Published in final edited form as:

J Nonparametr Stat. 2012 September 1; 24(3): 797–808. doi:10.1080/10485252.2012.672647.

A General Class of Signed Rank Tests for Clustered Data when the Cluster Size is Potentially Informative

Somnath Datta,

University of Louisville, Louisville, KY, USA

Jaakko Nevalainen, and

University of Turku, Turku, Finland

Hannu Oja

University of Tampere, Tampere, Finland

Somnath Datta: somnath.datta@louisville.edu; Jaakko Nevalainen: jaakko.nevalainen@utu.fi; Hannu Oja: Hannu.Oja@uta.fi

SUMMARY

Rank based tests are alternatives to likelihood based tests popularized by their relative robustness and underlying elegant mathematical theory. There has been a surge in research activities in this area in recent years since a number of researchers are working to develop and extend rank based procedures to clustered dependent data which include situations with known correlation structures (e.g., as in mixed effects models) as well as more general form of dependence.

The purpose of this paper is to test the symmetry of a marginal distribution under clustered data. However, unlike most other papers in the area, we consider the possibility that the cluster size is a random variable whose distribution is dependent on the distribution of the variable of interest within a cluster. This situation typically arises when the clusters are defined in a natural way (e.g., not controlled by the experimenter or statistician) and in which the size of the cluster may carry information about the distribution of data values within a cluster.

Under the scenario of an informative cluster size, attempts to use some form of variance adjusted sign or signed rank tests would fail since they would not maintain the correct size under the distribution of marginal symmetry. To overcome this difficulty Datta and Satten (2008; *Biometrics*, 64, 501–507) proposed a Wilcoxon type signed rank test based on the principle of within cluster resampling. In this paper we study this problem in more generality by introducing a class of valid tests employing a general score function. Asymptotic null distribution of these tests is obtained. A simulation study shows that a more general choice of the score function can sometimes result in greater power than the Datta and Satten test; furthermore, this development offers the user a wider choice. We illustrate our tests using a real data example on spinal cord injury patients.

Key words and phrases

nonparametric tests; clustered data; signed rank test; informative cluster size; symmetry

1. Introduction

Rank based tests are popular alternatives to likelihood based tests. They are appealing to practitioners for their relative robustness and they have been a popular field of study amongst the mathematical statisticians for their elegant underlying theory. This area of research has seen maximum activities in the sixties and seventies. There has been a surge in research activities in this area in recent years since a number of researchers are trying to

develop and extend rank based procedures to clustered dependent data. These include papers by Rosner and Grove, 1999; Rosner et al., 2003, 2006; Williamson et al., 2003; Larocque, 2005; Datta and Satten, 2005; Werner and Brunner, 2007; Datta and Satten, 2008; Larocque et al., 2007, Larocque, 2003; Nevalainen et al., 2007a,b; Haataja et al., 2009 and so on. Clearly, without any modification, the classical tests do not work for dependent data. Under cluster dependence, the data are grouped into dependent observations called clusters which are independent of one another. Some authors have proposed using the same classical test statistics as with independent data and obtained estimators of their asymptotic variance using the independence of clusters or by modeling the dependence structure within a cluster. This approach typically does not work for clusters with informative cluster size since the usual rank test applies to properties of marginal distributions of all data values irrespective of the cluster membership. This marginal distribution does not coincide with the correct marginal distribution of a two stage sampling where the clusters are the units of sampling in the first stage. We define informative clustering in more formal terms later in the paper. Basically it means that the distribution of outcomes (variable of interest) within a cluster depends on the size of the cluster. This type of situation typically arises when the clusters are defined in a natural way not controlled by the experimenter or statistician and in which the size of the cluster may carry information about the distribution of data values within a cluster. As for example, consider observations on all available teeth of a collection of randomly selected individuals. In this case, an individual is a primary unit (of sampling in stage one). The outcome (e.g., attachment loss) of a randomly selected tooth of a chosen individual denotes a unit selected by the second stage sampling and its distribution is certainly dependent on the overall oral health of the individual which in turn is indicated by the number of teeth (cluster size) the individual has. See, Nevalainen et al. (2011) for an overview of informative cluster size problems.

It is not obvious how to adapt a classical rank test to such situations. As for example, the two sample (group) Wilcoxon rank sum test has difficulty maintaining the correct size under clustering. Even when the group membership is constant within a cluster, naive attempts such as using the sum of average ranks do not work. Recently, Rosner, Glynn and Lee (RGL, 2003) proposed a rank test for clustered data for the case where all observations in the same cluster belong to the same group. In addition, the RGL test assumes that cluster members are exchangeable and the correlation structure within clusters is independent of group. The RGL stratifies on cluster size; after ranking all observations (ignoring clustering), it compares the rank sum of observations from group 1 having cluster size k to the fraction of group 1 clusters of size k times the total rank sum of observations from clusters of size k . The final statistic sums the comparisons over all cluster sizes k and divides by an appropriate standard deviation. Datta and Satten take a different approach that makes no assumptions on the nature of the clustering, extending an idea (called within cluster resampling) for parameter-estimation of Hofman et al. (2001) and Williamson et al. (2003) to hypothesis testing. The resulting test is valid (e.g., maintains correct size) in a wide variety of settings (e.g., when members of the same cluster belong to different groups or when the correlation structure depends on group membership).

Datta and Satten (2008) obtained a novel signed rank test for paired data when the pairs are clustered. Within clusters, the pairs may not be independent. For example, in a nuclear family with several offspring at least one of which has a disease under investigation and at least one of which is disease free, data from the possible pairs of affected and unaffected offspring are correlated due to shared genetic and environmental factors. For the same problem, Larocque (2005) proposed using the classical signed rank test with a natural variance estimate that is based on certain sums of squares over independent clusters. Rosner, Glynn and Lee (2006) proposed adjusting the variance of a signed rank test by assuming a common intra-class correlation and estimating it from the ranks of the absolute values of the

observations. In this problem, the cluster size is informative if the common distribution of the difference in each pair within a cluster depends on the cluster size. Using a simulation study, Datta and Satten demonstrated that only their version of the clustered data signed rank test maintains the proper size amongst five different tests (including the two mentioned above) when the cluster size is informative.

Here, we study this problem in more generality by introducing a class of valid tests employing a general score function. Asymptotic null distribution of these tests are obtained. A simulation study shows that a more general choice of the score function can sometimes result in greater power than the Datta and Satten test.

The rest of the paper is organized as follows. We introduce our family of signed rank tests in Section 2 and study their asymptotic null distribution, as the number of clusters approaches infinity. Section 3 presents the results of two simulation studies where we demonstrate that these tests maintain proper size under informative cluster size and possess decent power in a variety of settings. In Section 4, we apply our tests to a real data set on longitudinal functional outcome measures of spinal cord injury patients. Our paper ends with a discussion of our results and some future research directions (Section 5). All proofs are collectively presented in the Appendix.

2. The null hypothesis and the class of test statistics

The so called one sample problem often arises when one compares paired observations. The classical sign test is used to test whether the median of the difference of observations in a pair has zero median and the signed-rank test is used to test a further stringent null hypothesis that this difference has a symmetric distribution.

In the current setup, the paired data are clustered. Let i index clusters and let j index pairs formed within the clusters. For the j th pair in the i th cluster, we assume that interest centers on X_{ij} , the pair-specific difference in the outcome measure of interest. We assume that the data generation scheme, including that of the (random) cluster size n_i is independent and identical across the clusters $i = 1, \dots, M$ so that the collection $V_i = \{n_i, X_{i1}, \dots, X_{in_i}\}$, $1 \leq i \leq M$ are independent and identically distributed. Conditional on cluster size n_i let J_i be a random variable uniformly distributed over $\{1, \dots, n_i\}$. Treating J_i as a random index, we find that X_{iJ_i} has the marginal distribution function $F(x)$ is given by

$$F(x) = E \left\{ \frac{1}{n_i} \sum_{j=1}^{n_i} I(X_{ij} \leq x) \right\} \quad (1)$$

where the expectation is with respect to the joint distribution of the random vector V_i . In other words, F is the distribution of pair-wise differences for a randomly chosen pair in a randomly chosen cluster. The marginal hypotheses we want to test are

$$H_0: F \text{ is symmetric (around 0) against } H_1: F \text{ is not symmetric.}$$

Let $F(x; n_i) = E \left\{ \frac{1}{n_i} \sum_{j=1}^{n_i} I(X_{ij} \leq x) \mid n_i \right\}$ be the marginal distribution of the data in the i th cluster given its cluster size. The cluster size will be informative if $F(\cdot; n_i)$ depends on n_i ; otherwise, it is non-informative in which case $F(\cdot; n_i) \equiv F, \forall i$.

Like Datta and Satten (2008), our test statistic is motivated by the “Within-cluster-resampling” approach of Hoffman et al. (2001). Consider pseudo data $X_i^*, 1 \leq i \leq M$, where for each i , X_i^* is generated by resampling one observation from the i th cluster of original observations. In other words, conditionally given the original clustered data, generate X_i^* from F_i , the empirical distribution of $\{X_{i1}, \dots, X_{in_i}\}; 1 \leq i \leq M$. It then follows that unconditionally, $X_i^*, 1 \leq i \leq M$, are independent and identically distributed with the common marginal distribution F given by (1). Thus, by the classical theory of signed rank test based on independent and identically distributed data (see, e.g., Hájek, Šidák and Sen, 1999, Sec 3.3.2), a test statistic for H_0 against H_1 can be formed using the signed rank statistic based on the $\{X_i^*: 1 \leq i \leq M\}$ and a score function ϕ . These statistics can be averaged across all possible realizations of the pseudo-data to form a test based on the original data that maintains the size but eliminates the artificial randomization due to resampling. Formally, let E^* denote the conditional expectation on the X^* data given the original data $\underline{X} = (X_{ij}: 1 \leq j \leq n_i, 1 \leq i \leq M)$; our marginal signed-rank test statistic is then

$$T_{\text{WCR},M} = E^* \left(\sum_{i=1}^M \phi \left(\frac{R_i^{*+}}{M+1} \right) \text{sign}(X_i^*) \right), \quad (2)$$

where R_i^{*+} is the rank of $|X_i^*|$ amongst $|X_1^*|, \dots, |X_M^*|$ and sign is the standard signum function defined as $\text{sign}(x) = I(x > 0) - I(x < 0)$. Here ϕ is a score function on $(0, 1)$ satisfying $\int_0^1 \phi^2(u) du < \infty$.

By a formal two term Taylor expansion and taking expectation and ignoring the higher order (in M) terms we see (Appendix) that an approximate test statistic from (2) is given by the following that can be computed from data without resampling:

$$T_{\text{WCR},M} \approx \sum_{i=1}^M \frac{1}{n_i} \sum_{k=1}^{n_i} \text{sgn}(X_{ik}) \phi(\bar{F}^+(|X_{ik}|)) := T_M, \text{ say}, \quad (3)$$

where F_i^+ is the empirical distribution of $\{|X_{ik}|, 1 \leq k \leq n_i\}$, and $\bar{F}^+ = (M+1)^{-1} \sum_{i=1}^M F_i^+$.

The expression T_M on RHS of (3) will be used as our proposed test statistic for this problem. This statistic weights the contribution of each cluster equally regardless of differential cluster size. Also, $\bar{F}^+$ consists of an unweighted average of F_i^+ regardless of different cluster size. We can show that, under H_0 , it has an asymptotically (as $M \rightarrow \infty$) normal distribution with zero mean and therefore it will maintain the correct size when the number of clusters M is large. We need the following regularity conditions for establishing the asymptotic distribution.

Conditions

- (0) $V_i = \{n_i, X_{i1}, \dots, X_{in_i}\}, 1 \leq i \leq M$ are independent and identically distributed, $1 \leq i \leq M$. Note that the cluster size n_i is random and is part of V_i .
- (i) Given the cluster size n_i , the conditional distributions of the cluster members $X_{ij}, 1 \leq j \leq n_i$, are identical and continuous.
- (ii) ϕ is differentiable and ϕ' satisfies

$$\lim_{\rho \rightarrow 0} \int_0^1 V_\rho^{\varphi'}(u) du = 0,$$

where $V_\rho^{\varphi'}(u) = \sup \{ |\varphi'(x) - \varphi'(u)| : (u-\rho) \vee 0 < x < (u+\rho) \wedge 1 \}$.

$$(iii) \quad \sigma^2 = E \left[\left\{ \frac{1}{n_i} \sum_{k=1}^{n_i} \text{sgn}(X_{ik}) \varphi(F^+(|X_{ik}|)) \right\}^2 \right] \in (0, \infty), \text{ where } F^+(u) = 2F(u) - 1, u \in [0, 1].$$

Condition (0) says that the clusters are independent and the data generation scheme is identical for each cluster. Condition (i) is a very general condition for the distribution of data in each cluster that is needed for the marginal model formulation. It can be shown that (see Lemma 1 in the appendix) under this condition, indeed F defined in (1) is the marginal distribution of X_{i1} , for any $i > 1$. Note that it allows for quite arbitrary correlation structure within a cluster. Condition (ii) is needed for handling the Taylor expansion of the rank statistic with a general score. It is basically a smoothness condition. It is slightly stronger than assuming that the score φ is continuously differentiable. Condition (iii) guarantees that CLT applies and that the limit distribution is non-degenerate.

Theorem 1—Under H_0 and conditions (0)-(iii),

$$M^{-1/2} T_M \xrightarrow{d} N(0, \sigma^2),$$

and

$$M^{-1/2} T_M / \widehat{\sigma} \xrightarrow{d} N(0, 1)$$

as $M \rightarrow \infty$, where

$$\widehat{\sigma}^2 = \frac{1}{M} \sum_{i=1}^M \left\{ \frac{1}{n_i} \sum_{k=1}^{n_i} \text{sgn}(X_{ik}) \varphi(\overline{F}^+(|X_{ik}|)) \right\}^2.$$

Remark 1—It is to be noted that unlike the classical case, the exact sampling distribution of the class of signed rank tests is intractable since it will, in general, depend on the dependence structure, the within cluster distribution and on the cluster size mechanism. A simulation study in the next section shows that the large sample tests maintained the nominal size even when the number of clusters was ten.

3. Simulations

We conduct two simulation studies to demonstrate the performance of our tests. In each simulation, we make two choices of the score function. In both simulations, we include the identity score $\varphi(x) = x$ and denote the corresponding test statistic by T_1 . For the first simulation model, we also include a score function $\varphi(x) = -\log(1 - x)$ and denote the test statistic by T_2 . In the second simulation setting, T_2 corresponds to the score function $\varphi(x) =$

$x^{1/2}$. In both simulations, a modified (inverse cluster size weighted) t -test statistic $T_3 = \left(\sum_{i=1}^M n_i^{-1} \sum_{j=1}^{n_i} X_{ij} \right) / \left\{ \sum_{i=1}^M (n_i^{-1} \sum_{j=1}^{n_i} X_{ij})^2 \right\}^{1/2}$ is also included for benchmarking.

For the first simulation model, we use a moderately large number of clusters $M = 50$. In the second simulation model, we use a small and moderate number of clusters $M = 10$, and $M = 30$, respectively. Critical values based on the normal approximation are used in all cases.

In the first simulation model, we generate data from a cluster specific random effects model $X_{ij} = \beta + \mu_i + e_{ij}$, where the random effects μ_i are i.i.d. $\sim F$. The models errors e_{ij} are i.i.d. with the same symmetric distribution F and are independent of μ_i . Three choices of F are considered, namely, standard normal, $\text{Beta}(\frac{1}{2}, \frac{1}{2}) - \frac{1}{2}$ and a double (two-sided) Pareto distribution with shape parameter 2 and scale parameter 1/2. An informative cluster size is created by letting $n_i = 5$, if $\mu_i < 0$ and $n_i = 50$, if $\mu_i > 0$. Each size/power entry in Tables 1 is empirically obtained with a Monte Carlo size of 5000. Note that $\beta = 0$ corresponds to the null hypothesis (1) in this case.

For both normal and beta distributed error terms, the signed rank tests track the (modified) t test fairly well in terms of performance. Generally speaking, the t test has the largest power followed by the signed rank test with $\phi(x) = -\log(1 - x)$. The test with $\phi(x) = x$ (i.e., the Datta-Satten test) has the lowest power, albeit not by much. The situation is reversed for the heavy tailed data (Pareto) where the t -test is conservative and has the least power. In this setting, the signed rank test with the identity score has slightly higher power than that with $\phi(x) = -\log(1 - x)$.

In the second simulation model, the signs of all observations in a cluster are the same and is dependent on the cluster size. The basic model is similar to the one used by Datta and Satten (2008) and is given by

$$X_{ij} = S_{ij} \exp(|H_{ij}|),$$

where $H_{ij} = W_i + e_{ij}$, $W_i \sim N(0, \rho_i)$, $e_{ij} \sim N(0, 1 - \rho_i)$, and the variables W_i and e_{ij} are independent. Here $\rho_i = \gamma_1$, say, for $1 \leq i \leq M/2$ and $= \gamma_2$, say, for $M/2 + 1 \leq i \leq M$. The cluster size is informative of S_{ij} the sign of X_{ij} which is generated as follows. For each $1 \leq i \leq M$, generate a random fraction p_i from a Beta $(1, \beta)$ distribution; if $p_i \leq 0.5$, let $n_i = 2$, $S_{ij} = +1$, $1 \leq j \leq n_i$ else $n_i = 8$, $\rho_i = \gamma_2$, say, $S_{ij} = -1$, $1 \leq j \leq n_i$. Four different choices of the intra-cluster correlation combinations (γ_1, γ_2) were used. Note that when $\beta = 1$, the marginal null hypothesis H_0 stated in Section 2 holds.

The results overall show that the signed rank test with $\phi(x) = \sqrt{x}$ performs the best followed by the signed rank test with $\phi(x) = x$ (the Datta-Satten test) and they both beat the t -test in most cases (the only exception being the cases with $\gamma_1 = \gamma_2 = 0.2$, $\beta = 0.3$ and 0.5 and $M = 10$. The power of all tests increases as the number of clusters is increased from 10 to 30.

4. Applications to real data

We demonstrate the usefulness of our tests through an application to a data on functional recovery over time in a group of spinal cord injury (SCI) patients at the NeuroRecovery Network (NRN). Eligibility criteria for NRN patients include SCI with lesion at level T10 or above and non participation in inpatient rehabilitation programs (Harkema et al., 2011). Patients are discharged from the NRN for non-compliance with treatment, patient election, or if a plateau in the recovery of function is achieved. Since more severely impaired patients

tend to have more “room for improvement” in function recovery and hence remained enrolled in the NRN for longer periods of time, contributing more observations than those that enrolled with higher baseline function, the cluster size is potentially informative.

One key question that can be asked is whether these patients tend to improve over time in the rehabilitation program. To answer this we look at the temporal changes (between two consecutive visits) in walking function as measured by the walking speed in a ten meter walk test. We test the null hypothesis that the marginal distribution corresponding to these changes is symmetric around zero versus the alternative that there is a positive location shift. Besides the distance traveled in a ten meter test at each visit, our data set also consists of a baseline covariate called the “initial phase” which is a clinical assessment of patient’s functional ability (1=best, 3= worst). Note that in this setting, a single individual represents a cluster and the changes in walking speed in successive evaluations are the observations within a cluster. Thus, the cluster size is equal to the number of visits made by an individual minus one.

We only kept those patients who had at least three visits. The temporal changes of walking speed for the same patient formed a cluster. The total number of patients in our analysis was 176 and the cluster size ranged from 2 to 15. Figure 1 shows that Box-plots of our data categorized by the cluster size indicating a mild informative cluster size effect on the median shift. Patients with a small number of visits (≤ 6) tend to have a greater median change. A formal test of significance of informative cluster size is beyond the scope of this present paper; this problem may be studied elsewhere. A weighted histogram of the temporal changes (between two visits by all individuals) is created where a temporal change contributes an amount toward the frequency count of its class that is inversely proportional to the number of evaluations (minus one) by that individual. The overall weighted histogram displayed in Figure 2 indicates that the center of the distribution is shifted to the right of zero. Also, the skewed nature of the histogram indicates that a rank based test may be more appropriate than a normal based parametric test.

We have applied our tests with $\phi(x) = x$ and $\phi(x) = \sqrt{x}$, as well as the modified t -tests and all the tests are highly significant (Table 3) and positive, validating improvement over time for the NRN patients (Harkema *et al*, 2011). Table 3 also shows results of sub-analyses given the initial phase.

5. Discussion

In this paper, we introduce a general class of signed-rank test that can be used to test the symmetry of the marginal distribution of data even when they are clustered. Furthermore, the dependence within a cluster can be quite arbitrary and complex. While the signs of each observation enter the construction of the test statistic in the usual way, their ranks have to be defined appropriately which is not the same as the rank in the pooled sample. In particular, an inverse cluster size weighting appears naturally in their definition leading to an asymptotically valid (e.g., size preserving) test even when the cluster size is informative.

One might think that the use of the inverse sample size automatically make the resulting tests inefficient (e.g., less powerful) in situations when the cluster size is not informative. However, our past studies (Datta and Satten, 2008) show that is not the case. A detailed simulation study was undertaken in that paper for the identity score function under a variety of correlation structure. Even when the cluster size was non-informative, the results were mixed. Generally speaking, for unbalanced cluster size, our test outperformed tests that did not use inverse cluster size reweighting (e.g., Rosner *et al.*, 2006; Larocque, 2005) when the average correlation was high; however, when the correlation was low and the design was

unbalanced our test had slightly lower power than those two tests. Somewhat surprisingly, for balanced design they were comparable. Our test also outperformed that of Rosner et al. (2006) when the correlation within cluster was non-constant.

Unlike the classical rank tests for i.i.d. data, the exact distribution of these statistics is intractable due to the general nature of dependence with each cluster. Therefore, one must resort to large sample normal approximation in obtaining the critical values of the tests. In our simulation studies, decent size behavior is observed even for sample (in this case, the number of independent clusters) sizes as little as ten. The powers of the tests are very reasonable as well as compared with a benchmark (cluster size adjusted) t-test. For heavy tailed distributions, our tests with an appropriate score function (e.g., bounded) remain robust and powerful as compared with the modified t-test.

In practice, we anticipate that the use of the identity function will be most popular. As the simulation studies show, it tends to work well in most cases even though in some cases, one can obtain modest gain in power by using a non-linear score. Moreover, from a theoretical perspective, a study of a class of tests with a general score function is more satisfying than a single test; similar treatments are common for the classical rank tests for i.i.d. data.

It will be of interest to study the power properties of the class of tests further, either theoretically or through more extensive simulations to get a sense of what type of score functions to select in what situation. A difficulty towards this is that the general setting of the marginal testing problem is perhaps too broad. However, it might be possible to obtain some optimality results in certain particular models (e.g., the one way random effects ANOVA). This will be pursued elsewhere. Another future research problem will be to extend the within cluster resampling itself. Instead of sampling at random within each cluster, one can employ a non-uniform sampling mechanism. This will lead to a more general class of weights (than the inverse cluster size) for the resulting test statistics.

Acknowledgments

We thank an anonymous reviewer for many useful comments and a careful reading of the earlier versions of the manuscript. This research was supported by grants from the National Security Agency (Grant number H98230-11-1-0168), the National Institutes of Health (Grant number 1R03DE020839-01A1) and the Academy of Finland.

References

- Datta S, Satten GA. Rank-sum tests for clustered data. *Journal of American Statistical Association*. 2005; 100:908–915.
- Datta S, Satten GA. A signed-rank test for clustered data. *Biometrics*. 2008; 64:501–507. [PubMed: 17970820]
- Hájek, J.; Šidák, Z.; Sen, PK. *Theory of Rank Tests*. San Diego: Academic Press; 1999.
- Harkema S, Schmidt-Read M, Lorenz D, Edgerton VR, Behrman A. Balance and ambulation improvements in individuals with chronic incomplete spinal cord injury using Locomotor Training-based rehabilitation. *Archives of Physical Medicine and Rehabilitation*. 2011 In Press.
- Haataja R, Larocque D, Nevalainen J, Oja H. A weighted multivariate sign-rank test for cluster-correlated data. *Journal of Multivariate Analysis*. 2009; 100:1107–1119.
- Hoffman EB, Sen PK, Weinberg CR. Within-cluster resampling. *Biometrika*. 2001; 88:1121–1134.
- Larocque D. An affine-invariant multivariate sign test for cluster correlated data. *Canadian Journal of Statistics*. 2003; 31:437–455.
- Larocque, D. The Wilcoxon signed-rank test for cluster correlated data. In: Duchesne, P.; Rémillard, B., editors. *Statistical Modeling and Analysis for Complex Data Problems*. New York: Springer; 2005. p. 309-323.

- Larocque D, Nevalainen J, Oja H. A weighted multivariate sign test for cluster-correlated data. *Biometrika*. 2007; 94:267–283.
- Nevalainen, J.; Datta, S.; Oja, H. An overview of informative cluster size problems. 2011. Preprint
- Nevalainen J, Larocque D, Oja H. A weighted spatial median for clustered data. *Statistical Method and Applications*. 2007a; 15:355–379.
- Nevalainen J, Larocque D, Oja H. On the multivariate spatial median for clustered data. *The Canadian Journal of Statistics*. 2007b; 35:215–283.
- Rosner B, Grove D. Use of the Mann-Whitney U-test for clustered data. *Statistics in Medicine*. 1999; 18:1387–1400. [PubMed: 10399203]
- Rosner B, Glynn RJ, Lee M-LT. Incorporation of clustering effects for the Wilcoxon rank sum test: a large sample approach. *Biometrics*. 2003; 59:1089–1098. [PubMed: 14969489]
- Rosner B, Glynn RJ, Lee M-LT. The Wilcoxon signed rank test for paired comparisons of clustered data. *Biometrics*. 2006; 62:185–192. [PubMed: 16542245]
- Werner C, Brunner E. Rank methods for the analysis of clustered data in diagnostic trials. *Computational Statistics & Data Analysis*. 2007; 51:5041–5054.
- Williamson JM, Datta S, Satten GA. Marginal analyses of clustered data when cluster size is informative. *Biometrics*. 2003; 59:36–42. [PubMed: 12762439]

Appendix: Proofs of theorems

Lemma 1

Under condition (iii) of Theorem 1, the marginal distribution function of X_{i1} equals F given by (1).

Proof

$$\begin{aligned} F(x) &= E \left\{ \frac{1}{n_i} \sum_{j=1}^{n_i} I(X_{ij} \leq x) \right\} = E \left[E \left\{ \frac{1}{n_i} \sum_{j=1}^{n_i} I(X_{ij} \leq x) \middle| n_i \right\} \right] \\ &= E \left[\frac{1}{n_i} \sum_{j=1}^{n_i} P \{ X_{ij} \leq x | n_i \} \right] = E \left[\frac{1}{n_i} \sum_{j=1}^{n_i} P \{ X_{i1} \leq x | n_i \} \right], \end{aligned}$$

since the conditional distribution function of X_{ij} given n_i is the same for all $1 \leq j \leq n_i$ by condition (ii),

$$= E [P \{ X_{i1} \leq x | n_i \}] = P \{ X_{i1} \leq x \},$$

proving the lemma.

Proof of Theorem 1

Express the test statistic as

$$\begin{aligned}
T_M &= \sum_{i=1}^M \frac{1}{n_i} \sum_{k=1}^{n_i} \text{sgn}(X_{ik}) \varphi(\bar{F}^+(|X_{ik}|)) = \sum_{i=1}^M \frac{1}{n_i} \sum_{k=1}^{n_i} \text{sgn}(X_{ik}) \varphi(F^+(|X_{ik}|)) \\
&\quad + \sum_{i=1}^M \frac{1}{n_i} \sum_{k=1}^{n_i} \text{sgn}(X_{ik}) \varphi'(F^+(|X_{ik}|)) (\bar{F}^+(|X_{ik}|) - F^+(|X_{ik}|)) \\
&\quad + \sum_{i=1}^M \frac{1}{n_i} \sum_{k=1}^{n_i} \text{sgn}(X_{ik}) \left\{ \varphi'(z_{ik} \bar{F}^+(|X_{ik}|) + (1 - z_{ik}) F^+(|X_{ik}|)) - \varphi'(F^+(|X_{ik}|)) \right\} \times \\
&\quad \quad (\bar{F}^+(|X_{ik}|) - F^+(|X_{ik}|)), \quad \text{for some } 0 < z_{ik} < 1, \\
&= T_{1,M} + T_{2,M} + T_{3,M}, \text{ say.}
\end{aligned}$$

The first term $T_{1,M}$ is a sum of i.i.d. terms

$$T_{1,M} = \sum_{i=1}^M S_i, \text{ with } S_i = \frac{1}{n_i} \sum_{k=1}^{n_i} \text{sgn}(X_{ik}) \varphi(F^+(|X_{ik}|)).$$

Using condition (i), it follows from the same arguments as in the proof of Lemma 1 that

$$E(S_i) = E \text{sgn}(X_{i1}) \varphi(F^+(|X_{i1}|)),$$

which equals zero since under H_0 , X_{i1} has distribution F (due to Lemma 1) that is symmetric. Hence under condition (iii), $M^{-1/2} T_{1,M} \xrightarrow{d} N(0, \sigma^2)$.

The second term $M^{-1} T_{2,M}$ can be expressed as a V-statistic in the i.i.d. $V_i, 1 \leq i \leq M$, with kernel

$$h(V_i, V_{i'}) = \frac{1}{2n_i n_{i'}} \sum_{k=1}^{n_i} \sum_{k'=1}^{n_{i'}} \left[a(X_{ik}) \{I(|X_{i'k'}| \leq |X_{ik}|) - F^+(|X_{ik}|)\} + a(X_{i'k'}) \{I(|X_{ik}| \leq |X_{i'k'}|) - F^+(|X_{i'k'}|)\} \right]$$

where $a(x) = \text{sgn}(x) \varphi' \circ F^+(|x|)$, $x \in \mathcal{R}$.

Note that for $i \neq i'$,

$$\begin{aligned}
E(h(V_i, V_{i'}) | V_i) &= \frac{1}{2n_i} \sum_{k=1}^{n_i} a(X_{ik}) \left[E \left\{ \frac{1}{n_{i'}} \sum_{k'=1}^{n_{i'}} \{I(|X_{i'k'}| \leq u)\} \right\}_{u=|X_{ik}|} - F^+(|X_{ik}|) \right] \\
&\quad + \frac{1}{2n_i} \sum_{k=1}^{n_i} \left[E \left\{ \frac{1}{n_{i'}} \sum_{k'=1}^{n_{i'}} a(X_{i'k'}) I(v \leq |X_{i'k'}|) \right\}_{v=|X_{ik}|} - E \left\{ \frac{1}{n_{i'}} \sum_{k'=1}^{n_{i'}} a(X_{i'k'}) F^+(|X_{i'k'}|) \right\} \right]. \quad (\text{A.1})
\end{aligned}$$

Now by Lemma 1, the first term of RHS (A.1) equals zero. Also by similar conditioning arguments as in the proof of Lemma 1 using condition (i),

$$E \left\{ \frac{1}{n_{i'}} \sum_{k'=1}^{n_{i'}} a(X_{i'k'}) I(v \leq |X_{i'k'}|) \right\} = E a(X_{i'1}) I(v \leq |X_{i'1}|) = 0, \text{ for any } v,$$

since a is an odd function and the distribution of $X_{i'1}$ is F that is symmetric (under H_0). Similarly,

$$E \left\{ \frac{1}{n_{i'}} \sum_{k'=1}^{n_{i'}} a(X_{i'k'}) F^+(|X_{i'k'}|) \right\} = E a(X_{i'1}) F^+(|X_{i'1}|) = 0.$$

Thus, we have $E(h(V_i, V_{i'}) | V_i) \equiv 0$, which means the corresponding U-statistic is degenerate with mean zero and hence $M^{-1/2} T_{2,M} \xrightarrow{P} 0$.

Suppose $X_{i,1}^*, \dots, X_{i,M}^*$ are one set of within cluster resamples as in Hoffman *et al.* (2001); i.e., given V_i , X_i^* is drawn at random from $\{X_{i1}, \dots, X_{in_i}\}$, $1 \leq i \leq M$. Then

$$\begin{aligned} E \sup_t |\bar{F}^+(t) - F^+(t)| &= E \sup_t \left| E \left\{ M^{-1} \sum_{i=1}^M I(|X_i^*| \leq t) - F^+(t) \mid V_1, \dots, V_M \right\} \right| \\ &\leq E \left\{ \sup_t \left| M^{-1} \sum_{i=1}^M I(|X_i^*| \leq t) - F^+(t) \right| \right\} = O(M^{-1/2}), \end{aligned} \quad (\text{A.2})$$

from known results for the empirical cdf of i.i.d. random variables since, unconditionally, $|X_i^*|$ are i.i.d. $\sim F^+$.

Let $A_\rho = \{ \sup_t |\bar{F}^+(t) - F^+(t)| < \rho \}$, for $\rho > 0$. Since $0 < z_{ij} < 1$,

$$\sup_{i,k} \left| z_{ik} \bar{F}^+(|X_{ik}|) + (1 - z_{ik}) F^+(|X_{ik}|) - F^+(|X_{ik}|) \right| \leq \sup_t |\bar{F}^+(t) - F^+(t)| < \rho,$$

, on A_ρ . Therefore,

$$\begin{aligned} |T_{3,M}| I_{A_\rho} &\leq \left\{ \sup_t |\bar{F}^+(t) - F^+(t)| \right\} \times \\ &\sum_{i=1}^M \frac{1}{n_i} \sum_{k=1}^{n_i} \left| \varphi' \left(z_{ij} \bar{F}^+(|X_{ik}|) + (1 - z_{ij}) F^+(|X_{ik}|) \right) - \varphi'(F^+(|X_{ik}|)) \right| I_{A_\rho} \\ &\leq \left\{ \sup_t |\bar{F}^+(t) - F^+(t)| \right\} \sum_{i=1}^M \frac{1}{n_i} \sum_{k=1}^{n_i} V_\rho^{\varphi'}(F^+(|X_{ik}|)) \\ &\leq 2 \left\{ \sup_t |\bar{F}^+(t) - F^+(t)| \right\} \sum_{i=1}^M \frac{1}{n_i} \sum_{k=1}^{n_i} V_\rho^{\varphi'}(F^+(|X_{ik}|) \wedge (1 - \rho)) \end{aligned} \quad (\text{A.3})$$

Now

$$|M^{-1/2}T_{3,M}| = |M^{-1/2}T_{3,M}|_{I_{A_\rho^c}} + |M^{-1/2}T_{3,M}|_{I_{A_\rho}}, \text{ for any } \rho > 0.$$

Since $P(|M^{-1/2}T_{3,M}|_{I_{A_\rho^c}} \neq 0) \leq 1 - P(A_\rho^c) \rightarrow 0$, for each $\rho > 0$, by (A.2), $|M^{-1/2}T_{3,M}| = o_P(1)$, provided we can show that for any $\varepsilon > 0$, we can find $\rho = \rho(\varepsilon)$ such that

$$\limsup_{M \rightarrow \infty} P\{|M^{-1/2}T_{3,M}|_{I_{A_\rho}} > \varepsilon\} \leq \varepsilon.$$

Now by (A.2) and (A.3), this will hold if we can show the same for the part

$$T_{4,M} = \frac{1}{M} \sum_{i=1}^M \frac{1}{n_i} \sum_{k=1}^{n_i} V_\rho^{\phi'}(F^+(|X_{ik}|) \wedge (1-\rho)).$$

However, by Markov inequality

$$\begin{aligned} P\{|T_{4,M}| > \varepsilon\} &\leq \frac{ET_{4,M}}{\varepsilon} \\ &= \frac{V_\rho^{\phi'}(F^+(|X_{11}|) \wedge (1-\rho))}{\varepsilon}, \text{ by the same arguments as in Lemma 1,} \\ &= \frac{\int_0^{1-\rho} V_\rho^{\phi'}(u) du}{\varepsilon} \rightarrow 0, \text{ as } \rho \downarrow 0, \text{ by condition (ii).} \end{aligned}$$

This ends the proof.

Figure 1. Box-plots of changes in walking distance grouped by cluster size. The median changes by cluster size are indicated along the horizontal axis. The overall (unweighted) median change was 0.04.

Figure 2.

Inverse cluster size weighted histogram of the temporal changes. Each temporal change contributes an amount toward the frequency count of its class that is inversely proportional to the number of evaluations (minus one) by that individual. Here, sample size $M = 176$.

Table 1

Power simulation in a one-way random effects ANOVA for different error distributions. Data are generated from a cluster specific random effects linear model with 50 clusters. Cluster size is binary (5 or 50) and dependent on the realization of the random effects term. Critical values are based on the large sample normal approximations. Each entry in this table is based on a Monte Carlo size of 5000.

Tests	True Marginal mean				
	0	.1	.2	.3	.4
Normal (0,1) Error					
$\varphi(x) = -\log(1 - x)$	0.051	0.262	0.733	0.974	0.999
$\varphi(x) = x$	0.051	0.262	0.731	0.969	0.999
Modified t	0.050	0.265	0.738	0.974	0.999
Beta Error					
$\varphi(x) = -\log(1 - x)$	0.045	0.540	0.974	1.000	1.000
$\varphi(x) = x$	0.046	0.422	0.934	1.000	1.000
Modified t	0.046	0.450	0.959	1.000	1.000
Double Pareto Error					
$\varphi(x) = -\log(1 - x)$	0.051	0.124	0.350	0.601	0.810
$\varphi(x) = x$	0.051	0.181	0.521	0.815	0.950
Modified t	0.042	0.078	0.211	0.388	0.572

Table 2

Power simulation for a multiplicative nonlinear model where the signs of all observations in a cluster are the same and are determined by a latent factor related to the cluster size. The intra-cluster correlations are maintained at two values, γ_1 and γ_2 . Each entry in this table is based on a Monte Carlo size of 5000.

Design	Test	Size		Power				
		$\beta = 1$	$\beta = 0.3$	$\beta = 0.5$	$\beta = 1.5$	$\beta = 2.5$		
10 clusters								
$\gamma_1 = 0.2$	$\varphi(x) = \sqrt{x}$	0.048	0.498	0.234	0.131	0.551		
$\gamma_2 = 0.2$	$\varphi(x) = x$	0.051	0.510	0.244	0.125	0.541		
	Modified t	0.048	0.534	0.266	0.100	0.469		
$\gamma_1 = 0.8$	$\varphi(x) = \sqrt{x}$	0.051	0.501	0.236	0.134	0.556		
$\gamma_2 = 0.8$	$\varphi(x) = x$	0.050	0.466	0.223	0.131	0.522		
	Modified t	0.041	0.449	0.216	0.106	0.454		
$\gamma_1 = 0.2$	$\varphi(x) = \sqrt{x}$	0.050	0.497	0.231	0.139	0.562		
$\gamma_2 = 0.8$	$\varphi(x) = x$	0.049	0.491	0.230	0.134	0.538		
	Modified t	0.045	0.475	0.229	0.118	0.491		
30 clusters								
$\gamma_1 = 0.2$	$\varphi(x) = \sqrt{x}$	0.048	0.956	0.620	0.363	0.968		
$\gamma_2 = 0.2$	$\varphi(x) = x$	0.048	0.942	0.597	0.343	0.959		
	Modified t	0.048	0.908	0.584	0.284	0.924		
$\gamma_1 = 0.8$	$\varphi(x) = \sqrt{x}$	0.048	0.949	0.607	0.358	0.961		
$\gamma_2 = 0.8$	$\varphi(x) = x$	0.048	0.916	0.556	0.323	0.939		
	Modified t	0.042	0.852	0.497	0.265	0.866		
$\gamma_1 = 0.2$	$\varphi(x) = \sqrt{x}$	0.047	0.953	0.614	0.360	0.963		
$\gamma_2 = 0.8$	$\varphi(x) = x$	0.049	0.931	0.572	0.331	0.945		
	Modified t	0.045	0.877	0.523	0.285	0.880		

Results for the NRN data. The null hypotheses correspond to marginal symmetry of changes in the ten minute walking distance (given the initial phase of a patient).

Table 3

Initial phase of patient	Number of patients	Range of clusters	Studentized test statistic with		
			$\phi(x) = x$	$\varphi(x) = \sqrt{x}$	modified t
All	176	2–15	9.74	9.98	8.31
1	58	2–15	5.40	5.63	4.46
2	82	2–11	7.01	7.14	5.89
3	36	2–9	4.50	4.48	4.33