(F,F)$ is a function of F, so the null distribution of the test statistic T is not distribution-free, even for large sample sizes, unless ranking information is perfect or completely random. Since there is no available estimator for the variance of T under imperfect ranking the test procedure based on T uses $\sigma^2_{\lambda,(H,Q)}(F,F)$ in place of $\sigma^2_{\lambda,[H,Q]}(F,F)$. This inflates the size of the test when $\rho < 1$.

We inspected sensitivity of the test against a departure from the perfect ranking assumption in a small scale simulation study. In the simulation study,

Table 1. Simulated Type I error rates of rank sum test under varying degree of ranking information.

n	m	H	Q	$\rho = 1$	$\rho = 0.9$	$\rho = 0.75$	$\rho = 0.5$
15	15	3	3	0.052	0.074	0.100	0.142
15	15	3	5	0.054	0.086	0.120	0.157
15	15	5	3	0.046	0.077	0.116	0.163
15	15	5	5	0.054	0.082	0.123	0.180
15	30	3	3	0.051	0.075	0.109	0.144
15	30	3	5	0.046	0.072	0.108	0.168
15	30	5	3	0.052	0.080	0.124	0.159
15	30	5	5	0.056	0.084	0.125	0.180
30	15	3	3	0.051	0.076	0.104	0.133
30	15	3	5	0.051	0.078	0.114	0.166
30	15	5	3	0.054	0.076	0.113	0.165
30	15	5	5	0.051	0.088	0.128	0.175
30	30	3	3	0.051	0.078	0.110	0.140
30	30	3	5	0.052	0.086	0.126	0.180
30	30	5	3	0.051	0.078	0.128	0.165
30	30	5	5	0.048	0.093	0.157	0.215

we generated JPS samples with set sizes $H = 3, 5$, $Q = 3, 5$, and sample sizes $n = 15, 30$, $m = 15, 30$. Judgment ranks were generated using the Dell and Clutter (1972) model that assumes that each unit has a perceived size variable U_i related to the variable of interest X_i through a bivariate distribution. The bivariate distribution can be constructed in different ways. Let $\boldsymbol{X} = (X_1^*, X_2, \ldots, X_H)$ be an H-dimensional random vector from the underlying distribution F having variance σ_X^2, where X_1^* is the measured unit and $X_2, \ldots, X_H$ are the unmeasured units in a set in a JPS sample. We generate another H-dimensional random vector $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_H)$ from a normal distribution with mean zero and variance σ_ϵ^2, and add $\boldsymbol{X}$ and $\boldsymbol{\epsilon}$ to have

$$\boldsymbol{U} = \boldsymbol{X} + \boldsymbol{\epsilon}, \tag{2.1}$$

where random vector $\boldsymbol{U}$ is conditionally independent for a given value of $\boldsymbol{X}$. This model creates independent pairs of random vectors, $(U_1, X_1^*), (U_2, X_2), \ldots, (U_H, X_H)$, with correlation coefficient ρ, where $\rho = 1/\sqrt{1 + \sigma_\epsilon^2/\sigma_X^2}$. Units are ranked based on perceived sizes, $U_1 \cdots U_H$, and the rank of U_1 is recorded as judgment rank for the measured observation X_1^*. The quality of ranking information is controlled by the correlation coefficient ρ between the perceived size (U) and response variable (X).

In the simulation study, we used $\rho = 1$ (perfect ranking) and $\rho = 0.9, 0.75, 0.5$ (imperfect ranking). The simulation size was 5,000. Table 1 indicates that the

rank sum test achieves its nominal size (0.05) under perfect ranking information when $\rho = 1.0$, but the size of the test is inflated under even the minor ranking error $\rho = 0.9$. Simulated type I error rates are around 0.08 when $\rho = 0.9$ and jump to between 0.10 and 0.21 when ranking information is between $\rho = 0.75$ and $\rho = 0.5$. The main reason here is that the variance of the rank sum test statistic is inflated under imperfect ranking, $var(\sqrt{n+m}T) = \sigma^2_{\lambda,[h,Q]}(F,F) > \sigma^2_{\lambda,(H,Q)}(F,F)$. Since the rank-sum statistic uses $\sigma^2_{\lambda,(H,Q)}(F,F)$ as a variance, even though it is not the correct value under imperfect ranking, the test statistic is inflated and yields the larger type I error rates. One could estimate $\sigma^2_{\lambda,[h,Q]}(F,F)$ from the data and use this estimate to construct the test statistic. We find this approach unappealing since we would need to estimate a large number of judgment class parameters and some of these judgment classes may have very few observations. Another approach is to use the bootstrap distribution to compute the p-value of test statistic.

3. Rank Sum Test Based on Stratification

This section introduces a distribution-free test for any $n > 1$ and $m > 1$ under any consistent judgment ranking scheme. The proposed test extends the Fligner and MacEachern (2006) rank sum test from ranked set sampling design to a JPS setting. Here it is assumed that the set sizes H and Q in the X- and Y-samples are equal ($H = Q$), but the sample sizes n and m can be different. Let

$$T_\omega = \sum_{h=1}^{H} \omega_h I_{hx} J_{hx} I_{hy} J_{hy} T_{h,h},$$

where $\sum_{h=1}^{H} \omega_h I_{hx} I_{hy} = 1$. The test statistic T_ω is a linear combination of rank sum statistics of nonempty strata. The coefficient ω_h depends only on the sample size vectors $(N_1, \ldots, N_H)$ and $(M_1, \ldots, M_H)$. The weights ω_h are chosen to minimize the conditional variance of T_ω given the rank vectors of the X- and Y-samples. If $\omega = 1/d_{nm}$, we write $T_1 = T_{d_{nm}}$,where d_{nm} is the number of non-empty matching judgment classes from the X- and Y-samples, $d_{nm} = \sum_{h=1}^{H} I_{hx} I_{hy}$.

There are clear differences between the rank sum test statistics T and T_ω. The test statistic T makes nm comparisons by comparing all X-observations with all Y-observations, but T_ω compares only the X- and Y- observations from the same judgment class. Thus, T_ω makes $N_1M_1 + \cdots + N_HM_H$ comparisons between the X- and Y-sample observations. One might think that the test based on the statistic T_ω may have lower power than the test based on the statistic T, but the loss of efficiency is not very large. Fligner and MacEachern (2006) showed that these test statistics have the same efficiency under random ranking in a

two-sample ranked set sample. Under perfect and imperfect ranking information these two statistics are nearly equal, which of them performs better depends on the underlying distribution. We provide some preliminary results on the distributional properties of T_ω.

Lemma 4. *Let (X_i, R_i); $i = 1, \ldots, n$, and (Y_j, W_j); $j = 1, \ldots, m$, be two JPS samples with a consistent ranking scheme with $H = Q$. If $F = G$, then for any $n > 1$, $m > 1$, and $a \geq 0$, $b \geq 0$,*

$$E\left(\frac{I_{1x} I_{1y} J_{1x}^a J_{1y}^b}{d_{nm}^2}\right)$$
$$= \sum_{k=1}^{k^*} \sum_{t=0}^{t^*} \frac{\binom{H-1}{k-1}\binom{H-k}{t}}{H^{n+m} k^2} \left\{ \sum_{n_1=1}^{n-k-t+1} \sum_{j=1}^{k+t-1} (-1)^{j-1} \binom{k+t-1}{j-1} \binom{n}{n_1} \frac{(k+t-j)^{n-n_1}}{n_1^a} \right.$$
$$\left. \times \sum_{u=0}^{u^*} \binom{H-k-t}{u} \sum_{m_1=1}^{m-k-u+1} \sum_{i=1}^{k+u-1} (-1)^{i-1} \binom{k+u-1}{i-1} \binom{m}{m_1} \frac{(k+u-i)^{m-m_1}}{m_1^b} \right\},$$

where $k^* = min(n, m, H)$, $u^* = min(m-1, H-k-t)$, $t^* = min(n-1, H-k)$.

Lemma 5. *Let (X_i, R_i); $i = 1, \ldots, n$, and (Y_j, W_j);$j = 1, \ldots, m$, be two JPS samples from the X- and Y- populations. Suppose that the same ranking procedure is used in all sets.*

(i) $E(T_\omega) = \frac{1}{H} \sum_{h=1}^{H} \int \{1 - G_{[h]}(y)\} dF_{[h]}(y)$.

(ii) *If the null hypothesis is true ($F = G$), then the variance of T_ω is*

$$var(T_\omega) = \frac{H}{12} \left\{ E\left(\frac{\omega_1^2 I_{1x} I_{1y}}{N_1}\right) + E\left(\frac{\omega_1^2 I_{1x} I_{1y}}{M_1}\right) + E\left(\frac{\omega_1^2 I_{1x} I_{1y}}{N_1 M_1}\right) \right\}$$

for any n and m.

(iii) *The conditional variance of T_ω, given the judgment class sample size vectors $\boldsymbol{N}$ and $\boldsymbol{M}$, is*

$$var(T_\omega | \boldsymbol{R}, \boldsymbol{W}) = \sum_{h=1}^{H} \frac{\omega_h^2 I_{hx} I_{hy} (N_h + M_h + 1)}{12 N_h M_h}.$$

The choice of the weight

$$\omega_{h,o} = \frac{I_{hx} I_{hy} N_h M_h / (N_h + M_h + 1)}{\sum_{h=1}^{H} I_{hx} I_{hy} N_h M_h / (N_h + M_h + 1)} \tag{3.1}$$

minimizes the conditional variance of T_ω given the rank vectors $\boldsymbol{R}$ and $\boldsymbol{W}$. If T_o is the test statistic with optimal weight $\omega_{h,o}$ in equation (3.1), the variance

of T_o *is*

$$V(T_o) = \frac{1}{12}\sum_{h=1}^{H} E\left\{ \frac{I_{hx}I_{hy}N_hM_h/(N_h+M_h+1)}{\left(\sum_{h=1}^{H} I_{hx}I_{hy}N_hM_h/(N_h+M_h+1)\right)^2} \right\} = \frac{1}{12}\sum_{h=1}^{H}\Omega_h. \tag{3.2}$$

It is clear that the null variance of T_ω is distribution-free regardless of the quality of ranking information and the choice of the weight vector. For the optimal weight, we need to evaluate the expectation at (3.2) and an analytic expression for this expectation is a challenge. On the other hand, sample size vectors $\mathbf{N}$ and $\mathbf{M}$ have multinomial distributions regardless of the quality of the ranking information. We then estimate this expectation from a small simulation study by generating independent random vectors $\mathbf{N_i} = (N_{1i}, \ldots, N_{Hi})$ and $\mathbf{M_i} = (M_{1i}, \ldots, M_{Hi})$ from a multinomial distribution with parameters n, $(1/H, \ldots 1/H)$ and m, $(1/H, \ldots 1/H)$. Let

$$\omega^*_{hi} = \frac{I_{hxi}I_{hyi}N_{hi}M_{hi}/(N_{hi}+M_{hi}+1)}{\left(\sum_{h=1}^{H} I_{hxi}I_{hyi}N_{hi}M_{hi}/(N_{hi}+M_{hi}+1)\right)^2}, \quad i = 1, \ldots, B.$$

We estimate Ω_h with $\hat{\Omega}_h = \sum_{i=1}^{B}\omega^*_{hi}/B$, where B is selected large enough to have a consistent estimator. Our simulation study in Section 4 indicates that $B = 200$ provides a reasonably good estimator for the parameter Ω_h.

We now investigate the Pitman efficacies of the tests T, T_ω, and T_o for the location shift model $G(y) = F(y - \Delta)$,

$$eff(T) = \frac{(\mu'_T(0))^2}{var_0(T)}, \quad \mu_T(\Delta) = E_\Delta(T),$$

where $\mu'_T(0)$ is the derivative of $\mu_T(\Delta)$ at zero and $var_0(T)$ is the variance of T under the null hypothesis (Hettmansperger and McKean (2011)). Under this shift model, $G_{[q]}(y) = F_{[q]}(y - \Delta)$; $q = 1, \ldots, Q$. It is easy to see from Lemmas 2 and 5 that

$$\begin{aligned}
\mu_T(\Delta) &= E_\Delta(T) = \int (1 - F(y-\Delta))dF(y) \text{ and } \mu'_T(0) = \int f^2(y)dy,\\
\mu_{T_\omega}(\Delta) &= E_\Delta(T_\omega)\\
&= \frac{1}{H}\sum_{h=1}^{H}\int (1 - F_{[h]}(y-\Delta))dF_{[h]}(y) \text{ and } \mu'_{T_\omega}(0) = \frac{1}{H}\sum_{h-1}^{H}\int f^2_{[h]}(y)dy.
\end{aligned}$$

The efficacies of the tests, from the asymptotic null variances of test statistics in

Lemmas 2 and 5, are then

$$eff(BW) = eff(T) = \frac{(\int f^2(y)dy)^2}{\sigma^2_{\lambda,[H,Q]}(F,F)},\ eff(MW) = 12\lambda(1-\lambda)\Big\{\int f^2(y)dy\Big\}^2,$$

$$eff(FM) = eff(T_1) = eff(T_o) = 12\lambda(1-\lambda)\Big\{\frac{1}{H}\sum_{h=1}^{H} f^2_{[h]}(y)dy\Big\}^2.$$

The efficacy factors of BW and MF are given in Bohn and Wolfe (1992) and Fligner and MacEachern (2006), respectively. Under perfect ranking we replace the square brackets with parentheses. When the sample sizes n and m get large, both the X- and Y-population JPS samples approach to a balanced ranked set sample, with $\omega_{h,o} \approx 1/H$, and statistics T_1 and T_o yield the same asymptotic Pitman efficacy. Since the optimal weights $\omega_{h,o}$ approach to $1/H$ for large n and m, the asymptotic null variance of T_1 and T_o is the same as the asymptotic null variance of the simple random sample MW statistics

$$var_{H_0}(MW) = var_{H_0}(T_1) = var_{H_0}(T_o) = \frac{1}{12\lambda(1-\lambda)}.$$

The improved efficiency of T_1 and T_o over MW then comes from $\mu'_{T_1}(0)$ and $\mu'_{T_o}(0)$. The statistics MW and T have the same slope for the local power, but they have different null variances.

The asymptotic Pitman relative efficiency of a test T_1 with respect to a test T_2 is

$$RE(T_1,T_2) = \frac{eff(T_1)}{eff(T_2)}.$$

Table 2 provides the asymptotic Pitman relative efficiencies of $T_o = T_1$ with respect to T and MW for $H = Q = 2,3,4,5$. We also provide $RE(T,MW)$ in the last column of Table 2 for comparison purposes. Relative efficiencies are evaluated for standard normal, Student's t-distribution with 3 degrees of freedom ($t(3)$), and uniform distributions under a perfect ranking scheme. The proposed testing procedures T_1 and T_o are asymptotically better than MW procedures for all distributions in Table 2, the amount of improvement depending on the underlying distribution.

4. Finite Sample Comparisons

In this section, we compare the performance of the testing procedures T_1, T_o, and T based on JPS, T_{FM} based on RSS, and MW based on simple random samples in finite sample settings. Our simulation study considered a variety of set sizes, sample sizes, and underlying distributions; the JPS and RSS samples were constructed under a wide range of quality of ranking information, ranging

Table 2. Asypmtotic Pitman relative efficiencies of $T_1 = T_o$ with respect to T and MW.

Dist	H	$RE(T_1, T)$	$RE(T_1, MW)$	$RE(T, MW)$
Normal	2	0.986	1.479	1.5
Normal	3	0.976	1.952	2.0
Normal	4	0.969	2.421	2.5
Normal	5	0.963	2.889	3.0
t(3)	2	0.931	1.396	1.5
t(3)	3	0.901	1.802	2.0
t(3)	4	0.885	2.213	2.5
t(3)	5	0.876	2.627	3.0
Uniform	2	1.185	1.778	1.5
Uniform	3	1.280	2.560	2.0
Uniform	4	1.337	3.344	2.5
Uniform	5	1.376	4.128	3.0

from perfect to random ranking. For all tests the critical values were obtained from a simulation under the null hypothesis to match their type I error rates at $\alpha = 0.05$.

Data sets were generated from the standard normal, $t(3)$, and uniform $(0, 1)$, using the additive perceptual-error model (Dell and Clutter (1972), Fligner and MacEachern (2006)). The construction of this model is given in Section 2. In each case we tested the null hypothesis $H_0 : \Delta = 0$ against $H_a : \Delta > 0$. The Δ parameter was taken to be $\Delta = 0, 0.1, 0.2, \ldots, 1$. The quality of ranking information was controlled by the correlation coefficient between the U_i and X_i in model (2.1), using $\rho = 1$ for perfect ranking, and $\rho = 0.9$ and $\rho = 0.75$ for imperfect ranking. For a fourth probability model, the data sets were generated from the Weibull distribution with gamma perception. This is a non-additive perceptual model suggested by Fligner and MacEachern (2006). The model is parametrized based on

$$X|\theta \sim Weibull(\theta, \beta) \text{ and } U|X \sim gamma(uX, 1),$$

where θ and uX are the shape parameters of Weibull and gamma, respectively, and β is the scale parameter of the Weibull distributions. The quality of ranking information is governed by the parameter u: the larger the u better the ranking. For $\theta = 1$, the correlation coefficient between X and U is $\rho = 1/\sqrt{1 + 1/u}$. Equivalently, we select the parameter u for a given value of ρ, $u = \rho^2/(1 - \rho^2)$. In this model we first generated a random vector $\boldsymbol{X} = (X_1^*, X_2, \ldots, X_H)$ of size H from $Weibull(1, \beta)$ and then generated another random vector $\mathbf{U}$ from $gamma(u\mathbf{X}, 1)$, where the first component of $\boldsymbol{X}$, X_1^*, is considered as the measured observation in a set in the JPS sample. The components of vector $\mathbf{U}$ were ranked and the

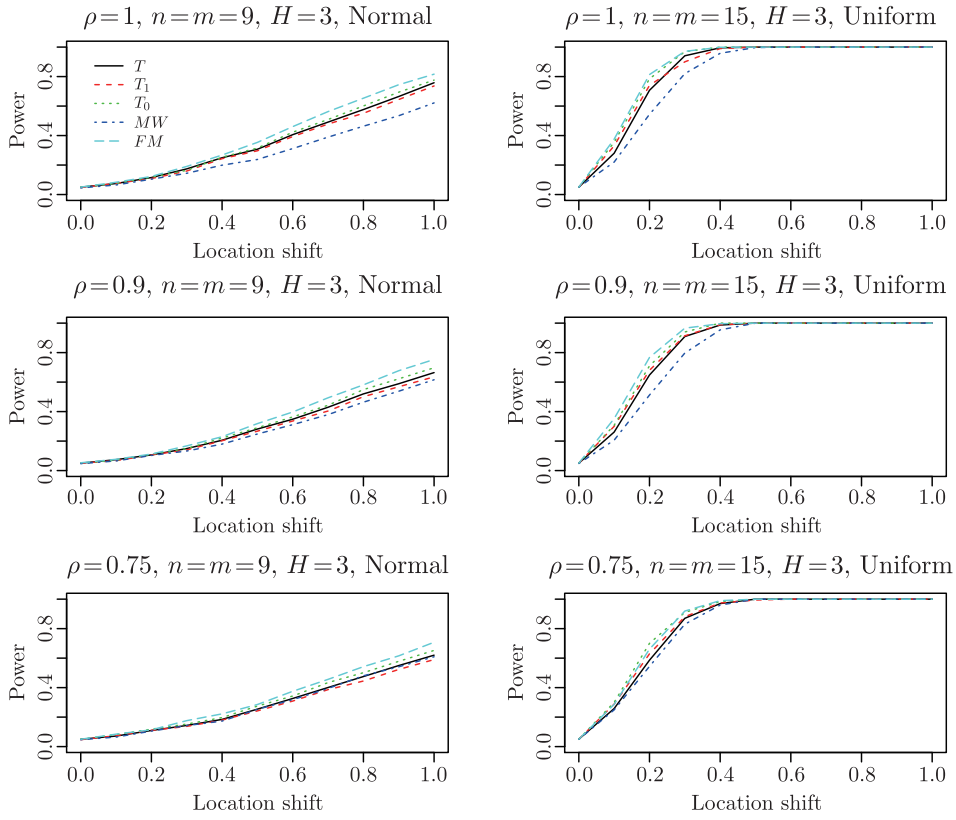

Figure 2. Power plots of five tests based on statistics T, $T_\omega = T_1$, T_o, MW and FM.

rank of U_1 was taken as the judgment rank for the measured observation X_1^*. The Weibull is an asymmetric distribution. In this model, the X-sample data were generated with $\beta_x = 1$ and the Y-sample data were generated with $\beta_y > 1$. We tested the null hypothesis $H_0 : \beta_y = \beta_x$ against the alternative $\beta_y > \beta_x$, equivalent to testing $H_0 : F \stackrel{s}{=} G$ against $H_0 : F \stackrel{s}{>} G$, where $F = Weibull(1, \beta_x)$ and $G = Weibull(1, \beta_y)$.

The power curves of the tests T, T_o, T_1, FM, and MW are given in Figures 2 and 3 with each power curve computed based on 5,000 iterations. Each panel shows five power curves: the solid line is for T, short-dashed line is for T_1, dotted line is for T_o, the dashed-dotted line is for MW, and long-dashed line is for FM. The lowest power curve in these panels belong to MW, as expected. Since the ranking process in an RSS sample induces stronger data structure, the FM test is slightly more efficient than T_1, T_0, and T_{BW} for small samples (n=m=9, H=3), but the loss of efficiency in test T_ω is not that large. For large sample sizes ($n = m = 24$ and $n = m = 36$), all RSS and JPS tests appear to have

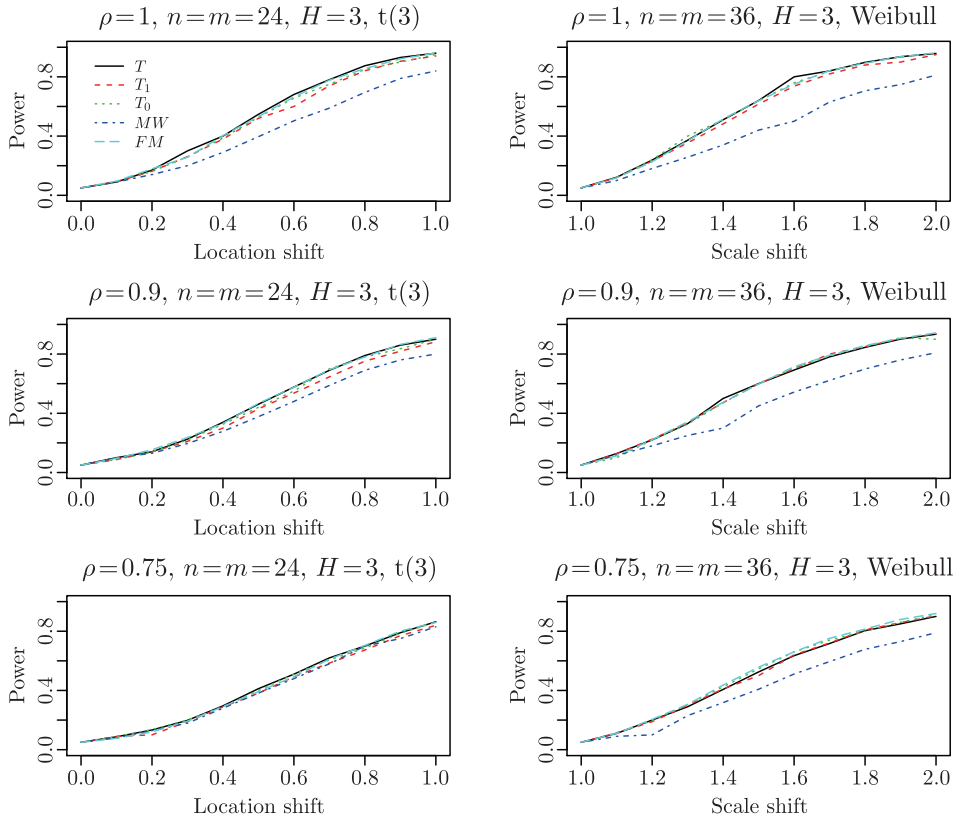

Figure 3. Power plots of five tests based on statistics T, $T_\omega = T_1$, T_o, MW and FM.

similar efficiencies. In this case, JPS sample has smaller probability of having empty judgment classes and in the limit it becomes an RSS sample. For the uniform distribution, while the power curve of FM is still the highest, the power curve of T_1 lies in between those of T_o and T. The power curve of MW is the lowest. These results are consistent with Pitman efficiency results except that T_o is slightly more efficient than T_1 for the uniform. Table 2 has T_o and T_ω with the same Pitman asymptotic efficiency, though the panels in Figure 2 show higher powers for T_o. This is a finite sample effect. The optimality of T_o is established based on the minimization of the finite sample variance of the conditional distribution of T_w given the sample size vectors $\boldsymbol{N}$ and $\boldsymbol{M}$. Even though they asymptotically have the same efficiency, their efficiencies differ for finite sample sizes, as in the uniform.

Figure 3 presents the power curves of the five tests for Student's t-distribution with 3 degrees of freedom and Weibull distributions. The power curves for T, FM, T_0 for the t-distribution appear to be identical and slightly higher than

the power curves of T_1. The same pattern holds for the Weibull distribution: T, T_1, T_o, and FM have almost identical power curves and are all above the power curve of MW test based on an SRS sample.

The size of the test T is usually inflated under imperfect ranking unless its critical region is simulated under the null hypothesis. The tests T_1 and T_o are distribution free and achieve their nominal sizes for any sample sizes n and m. The simulation study has the proposed distribution-free test T_o preferable over T and MW since its null distribution does not depend on the underlying distribution, it has almost the same power as the power of the test T, and always higher power than the power of MW.

5. Example

A pilot study was conducted by the researchers at the Horticulture Research International and the University of Kent in 1997 to investigate the efficiency of a ranked-set sample design over a simple random sample design. Murray, Ridout, and Cross (2000) provided a detailed description of the pilot study and reported that an RSS design had a significant amount of efficiency gain over an SRS design in the estimation of a population mean. The experiment compared the coverage of spray deposits on the leaves of apple trees under two sprayer settings. The researchers sprayed two plots of nine trees with a water-soluble fluorescent tracer at 2% concentration. The trees in the first plot was sprayed at a high volume with a coarse nozzle sprayer setting to produce large droplet sizes on the surface of the leaves. The second plot of nine trees was sprayed at a low volume with a fine nozzle sprayer setting to produce small droplet sizes. Twenty-five sets of five leaves (125 leaves) were sampled from the middle five trees in each plot. The leaves in the sets were inspected by four observers, independently, under an ultraviolet light to rank the percentage of upper- and lower-surface area coverage (%Cover) of the spray deposits. The measurement for %Cover was made by using an image analysis (Optimax V). The experiment was run for two randomizations and a total of 125 leaves were measured for each treatment. The precise measurement of %Cover is a time-consuming and and expensive process, so subjective ranking provided additional information to improve the statistical inference. Murray, Ridout, and Cross (2000) also measured the amount of spray deposit (Deposit) for all leaves in each set to identify the true rank of the leaves, hence to evaluate the ranking accuracy of the four rankers. The true amount of deposit was measured by washing the leaf surface with 5mL (mili-liter) of water and measuring the relative concentration of tracer deposit.

We use %Cover on the upper surface area of the leaves as the variable of interest and only the ranking information of the first ranker. We treated 125 measurements (from first randomization) from the coarse and fine-treatment groups

as the X- and Y- sample populations, respectively. For the location shift between the X and Y populations, The Hodge-Lehman estimate is 0.051. The variances (σ_x^2, σ_y^2) of the populations are $\sigma_x^2 = 0.0132$ and $\sigma_y^2 = 0.0262$. Using the populations, we simulated many SRS and JPS samples of sample size $n = m = 15, 25$ and set size $H = Q = 3, 5$, and compared the empirical powers of the testing procedures T, T_1, T_o, and MW. The first part of the simulation looked at the type I error rates of the tests under the null hypothesis for a two-sided test. Test statistics were computed after subtracting Δ from each Y-population sample observations. The critical regions for the tests were computed from the normal approximation to the null distributions of the test statistics. The second part of the simulation considered the power of the tests. Here the Y-population observations were not centered so that shift parameter (alternative hypotheses) between X- and Y- population was 0.051.

To draw a JPS sample we used the Dell-Clutter perceived size model. For this data set, the perceived sizes were not observable so we first need to link the ranking quality of the observer to the correlation coefficient ρ in the Dell-Clutter model. Fligner and MacEachern (2006) reported that the average value of the Kendall tau distance between the observer's ranking and the true ranking was 1.48 for coarse nozzle setting and 1.40 for the fine nozzle setting. They also simulated the Dell-Clutter model with set size 5, $\rho = 0.9$, and the normal distribution, and reported the average Kendal tau distance between the perceived ranks and true ranks as 1.43 based on 10,000 replication. The Dell-Clutter model with $\rho = 0.9$ can generate the JPS data by approximately matching the ranking quality in RSS data in Murray, Ridout, and Cross (2000). We first took a simple random sample of n leaves with replacement. For each selected leaf, $X_i; i = 1, \ldots, n$, another $H - 1$ leaves were sampled from the remaining leaves to form a set $\boldsymbol{X}_i = (X_i^*, X_2, \ldots, X_H)$, of size H, where X_i^* is the measured observation in the set $\boldsymbol{X}_i$. A random vector, $\boldsymbol{\epsilon}_i = (\epsilon_1, \ldots \epsilon_H)$, of size H was generated from the normal with variance $\tau_\epsilon^2 = \sigma_x^2(1 - \rho^2)/\rho^2$. Vectors $\boldsymbol{X}_i$ and $\boldsymbol{\epsilon}_i$ were added to form the perceived size vector $\boldsymbol{U}_i = \boldsymbol{X}_i + \boldsymbol{\epsilon}_i$. The components of the vector $\boldsymbol{U}_i$ were ranked and the rank of U_1 in $\boldsymbol{U}_i$ was taken as judgment rank of X_i^*. The JPS sample for the Y-population was constructed in similar fashion.

We simulated 10,000 JPS and SRS data from the $X-$ and Y-populations with sample size $n = m = 15, 25$, set size $H = Q = 3, 5$, $\rho = 1, 0.9, 0.75$ and $\Delta = 0, 0.051$. Since we were sampling with replacement from a finite population, ties were possible among the sampled units, we broke them at random when they existed. The empirical powers of T, T_1, T_o, and MW are presented in Table 3. For a two-sided level 0.05 test, the rejection regions of the tests T_1, T_o, and MW are approximately equal to the nominal type I error rate (0.05) under the null hypothesis, regardless of the quality of ranking information. While the testing

Table 3. Simulated empirical powers of the tests T, T_1, T_o, and MW from the apple tree experiement in Murray, Ridout, and Cross (2000). The simulation size is 10,000.

n	m	H	$\rho = 1$	shift	T	Tw1	Topt	MW
15	15	3	1.00	0.000	0.044	0.039	0.040	0.059
15	15	3	0.90	0.000	0.080	0.044	0.048	0.063
15	15	3	0.75	0.000	0.102	0.047	0.050	0.064
15	15	5	1.00	0.000	0.049	0.036	0.047	0.062
15	15	5	0.90	0.000	0.088	0.038	0.045	0.060
15	15	5	0.75	0.000	0.130	0.042	0.050	0.059
15	15	3	1.00	0.051	0.280	0.278	0.307	0.216
15	15	3	0.90	0.051	0.313	0.240	0.262	0.217
15	15	3	0.75	0.051	0.325	0.208	0.227	0.220
15	15	5	1.00	0.051	0.307	0.329	0.378	0.217
15	15	5	0.90	0.051	0.346	0.251	0.290	0.209
15	15	5	0.75	0.051	0.376	0.204	0.231	0.226
25	25	3	1.00	0.000	0.042	0.037	0.042	0.052
25	25	3	0.90	0.000	0.068	0.037	0.038	0.054
25	25	3	0.75	0.000	0.103	0.042	0.043	0.053
25	25	5	1.00	0.000	0.042	0.035	0.038	0.059
25	25	5	0.90	0.000	0.088	0.033	0.037	0.055
25	25	5	0.75	0.000	0.151	0.040	0.043	0.056
25	25	3	1.00	0.051	0.505	0.494	0.519	0.314
25	25	3	0.90	0.051	0.502	0.424	0.445	0.317
25	25	3	0.75	0.051	0.507	0.363	0.384	0.314
25	25	5	1.00	0.051	0.608	0.587	0.651	0.316
25	25	5	0.90	0.051	0.588	0.457	0.518	0.325
25	25	5	0.75	0.051	0.586	0.353	0.401	0.318

procedure T yielded the nominal type I error rates under the null hypothesis for $\rho = 1$, it had inflated error rates for $\rho < 1$. For the power comparisons, we ignored the test T because it is not valid under imperfect ranking.

Powers of the tests T_1 and T_o, and MW, when the shift parameter was $\Delta = 0.051$, were consistent with the simulation results in Section 4. The test (T_o) based on optimal weight yielded the highest power. The power of MW test gave the lowest power, as expected. The type I error rates of T_o and T_1 were not effected by the quality of ranking information.

6. Concluding Remark

A judgement post-stratified sample has been shown, in many contexts, to yield desirable properties over a simple random sample. A JPS sample provides additional information in the form of judgment ranks associated with measured observations, allowing one to construct more powerful tests and more accurate

estimators. We considered two issues in developing inference in JPS sample. The first is that a JPS sample is prone to produce empty judgment classes and, if the statistics are not properly adjusted, the validity of the statistical inference is in question. The secondly, many procedures rely heavily on perfect ranking. It is then important to develop inferential procedures that are robust against ranking error and the presence of empty judgment classes.

This article develops a class of two-sample test procedures that are robust against the presence of ranking error and empty judgment classes. All tests in this class achieve their nominal levels regardless of the quality of ranking information and the degree of imbalance in the JPS samples (including the empty judgment classes). Tests are distribution-free for all set and sample sizes. Within this class, we construct an optimal test procedure by assigning a weight that depends on the judgment class sample sizes. This test provides a substantial improvement over a Mann-Whitney-Wilcoxon test based on a simple random sample.

Acknowledgement

The author thanks three anonymous reviewers for their constructive comments.

References

Bohn, L. L. (1998). A ranked-set sample signed-rank statistic. *J. Nonparametr. Stat.* **9**, 295-306.

Bohn, L. L. and Wolfe, D. A.(1992). Nonparametric two-sample procedures for ranked-set sample data. *J. Amer. Statist. Assoc.* **87**, 552- 561.

Chen, Z. (2001). The optimal ranked-set sampling scheme for inference on population quantiles. *Statist. Sinica* **11**, 23-37.

Dell, T. R. and Clutter, J. L. (1972). Ranked-set sampling theory with order statistics background. *Biometrics* **28**, 545-555.

Fligner, M. A. and MacEachern, S. N. (2006). Nonparametric two-sample inference for ranked-set sample data. *J. Amer. Statist. Assoc.* **101**, 1107-1118.

Frey, J. and Feeman, T. G. (2012). An improved mean estimator for judgment post-stratification. *Comput. Statist. Data Anal.* **56**, 418-426.

Frey, J. and Feeman, T. G. (2013). Variance estimation using judgment post-stratification. *Ann. Inst. Statist. Math.* **65**, 551-569.

Frey, J. and Ozturk, O. (2011). Constrained estimation using judgment post-stratification. *Ann. Inst. Statist. Math.* **63**, 769-789.

Hettmansperger, T. P. and McKean, J. W. (2011). *Robust Nonparametric Statistical Methods.* CRC Press, New York.

Hollander, M., Wolfe, D. A. and Chicken, E. (2014). *Nonparametric Statistical Method.* Wiley, New Jersey.

Kaur, A., Patil, G.P., Taillie, C. and Wit, J. (2002). Ranked set sample sign test for quantiles. *J. Statist. Plann. Inference* **100**, 337-347.

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

Murray, R. A., Ridout, M. S. and Cross, J. V. (2000). The use of ranked set sampling in spray deposit assessment. *Aspect of Appl. Biology* **57**, 141-146.

Ozturk, O. (2002). Ranked set sample rank regression estimator. *J. Amer. Statist. Assoc.* **97**, 1180-1191.

Ozturk, O. (2012). Combining ranking information in judgment post stratified and ranked set sampling designs. *Environ. Ecolog. Statist.* **19**,73-93.

Ozturk, O. (2013). Combining multi-observer information in partially rank-ordered judgment post-stratified and ranked set samples. *Canad. J. Statist.* **41**, 304-324.

Ozturk, O. (2014a). Estimation of population mean and total in a finite population setting using multiple auxiliary variables. *J. Agric. Biol. Environ. Stat.* **19**, 161-184 .

Ozturk, O. (2014b). Statistical inference for population quantiles and variance in judgment post-stratified samples. *Comput. Statist. Data Anal.* **77**, 188-205.

Ozturk, O. and Wolfe, D. A. (2000a). Alternative ranked set sampling protocols for the sign test. *Statist. Probab. Lett.* **47**, 15-33.

Ozturk, O. and Wolfe, D. A. (2000b). Optimal allocation procedure in ranked set sampling for unimodal and multi-model distributions. *Environ. Ecolog. Statist.* **7**, 343-356.

Ozturk, O. and Wolfe, D. A. (2000c). An improved ranked set two-sample Mann-Whitney-Wilcoxon test. *Canad. J. Statist.* **28**, 123-135.

Stokes, S. L., Wang, X. and Chen, M. (2007). Judgment post stratification with multiple rankers. *J. Stat. Theory Appl.* **6**, 344-359.

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

Wang, X., Stokes, L., Lim, J. and Chen, M. (2006). Concomitant of multivariate order statistics with application to judgment poststratification. *J. Amer. Statist. Assoc.* **101**, 1693-1704.

Wang, X., Wang, K. and Lim, J. (2012). Isotonized CDF estimation from judgment post-stratification data with empty strata. *Biometrics* **68**, 194-202.

Wolfe, D. A. (2012). Ranked set sampling: its relevance and impact on statistical inference. *ISRN Probability and Statistics*, doi:10.5402/2012/568385.

Department of Statistics, The Ohio State University, 1958 Neil Avenue, Columbus, OH 43210, USA.

E-mail: omer@stat.osu.edu

(Received May 2014; accepted November 2014)