Rank-sum Test Based on Order Restricted Randomized Design

Omer Ozturk and Yiping Sun *

Abstract— **One of the main principles in a design of experiment is to use blocking factors whenever it is possible. On the other hand, if the blocking information is not precisely defined and subjective in nature, it is usually discarded. This paper introduces a new design that utilizes available subjective information on experimental units to create artificial blocking factors and develops a rank-sum test to test the difference between two population medians.**

Keywords: Order restricted randomized design, ranking error , Wilcoxon test, ranked set sampling, Pittman efficiency.

1 Introduction

In many experimental settings, experimental units (EU) based on inherent variation may provide two types of information, formal measurements or informal and subjective observations. While the formal measurements are successfully used in traditional analysis as covariates, informal observations are usually ignored. A new design , order restricted randomized design (ORRD), exploits the use of this informal and subjective information to design an experiment. Sets of experimental units, each of size H are recruited from a potential population along with subjective information that they may have. This subjective information is used to judgement rank the EUs from smallest to largest in each set to create artificial covariates. Ranking process induces a positive correlation among within-set experimental units. The ORRD then uses a restricted randomization of the treatment regimes to the ordered units in each set to facilitate negative correlation between the responses coming from two different treatment group in the same set. This negative correlation acts as a variance reduction technique in the estimation of the contrast parameter.

Section 2 describes the ORR designs for a two-treatment setting. Section 3 introduces a rank-sum test based on ORR design to draw inference for the equality of the population medians. The Pittman efficacy of the test is computed and compared with its competitors. Section 4 develops asymptotic null distribution of the test statistics under a consistent judgment ranking scheme and provides empirical evidence that the new test outperforms Mann-Whitney-Wilcoxon rank-sum test based on simple random sampling design.

2 The order restricted randomized designs

We assume that the EUs enter study either sequentially or H at a time. In either case, we need to have H units to form a set. One replication of the basic ORR design requires two sets, each of size H and can be constructed in a three step procedure.

Step I. We first identify the design parameters H, α and β, where H is the set size, α and β are two disjoint sets that partition the integers $1, \cdots, H$.

Step II. In each of the two sets, we pre-experimentally judgment rank the EUs from smallest to largest based on available subjective information on within-set EUs, and obtain the ranks $R_1, \cdots, R_H$.

Step III. In one of the sets, we perform a randomization to assign the treatment level T_1 to EUs whose ranks are in the α-set and the treatment level T_2 to the EUs whose ranks are in the β-set. We perform an opposite allocation in the other set without a randomization so that each treatment level is applied to all the ranks $R_1, \cdots, R_H$.

This basic design is repeated n times to increase the sample size. We use $X_{[h]j}$ and $Y_{[h]j}$ to denote the response measurements from the h-th ranked unit, j-th repetition, and the treatments T_1 and T_2, respectively. We assume that response measurements can be modeled as follows

$$\begin{aligned} X_{[h]j} &= \mu_1 + \gamma_{[h]j}, h = 1, \cdots, H, j = 1, \cdots, n, \\ Y_{[h]j} &= \mu_2 + \gamma_{[h]j}, h = 1, \cdots, H, j = 1, \cdots, n, \end{aligned} \qquad (1)$$

where μ_1, μ_2 are the medians of the treatment groups T_1, T_2, respectively, and $\gamma_{[h]j}$ is the random error associated with the experimental unit in replicate j and judgment rank h.

To illustrate the construction of the design we consider an

*Omer Ozturk is Professor, Department of Statistics, The Ohio State University, Columbus OH 43210,. Email: omer@stat.osu.edu, Tel/Fax: 614-292-3346/2096 . Yiping Sun, Senior Statistician, Forest Research Institute, Harborside Financial Center Plaza V, Jersey City, NJ 07311, Email:yiping.sun@frx.com. This work is partially supported by NSA Grand MSPF-04G-10 and NSF grand DMS 06-05041.

example. Assume that we wish to conduct an experiment to compare the efficacy of two drugs, drug T_1 and drug T_2. We set $H = 3$ and $\alpha = \{2\}$ and $\beta = \{1, 3\}$. For just one repetition of the basic design we need two sets, each of which has three patients. We rank the patients in each set separately based on general quality of health measure, pre-medical history, etc. In one of the set, we randomly assign T_1 to a patient whose rank is 2 (in α-set) and T_2 to patients whose ranks are 1 and 3 (in set β). In the second set, we do the opposite allocation so that T_1 and T_2 are applied to patients whose ranks are 1, 3, and 2, respectively. It is clear from this example that each treatment group is applied to all the ranks. Thus, it is a balanced design. The basic ORR design for this example is illustrated in Table 1.

Table 1: Basic ORR design when $H = 3$, $\alpha = \{2\}$, and $\beta = \{1, 3\}$.

Set 1	β_1, T_2, $Y_{[\beta_1]1}$	α_1, T_1,$X_{[\alpha_1]1}$	β_2, T_2,$Y_{[\beta_2]1}$
Set 2	β_1, T_1,$X_{[\beta_1]1}$	α_2, T_2,$Y_{[\alpha_1]1}$	β_2, T_1,$X_{[\beta_2]1}$

The main features of this design can be summarized as follows:

- In each treatment group, there are n observations that has judgment rank h for $h = 1, \cdots, H$. Thus, each rank is equally represented within each treatment group. This property is useful to have an unbiased estimator for the treatment mean.
- Within set measurements, due to judgment ranking of experimental units, are not independent. Under some mild assumptions (for example under the additive model (1)), they are positively correlated.
- This design puts emphasis on the contrast parameter $\Delta = \mu_1 - \mu_2$. Within-set judgment ranking process creates an error structure with positive covariances between within set responses. The restricted randomization turns this positive covariances into negative one in the estimation of the contrast parameter Δ.
- This design is unique for $H = 2$. The number of designs increases with set size H. In this case, an efficient design can be found by selecting optimal design parameters α and β.

The previous works in ORR design demonstrated that use of subjective information along with restricted randomization yields highly efficient inference for control versus treatment comparison [1] and two-sample inference [2]. Close inspection of ORRD shows that judgment ranking process creates a kind of stratified sample. In this regard, each judgement class can be considered as a stratum. Borrowing the idea of post-stratification from the theory of sampling survey, judgment ranking can be done post-experimentally. This judgment post stratified ORR design also yields inference with high efficiency [3]. Two-sample inference based on ORR design is also considered in the context of median confidence intervals [4]. The detailed development of the theory in this paper is provided in a PhD dissertation at the Ohio State University. The proofs of the theorems can be found in [5].

3 Rank-sum Test

Let $F(x)$ and $G(y - \Delta)$ be the cumulative distribution function of the treatment populations T_1 and T_2, respectively. The parameter $\Delta = \mu_2 - \mu_1$ denotes the location shift between these two distributions, where μ_1 and μ_2 are the medians of F and G. Let $X_{[h]j}$ and $Y_{[h]j}$, $h = 1, \cdots, H$, $j = 1, \cdots, n$ be the samples from treatment populations T_1 and T_2 generated by ORR design in Section 2 . Let $F_{[h]}$ and $G_{[h]}$ be the cdf of the judgment order statistics $X_{[h]j}$ and $Y_{[h]j}$, respectively. In this notation, square brackets indicate the quality of ranking information. If there is no ranking error, we replace the square brackets with the round one and judgment ranked order statistics then become usual order statistics from a set of size H.

We now wish to develop a nonparametric test for the hypothesis

$$H_0 : \Delta = 0 \quad H_A : \Delta \neq 0.$$

Even though we use two-sided alternatives here test can be applied to one sided alternatives with minor modification. Let

$$T = \sum_{i=1}^{H} \sum_{j=1}^{n} \sum_{k=1}^{H} \sum_{t=1}^{n} I(X_{[i]j} \leq Y_{[k]t}). \tag{2}$$

We reject the null hypothesis for extreme values of T. For an arbitrary judgment ranking scheme, the exact null distribution of T is not possible. We then look at the null distribution of T as n goes to infinity.

Theorem 1 *Let $\bar{T} = T/(n^2H^2)$. For an arbitrary, but consistent ranking scheme, $E\bar{T} = 1/2$.*

Theorem 2 *Under a consistent ranking scheme, as n goes to infinity, the asymptotic null distribution of $\sqrt{2nH}(\bar{T} - 1/2)$ converges to a normal distribution with mean zero and variance $\sigma^2 = \frac{2}{H}(\sigma_1^2 + \sigma_2^2)$, where*

$$\begin{aligned} \sigma_1^2 &= Var\left[\sum_{i=1}^{u}(1 - F(X_{[\alpha_i]1}) - \bar{\tau}_{\alpha_i\cdot}) \right. \\ &\qquad \left. + \sum_{k=1}^{H-u}(F(Y_{[\beta_k]1}) - \bar{\tau}_{\cdot\beta_k})\right], \end{aligned}$$

$$\begin{aligned}\sigma_2^2 &= Var\left[\sum_{i=1}^{H-u}(1-F(X_{[\beta_i]1})-\bar{\tau}_{\beta_i\cdot})\right.\\ &\qquad \left.+\sum_{k=1}^{u}(F(Y_{[\alpha_k]1})-\bar{\tau}_{\cdot\alpha_k})\right],\end{aligned}$$

$\tau_{ik} = EI(X_{[i]j} \leq Y_{[k]t})$, $\bar{\tau}_{\cdot k} = \sum_{i=1}^{H}\tau_{ik}/H$, $\bar{\tau}_{i\cdot} = \sum_{k=1}^{H}\tau_{ik}/H$, and u is the number of elements in set α.

The results of Theorem 2 hold for an arbitrary ranking scheme as long as it is consistent in each set. The consistency here is used to denote that the same ranking mechanism is used in each set. We note that the asymptotic null distribution of $\bar{T}$ is not distribution free. It depends on the judgment ranking scheme. Under perfect ranking, considerable simplification is possible in Theorem 2.

Corollary 3 *Under perfect ranking , the asymptotic null distribution of $\sqrt{2nH}(\bar{T}-1/2)$ converges to a normal distribution with mean zero and variance σ_P^2, where*

$$\begin{aligned}\sigma_p^2 &= \frac{4}{H}\left\{\sum_{i=1}^{u}\frac{\alpha_i(H+1-\alpha_i)}{(H+1)^2(H+2)}\right.\\ &\quad +2\sum_{i=1}^{u}\sum_{j=1}^{u}I(\alpha_i<\alpha_j)\frac{\alpha_i(H+1-\alpha_j)}{(H+1)^2\,(H+2)}\\ &\quad +\sum_{k=1}^{H-u}\frac{\beta_k(H+1-\beta_k)}{(H+1)^2(H+2)}\\ &\quad +2\sum_{k=1}^{H-u}\sum_{t=1}^{H-u}I(\beta_k<\beta_t)\frac{\beta_k(H+1-\beta_t)}{(H+1)^2\,(H+2)}\\ &\quad -2\sum_{i=1}^{u}\sum_{k=1}^{H-u}I(\alpha_i<\beta_k)\frac{\alpha_i(H+1-\beta_k)}{(H+1)^2\,(H+2)}\\ &\quad \left.-2\sum_{i=1}^{u}\sum_{k=1}^{H-u}I(\beta_k<\alpha_i)\frac{\beta_k(H+1-\alpha_i)}{(H+1)^2\,(H+2)}\right\}.\end{aligned}$$

It is now clear that the asymptotic null distribution of $\bar{T}$ is distribution free under perfect ranking.

When $H > 2$ the number of designs that we can select for ORRD is not unique. In this case, we select a design so that the Pittman efficiency of the test is larger than any other design in its class. Let $G_{\Delta_n}(t) = F(t-\Delta_N)$, where $\Delta_n = a/\sqrt{n}$, $a > 0$. Under this local alternative, the Pittman efficacy of the test based on the design parameters α and β is given by

$$c^2(\alpha,\beta) = \frac{\mu'(0)}{\sigma_P^2}, \quad \mu'(0) = \frac{d}{d\Delta}E_{\Delta}\bar{T}|_{\Delta=0}.$$

For a general α and β, the Pittman efficacy of the ORR design is then given by

$$c^2(\alpha,\beta) = \frac{(\int f^2(y)dy)^2}{\sigma_P^2}.$$

In order to find the optimal design, we need to maximize this equation with respect to sets α and β. This is equivalent to minimizing the asymptotic null variance of $\bar{T}$.

Theorem 4 *Let $H > 2$ be any fixed integer. Then the Pittman efficacy of the test $\bar{T}$ is maximized when set α contains odd integers only and set β contains even integers only, or vice versa.*

The theorem 4 indicates that the optimal design is the one that distributes integers to set α and β as evenly as possible. This can be achieved by putting odd integers in set α and even integers in set β.

The asymptotic variance of the test statistics $\bar{T}$ reduces to a simple form for the optimal design.

Corollary 5 *Assume that set α and β contains odd and even integers. Under perfect ranking assumption, the asymptotic null variance of $\bar{T}$ based on optimal design, σ_{Opt}^2 reduces to*

$$\sigma_{Opt}^2 = \begin{cases}\frac{1}{(H+1)^2} & \textit{if } H \textit{ is even}\\ \frac{1}{H(H+2)} & \textit{if } H \textit{ is odd.}\end{cases}$$

The Pittman efficacy of the Mann-Whitney-Wilcoxon (MWW) test based on simple random sampling can be found in [6]. For equal sample sizes, it reduces to

$$c^2(MWW) = 3\{\int f^2(x)dx\}^2.$$

We now compare the asymptotic Pittman relative efficiency of the rank-sum test based on optimal ORRD with respect to MWW test based on simple random sample

$$eff(Opt, MWW) = \frac{c^2(Opt)}{c^2(MWW)} = \frac{1}{3\sigma_{opt}^2}.$$

For $H = 2, 3$, and 4, relative efficiencies are $3, 5$, and $25/3$, respectively.

The point and interval estimate of the shift parameter Δ can be constructed from pairwise differences of X- and Y-sample observations,

$$\hat{\Delta} = median\{Y_{[k]t} - X_{[i]j}\}.$$

The estimator has the same form of Hodges-Lehman estimator as in a simple random sample. On the other, its distributional properties are different due to within-set correlation structure.

Distribution-free confidence interval of Δ follows directly from the inversion of the null distribution of T. We first note that the null distribution of T is symmetric around $(nH)^2/2$. Let $D_{(1)} \leq \cdots \leq D_{(n^2H^2)}$ be the ordered differences of $Y_{[k]t} - X_{[i]j}$, for $k, i = 1, \cdots, H$ and $t, j = 1, \cdots, n$. If we select a k^* such that $P_0(T \leq k^*) = \eta/2$, from the symmetry of T we have that

$$[D_{(k^*+1)}, D_{(n^2H^2-k^*)}]$$

is an $100(1-\eta)\%$ confidence interval for Δ. For large n, k^* can be approximated form the asymptotic null distribution of T,

$$k^* = (nH)^2/2 - 0.5 - z_{\eta/2}\sigma_T,$$

where $\sigma_T^2 = (nH)^3\sigma_P^2/2$ is an estimate of the variance of T and z_a is the a-th upper quantile of the standard normal distribution.

In order to investigate the convergence rate of the asymptotic distribution of $\bar{T}$ we performed a simulation study. Simulation setting consists of different set (H), replication (n) sizes, varying degree of ranking information, and some common underlying distribution (F). Judgment ranking information is modeled through Dell and Clutter [7] model. This model uses an additive perceptual error model

$$u_i = \gamma_i + w_i,$$

where the residual γ_i assumed to have a distribution F with mean zero and variance 1. The random components w_i is generated from a normal distribution with mean zero and variance θ^2, and γ_i and w_i are independent. In order to generate judgment order statistics from this model, we generate two sets of random variate γ and w , each of size H. We add these vectors to obtain $u = \gamma + w$. This vector is sorted and corresponding γ values are taken as judgment order statistics. In this model, the quality of judgment ranking information is controlled by the correlation coefficient $\rho = corr(u, \gamma)$. The correlation $\rho = 1$ and $\rho = 0$ correspond to perfect and random ranking, respectively. Intermediate value of judgment ranking information can be considered by selecting $0 < \rho < 1$.

Table 2: Estimated Type I error rates when $n = 5$ and underlying distribution is standard normal.

H	$\rho = 1$	$\rho = 0.9$	$\rho = 0.75$	$\rho = 0.5$
2	0.042	0.112	0.178	0.240
3	0.038	0.161	0.259	0.339
4	0.043	0.242	0.371	0.469
5	0.044	0.283	0.447	0.523

Table 2 presents the estimated Type I error rates for different values of ρ. It is clear that simulated type I error rates are close to nominal Type I error rate of 0.05 if there is no ranking error. On the other hand even a small ranking error inflates the Type I error rates considerably.

4 Asymptotic Null Distribution under Imperfect Ranking

Simulation study in Section 3 indicates that even a slight departure from perfect ranking inflates the Type I error rates. Hence, test looses its distribution free property. Under imperfect ranking, the quantity σ_P^2 under estimates the variance of the test statistics. In order to correct this problem, It is important to have a consistent estimator for σ^2. Let Z_i, $i = 1, \cdots, 2n$ be the H-dimentional within set correlated observations in each replication

$$Z_i = (X_{[\alpha_1]i}, \cdots, X_{[\alpha_u]i}, Y_{[\beta_1]i}, \cdots, X_{[\beta_{H-u}]i}).$$

Theorem 6 *For a fixed set size H, under a consistent ranking scheme, unbiased and consistent estimator of σ^2 is given by $\hat{\sigma}^2 = \frac{4}{H}(H/3 + A - B - C)$, where*

$$\begin{aligned}
A &= -\sum_{i=1}^{H} \hat{\mu}_{[i]} + 2\sum_{i=1}^{H}\sum_{j=1}^{H} I(i<j)(\hat{\nu}_{[i,j]} - \hat{\mu}_{[i]}\hat{\mu}_{[j]}), \\
B &= 4\sum_{i=1}^{u}\sum_{k=1}^{H-u} I(\alpha_i < \beta_k)(\hat{\nu}_{[\alpha_i,\beta_k]} - \hat{\mu}_{[\alpha_i]}\hat{\mu}_{[\beta_k]}) \\
C &= 4\sum_{i=1}^{u}\sum_{k=1}^{H-u} I(\alpha_i > \beta_k)(\hat{\nu}_{[\alpha_i,\beta_k]} - \hat{\mu}_{[\alpha_i]}\hat{\mu}_{[\beta_k]}) \\
\hat{\mu}_{[i]} &= \frac{1}{2n(2n-1)H}\sum_{j=1}^{2n}\sum_{k\neq j}^{2n}\sum_{s=1}^{H} I(Z_{[s]k} \leq Z_{[i]j}) \\
\hat{\nu}_{[i,j]} &= \frac{\sum_{l=1}^{2n}\sum_{k\neq l}^{2n}\sum_{t\neq l,k}^{2n} T_{i,j,k,t,l}}{4n(2n-1)(n-1)H^2} \\
T_{i,j,k,t,l} &= \sum_{s=1}^{H} I(Z_{[s]k} \leq Z_{[i]l})\sum_{s=1}^{H} I(Z_{[s]t} \leq Z_{[j]l}).
\end{aligned}$$

By using the consistent estimator of σ^2 we can easily establish from the Slutsky's theorem that $\sqrt{2nH}(\bar{T} - 1/2)/\hat{\sigma}$ converges to a standard normal distribution as the repetition number n goes to infinity. Even though, this result holds for large n, it may not provide a satisfactory solution for small n. Since we estimate σ^2 consistently from the data, Student's t-distribution with $2n-2$ degrees of freedom provides better approximation for small n.

Table 3 presents the Type I error rate estimates based on Student's t-approximation under different judgment quality information. Underlying distributions are taken as standard normal (N), Student's t-distribution with 3-degrees of freedom (t_3) and lognormal distribution (LN).

It is now clear from Table 3 that estimates of the Type I error rates are relatively close to the nominal values for all ρ values and distributions in Table 4. Based on this result, for all practical purposes we may claim that the

Table 3: Estimated Type I error rates based on Student's t- approximation. Simulation size is 5,000 and $n = 5$.

Dist	H	$\rho = 1$	$\rho = 0.9$	$\rho = 0.75$	$\rho = 0.5$
N	2	0.041	0.051	0.053	0.053
	3	0.053	0.062	0.058	0.060
	4	0.045	0.059	0.056	0.055
t_3	2	0.040	0.056	0.056	0.059
	3	0.050	0.061	0.058	0.059
	4	0.039	0.061	0.056	0.054
	5	0.049	0.053	0.058	0.055
LN	2	0.038	0.056	0.058	0.057
	3	0.051	0.064	0.063	0.064
	4	0.040	0.049	0.051	0.052
	5	0.040	0.051	0.056	0.049

proposed test is asymptotically distribution free irrespective of the quality of judgment ranking information.

We next investigate the empirical power of the test. Simulation study considered set size $H = 3$, the number of replication $n = 5$ and varying degree of judgment ranking information. Residual for the ORRD are again generated from Dell and Clutter model for $\rho = 1, 0.9, 0.75$, and 0.50. For the alternative hypothesis we considered location shift of $\Delta = 0(0.1)1$. The empirical powers of the rank-sum test of ORRD along with classical Mann-Whitney-Wilcoxon test is given in Figure 1.

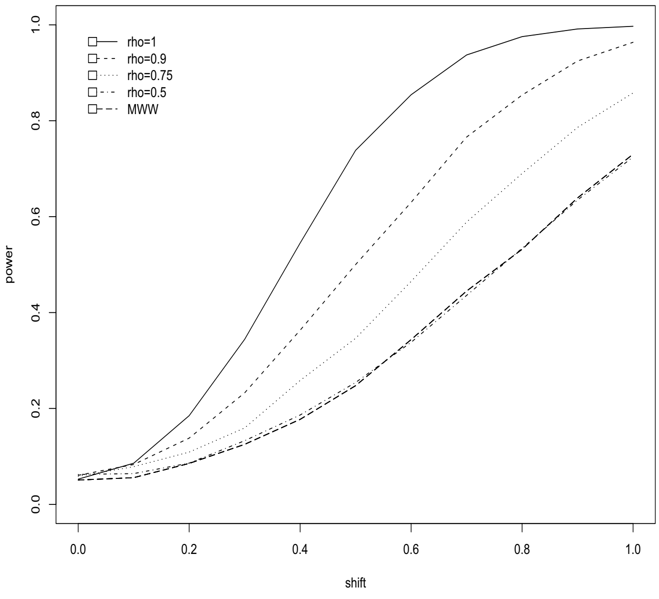

Figure 1: Empirical power curves of the rank-sum test based on ORRD for selected ρ(rho). Set size $H = 3$, replication size $n = 5$ and simulation size is $5,000$.

Figure 1 illustrates that the new test has substantially higher power than the the power of the Mann-Whitney-Wilcoxon test as long as there is some information to judgment rank the units prior to experimentation. If judgment ranking information is week, the correlation coefficient is less than 0.5, the ORR design is as good as simple random sampling design. This indicates that the proposed test does not loose its power if our ranking information leads to a random ranking.

5 Conclusion

This papers develops distribution-free inference based on ORR design for the location shift between two distributions. New design exploits the use of subjective information to rank the experimental units to produce more accurate inference for the contrast parameter. The approach that we have taken in this paper extends to more complex treatment structure with k-treatments. A test, similar to Kruskal-Wallis test, can be constructed. In this case, interesting design issues appear. In current work, the authors are pursuing the extension of ORR design to this k-treatment structure.

References

[1] Ozturk, O. and MacEachern, S. N. Order restricted randomized design for control versus treatment comparison. *Annals of the Institute of Mathematical Statistics*, 56, 701-720, 2004.

[2] Ozturk, O. and MacEachern, S. N. Order restricted randomized designs and two-sample inference, *Journal of Environmental and Ecological Statistics*,14, 365-381, 2007.

[3] Du. J. and MacEachern, S. N. Judgment post-stratification for deigned experiment. *Biometrics*, 64, 345-354, 2008.

[4] Two sample median test for order restricted randomized designs. *Statistics and Probability Letters*, 17, 131-141, 2007.

[5] Rank-sum test for two-sample location problem under order restricted randomized design. The PhD dissertation, Department of Statistics, The Ohio State University, 2007.

[6] , Hettmansperger, T. P. *Statistical Inference Based on Ranks*, Reprint Edition, Krieger Publishing Company, 1991.

[7] , Dell, T. R., and Clutter, J. L. Ranked-set sampling theory with order statistics background. *Biometrics,* 28, 545-555, 1977.