Published in final edited form as:

Biometrics. 2016 June ; 72(2): 432–440. doi:10.1111/biom.12447.

A Rank-Sum Test for Clustered Data when the Number of Subjects in a Group within a Cluster is Informative

Sandipan Dutta¹ and Somnath Datta²

Sandipan Dutta: sandipan.dutta@louisville.edu

¹Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, Kentucky 40202, U.S.A.

²Department of Biostatistics, University of Florida, Gainesville, Florida 32610, U.S.A.

SUMMARY

The Wilcoxon rank-sum test is a popular nonparametric test for comparing two independent populations (groups). In recent years, there have been renewed attempts in extending the Wilcoxon rank sum test for clustered data, one of which (Datta and Satten, 2005, *Journal of the American Statistical Association* 100, 908–915) addresses the issue of informative cluster size, i.e., when the outcomes and the cluster size are correlated. We are faced with a situation where the group specific marginal distribution in a cluster depends on the number of observations in that group (i.e., the intra-cluster group size). We develop a novel extension of the rank-sum test for handling this situation. We compare the performance of our test with the Datta-Satten test, as well as the naive Wilcoxon rank sum test. Using a naturally occurring simulation model of informative intra-cluster group size, we show that only our test maintains the correct size. We also compare our test with a classical signed rank test based on averages of the outcome values in each group paired by the cluster membership. While this test maintains the size, it has lower power than our test. Extensions to multiple group comparisons and the case of clusters not having samples from all groups are also discussed. We apply our test to determine whether there are differences in the attachment loss between the upper and lower teeth and between mesial and buccal sites of periodontal patients.

Keywords

Correlated data; Dental data; Nonparametric tests; Wilcoxon rank-sum test; Within-cluster resampling

1. Introduction

Rank based tests are very popular nonparametric methods for comparing two groups or populations. They are particularly useful when the underlying distributions are suspected to

Correspondence to: Sandipan Dutta, sandipan.dutta@louisville.edu.

Supplementary Materials

Web Appendices referenced in Sections 2, 3, 4, and an R code for implementing the novel rank-sum test are available with the paper at the Biometrics website on Wiley Online Library.

be non-normal. One such widely used test for comparing two groups is the Wilcoxon rank-sum test (Wilcoxon, 1945). One important assumption for applicability of Wilcoxon rank-sum test is that all the observations under the study are independent. However, this assumption may be violated under certain circumstances. In many practical situations we have clustered data where the observations within the clusters are correlated. An example of such clustered data is the data on attachment loss measurement of different teeth of the same individual. Wilcoxon rank-sum test may not be a good option for this type of clustered data. Rosner, Glynn, and Lee (2003) proposed a rank sum test for clustered data for the cases where all the cluster members are from the same group and the correlation structure within a cluster is common across groups. But this approach would not work when the members from a single cluster do not necessarily belong to the same group. Also, this will not maintain the nominal size when the number of observations in a cluster (cluster size) is associated with the outcome of interest from that cluster in some way. This is a case of informative cluster size, where the informativeness comes from the fact that the number of observations in a given cluster (i.e., the cluster size) may be affected by some latent (cluster-specific) factor that affects the outcome variable in that cluster as well. Datta and Satten (2005) proposed a rank-sum test for clustered data that does not make any assumption on the nature of clustering and performs reasonably well in case of informative cluster sizes. But, even the test by Datta and Satten (2005) does not seem to perform well in situations where the outcome of interest belonging to a group in a given cluster appears to be correlated with the number of observations having the same group membership (i.e., the intra-cluster group size) within that cluster. This scenario, following the idea of informative cluster size, can be thought of as informative intra-cluster group (ICG) sizes. This notion of informative ICG sizes can occur in a dental study when one is interested in comparing the nature of attachment losses of teeth between the upper and lower jaws. This is because, the difference (if any) between the nature of attachment loss of the teeth of upper and lower jaws can be suspected to be associated with the difference between the number of teeth present in the upper and lower jaws. Another interesting situation, where informative ICG sizes come into play, can be found in studies relating to hereditary diseases. In many genetic studies it has been observed that an inherited disease is often diagnosed at a younger age in a later generation than that in an earlier generation. This phenomenon of an earlier onset of a disease in each successive generation of a family, called anticipation, is prevalent in diseases like non-Hodgkins lymphoma, breast and ovarian cancer, Huntington's disease among others. In case of testing for this anticipation phenomenon of a disease, or in general, to test whether the age at onset of a disease differs in two different generations of large pedigrees, an interesting information might be the number of affected individuals, belonging to a certain age interval, that are present in each of the two generations under study. "Affected" individuals include subjects who are currently diseased at the time of the study as well as those who were known to be diseased at some point of time before the study. If we find that there is a large difference in the number of affected individuals (belonging to that certain age group) between the two generations, then one may relate this difference to be associated to the difference in the onset age between the two generations. So, this might be a case of informativeness in the number of subjects (affected individuals in a certain age interval) in a group (generation) within a cluster (a large pedigree). Motivated by these, we develop a rank-sum test for comparing the marginal distribution of outcomes from different groups

under the cases of informative ICG sizes. We return to the example of tooth attachment loss in the application section.

Extending the idea of within-cluster resampling (Hoffman, Sen, and Weinberg, 2001; Williamson, Datta, and Satten, 2003; Datta and Satten, 2005; Datta and Satten, 2008), we obtain a rank-sum test for clustered data, with observations from both groups being present in every cluster. Our resampling scheme is an extension of the usual within-cluster resampling because instead of resampling one observation at random from each cluster, we first resample one group membership (out of the two possible groups) for a cluster and then resample an outcome from that group belonging to that cluster. We repeat this resampling for each cluster and obtain a rank-sum statistic based on the resampled observations. Then, following the approaches of Datta and Satten (2005), we derive our test statistic by averaging the rank sum statistic over all possible choices of the resampled observations given the data. After constructing our test, we compare it with three other existing tests, including the test by Datta and Satten (2005), under naturally occurring simulation scenarios of informative ICG sizes. We show that our test maintains the correct size under the null hypothesis of marginal symmetry, unlike the test by Datta and Satten (2005). Moreover, our test has better power performances than the three other tests under this simulation study. Besides, we show that our test also has acceptable size and power in simulation settings where we have informative cluster sizes but non-informative sizes and also in simulation scenarios ICG having both noninformative cluster and ICG sizes. Additionally, we extend our test statistic for two group comparison to the cases when some of the clusters may have observations from only one of the two groups (i.e., the intra-cluster group structures are incomplete). We present a simulation study to show that our test still maintains the appropriate size and has reasonable power under this scenario of incomplete ICG structures within a cluster. We also discuss an extension to our test where there are observations from more than two groups in every cluster.

The rest of the article is organized as follows. In Section 2, we introduce the necessary notations, formulate our testing problem, and develop a test statistic for comparing the outcomes from two groups when outcomes from both the groups are present in every cluster. This section also contains some variant forms of our test under different clustered data settings including expressions of the test statistic with other required quantities that generalize our new rank-sum test to more than two groups and an extension of our test statistic for two group comparison where some clusters may have observations from only one of the two groups. Section 3 contains simulation studies that evaluate the empirical performance of our test compared to three other tests on the basis of size and power. In Section 4, we return to the dental data discussed before and apply our testing procedure to compare the difference between the tooth attachment loss in the upper and lower jaws. Besides this, we apply our test for a different comparison in dental data where some other rank-sum tests can also be applied. The article ends with a discussion in Section 5. Detailed steps for deriving our test statistic are discussed in the Web Appendix (Supplementary Web Materials).

2. Notations, Formulation of the Problem and Proposed Test Statistic

Let M denote the number of clusters and let X_{ik} denote the k^{th} observation in the i^{th} clusters $1 \leq k \leq N_i, 1 \leq i \leq M$, where N_i denotes the number of observations in the i^{th} cluster. Let G_{ik} be the indicator denoting the binary group membership (0 or 1) of the k^{th} observation in the i^{th} cluster. Thus the entire data set consists of $\{\mathbb{V}_i: 1 \leq i \leq M\}$, with $\mathbb{V}_i = \{N_i, X_{ik}, G_{ik}, 1 \leq k \leq N_i\}$ corresponding to the i^{th} cluster. Also, let N_{i1} and N_{i0} be the numbers of observations in the i^{th} cluster belonging to group 1 and group 0, respectively. Thus, we have $N_{i1} + N_{i0} = N_i$. We consider the possibility that the cluster size N_i as well as the group memberships G_{ik} are random (and thus, so are the $N_{id}, d = 0, 1$). The members in a cluster could have an arbitrary dependence structure; however, members in different clusters are statistically independent and hence the entire $\mathbb{V}_i$ and $\mathbb{V}_j$ are independent. For mathematical convenience, we further assume that $\mathbb{V}_i, 1 \leq i \leq M$, are independent and identically distributed (iid).

The null hypothesis we consider is that the observations from the two groups follow the same marginal distribution. Mathematically, it is written as

$$H_0: P(X_{ik} \leq x | G_{ik}=0) = P(X_{ik} \leq x | G_{ik}=1) (= \mathcal{F}(x), \text{ say}), \text{ for } \text{all } x.$$

However, the empirical analogue of the above “group specific” (e.g., conditional) marginal distributions can be constructed in three possible ways resulting in three different statistical comparisons:

i.

$$\hat{\mathcal{F}}_1(x|d) = \frac{\sum_{i=1}^M \sum_{k=1}^{N_i} I(X_{ik} \leq x, G_{ik}=d)}{\sum_{i=1}^M \sum_{k=1}^{N_i} I(G_{ik}=d)}, d=0, 1;$$

ii.

$$\hat{\mathcal{F}}_2(x|d) = \frac{\sum_{i=1}^M \frac{1}{N_i} \sum_{k=1}^{N_i} I(X_{ik} \leq x, G_{ik}=d)}{\sum_{i=1}^M \frac{1}{N_i} \sum_{k=1}^{N_i} I(G_{ik}=d)}, d=0, 1;$$

iii.

$$\hat{\mathcal{F}}_3(x|d) = \frac{\sum_{i=1}^M \frac{1}{2N_{id}} \sum_{k=1}^{N_i} I(X_{ik} \leq x, G_{ik}=d)}{\sum_{i=1}^M \frac{1}{2N_{id}} \sum_{k=1}^{N_i} I(G_{ik}=d)}, d=0, 1.$$

Note that (i) represents the (empirical) distribution of group d ($d = 0, 1$) data values in the entire sample irrespective of their cluster membership. Calculation (ii) is based on sampling a single paired (e.g., (X, G)) observation from each cluster. In other words, $\hat{\mathcal{F}}_2(\cdot|d)$ represents the conditional distribution of a typical outcome value X_{iJ_i} for a typical cluster i , given the corresponding group membership G_{iJ_i} equals d . Here, J_i is a discrete uniform on $\{1, \dots, N_i\}$. Calculation (iii) is based on computing the proportion of outcomes belonging to group d in a

typical cluster i which are less than or equal to x and then taking the average of these proportions over all the clusters. Each of quantities in the right hand sides of (i), (ii), and (iii), can be written as an estimate of $P(X_{ik} \leq x, G_{ik} = d)/P(G_{ik} = d)$, but the difference lies in construction of the estimates of the probabilities. In (i) the probabilities are estimated by pooling all the observations together irrespective of their cluster membership, while in (ii), and (iii), the estimates are constructed by conditioning on N_i and N_{id} respectively. Every outcome, belonging to group d and having value less than x , contributes equally in the construction of $\hat{\mathcal{F}}_1$, but in constructions of $\hat{\mathcal{F}}_2$ and $\hat{\mathcal{F}}_3$ we have different contributions from the different outcomes depending on their cluster memberships.

Let $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3$ be the distribution functions which are estimated by $\hat{\mathcal{F}}_1, \hat{\mathcal{F}}_2, \hat{\mathcal{F}}_3$ respectively. When the cluster sizes as well as the ICG sizes formed by the two groups within each cluster are not suspected to be associated to the outcome variable in any way, then hypotheses involving $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3$ become equivalent and one can test any one of these three hypotheses. If there is some association between the cluster size and the outcome variable in that cluster, one can think of testing hypothesis involving $\mathcal{F}_2$ for appropriate comparison. This is a situation of informative cluster sizes. Again, if the ICG sizes formed by the two groups in a cluster appears to be correlated (even after conditioning on the overall cluster size) with the outcomes from the respective groups in that cluster, one may think of testing hypothesis comprising of distribution $\mathcal{F}_3$ instead of $\mathcal{F}_1$ and $\mathcal{F}_2$ to get more meaningful results. We can refer to this as a case of informative ICG sizes. In the absence of this informativeness in the ICG sizes, one can test the null hypotheses of equality of marginal distributions involving any one of the marginal distributions $\mathcal{F}_2$ and $\mathcal{F}_3$, possibly leading to similar conclusion in each case.

In this paper we are interested in comparing $\mathcal{F}_3$ in the two groups when the ICG sizes are potentially informative. Currently, no rank based tests are available for testing group differences for clustered data that takes into account the informativeness of the ICG sizes formed by the groups under study. We denote the common marginal distribution under the null hypothesis as $\mathcal{F}(\cdot)$. It is perhaps worth pointing out that the estimation of the marginal regression parameters via weighted estimating equations in presence of informative ICG size has been considered by Huang and Leroux (2011).

2.1 Development of the Test Statistic

For the sake of simplicity, let us relabel the observations according to their group membership within each cluster in the following way. In the i^{th} cluster, let

$\{X_{i1}^{(1)}, X_{i2}^{(1)}, \dots, X_{iN_{i1}}^{(1)}\}$ represent the set of observations belonging to the group indexed by

1, while $\{X_{i1}^{(0)}, X_{i2}^{(0)}, \dots, X_{iN_{i0}}^{(0)}\}$ represents the set of observations belonging to the group indexed by 0. We denote these sets as $\underline{X}_i^{(1)}$ and $\underline{X}_i^{(0)}$ respectively. Thus, $\underline{X}_i^{(1)}$ and $\underline{X}_i^{(0)}$ form a partition of $\underline{X}_i = \{X_{i1}, \dots, X_{iN_i}\}$, the set of all observations in cluster i . The number of

observations belonging to the set $\underline{X}_i^{(d)}$ ($d=0, 1$) is the intra-cluster group size of group d in the i^{th} cluster. Till the end of the Section 2.2, we would assume that at least one observation from each group is present in every cluster. In this Section, this assumption means that $N_{i1} >$

0 and $N_{j0} > 0$ with probability one for every cluster i . A relaxation of this condition is discussed in Section 2.2.

Our test statistic, for testing the hypothesis involving marginal distributions $\mathcal{F}_3$ as estimated in (iii), can be generated from a resampling scheme which is an extension of the within-cluster resampling (WCR). An outline of the resampling scheme is as follows: For each cluster i , let us resample group membership as G_i^* , where G_i^* takes value 0 or 1 with equal probability $\frac{1}{2}$. If $G_i^* = 1$, we resample one observation for the i^{th} cluster from the set of observations $\underline{X}_i^{(1)}$ and name it X_i^* . If $G_i^* = 0$, resample X_i^* from the set $\underline{X}_i^{(0)}$.

The fact that the outcomes are resampled from the subsets formed by the two groups in a cluster and not from the whole cluster makes this resampling scheme different from the usual WCR technique. Now, this resampling gives us M pairs of independent observations (X_i^*, G_i^*) . If R_i^* be the rank of X_i^* among the set $\{X_j^*, 1 \leq j \leq M\}$, i.e.,

$R_i^* = 1 + 0.5 \left\{ \sum_{j \neq i} I(X_j^* \leq X_i^*) + \sum_{j \neq i} I(X_j^* < X_i^*) \right\}$, then the Wilcoxon rank sum statistic based on these M pairs of resampled observations (X_i^*, G_i^*) would be of the form :

$S^* = \sum_{i=1}^M G_i^* R_i^*$. One can use S^* as a valid test statistic and carry out the test based on S^* . But that test would be inefficient as the test statistic would depend too much on one particular observation chosen from each cluster. So to get rid of the imposed randomization due to resampling we propose a test statistic based on earlier approaches of Williamson et. al (2003), Datta and Satten (2005), and Datta and Satten (2008), that corresponds to averaging S^* over all possible choices of (X_i^*, G_i^*) values given the data.

Thus, our test statistic is $T = E(S^* | X, G)$, where $X = \{X_{ik} : 1 \leq k \leq N_i; 1 \leq i \leq M\}$ and $G = \{G_{ik} : 1 \leq k \leq N_i; 1 \leq i \leq M\}$. We can calculate the theoretical expression of T . After some necessary steps, a convenient expression of T (see Web Appendix A for the detailed steps) turns out to be

$$T = \sum_{i=1}^M \left(\sum_{k=1}^{N_{i1}} \frac{1}{2N_{i1}} \left[1 + \frac{1}{2} \sum_{j \neq i} \left\{ F_j(X_{ik}^{(1)}) + F_j(X_{ik}^{(1)} -) \right\} \right] \right),$$

$$\text{where } F_j(x) = \frac{1}{2N_{j1}} \sum_{h=1}^{N_{j1}} I(X_{jh}^{(1)} \leq x) + \frac{1}{2N_{j0}} \sum_{h'=1}^{N_{j0}} I(X_{jh'}^{(0)} \leq x).$$

Besides T , we need to know its expected value $E(T)$ and its variance estimate $V(\hat{T})$ to properly carry out inference based on T . To get $E(T)$ we note that $E(T) = E(S^*)$. The unconditional expectation of S^* can be calculated easily through conditioning on the vector of group membership indicator $G^* = (G_1^*, G_2^*, \dots, G_M^*)$. So we get,

$$E(T) = E(S^*) = E(E(S^* | G^*)) = E\left(\sum_{i=1}^M G_i^* \left(\frac{M+1}{2}\right)\right) = \left(\frac{M+1}{2}\right) \sum_{i=1}^M \frac{1}{2} = \frac{M(M+1)}{4}.$$

The next step is to find a variance estimate $\hat{V}(T)$. To get the variance estimate of T , we employ the jackknife technique. Here the clusters can be thought of as iid units and thus we can use a ‘delete-1-cluster’ jackknife approach to get the necessary results. Mathematically, this can be formulated as follows. Let T_{-i} be the value of the statistic T calculated after deleting the i^{th} cluster. Let us define, $T_i^* = T - T_{-i}$. Then the estimate of variance of T , which is the jackknife variance estimate, is given by

$$\hat{V}(T) = \hat{V}_{JK} = M \cdot \text{Var}(T_i^*) = \frac{M}{M-1} \sum_{i=1}^M (T_i^* - \bar{T}^*)^2.$$

Now that we have the expressions for T , $E(T)$, $\hat{V}(T)$, we can carry out the testing using the absolute value of the standardized statistic $Z = (T - E(T)) / (\hat{V}(T))^{1/2}$.

The asymptotic distribution of Z is established through the following theorem. An outline of its proof is given in the Web Appendix B.

THEOREM 1 (Asymptotic normality)—Under H_0 , as $M \rightarrow \infty$, $Z \xrightarrow{d} N(0, 1)$ under certain regularity conditions of a Lindeberg Central Limit Theorem.

The p-value for the test is computed as the probability that, under H_0 , the absolute value of the Z-statistic exceeds its observed value in magnitude. We would reject the null hypothesis H_0 at a $100\alpha\%$ level of significance if the p-value is less than α .

Till this point we have assumed the existence of only two groups in every cluster. In Web Appendix C, we have discussed a more general situation where there are m groups in every cluster, such that $m > 2$.

2.2 Extension to Incomplete Intra-cluster Group Structure in One or More Clusters

In case of binary grouping, (i.e., $G_{ik} = 0$ or 1), we have assumed that there is at least one observation from each group in every cluster. In practice, one may encounter a few clusters (not all) with one group of observations completely missing. In other words, there may be some clusters having outcomes from only one of the two possible groups. We call such a case as incomplete informative intra-cluster group structure within a cluster. The hypothesis of interest remains the same, viz., whether the marginal distributions of outcomes are same for the two groups. We cannot directly apply the test statistic in the form described in Section 2.1 to this setting. This is mainly because of the fact that the test statistic developed in Section 2.1 is only applicable under the assumption that outcomes from both groups are available within each cluster. We extend the approach described in Section 2.1, to get a valid test statistic in this setting.

Here we follow the same notations as described in Section 2.1. In cases of incomplete ICG structures within a cluster, the empirical analogue of the “group specific” marginal distributions of our interest can be constructed as a modification of $\hat{\mathcal{F}}_3$ as

$\hat{\mathcal{F}}_4(x|d) = \sum_{i=1}^M W_{id} \sum_{k=1}^{N_i} I(X_{ik} \leq x, G_{ik}=d) / \{ \sum_{i=1}^M W_{id} \sum_{k=1}^{N_i} I(G_{ik}=d) \}$, where $W_{id} = (2N_{id})^{-1}$, or N_{id}^{-1} , or 0, according to whether the i^{th} cluster has observations from both groups, the d^{th} group only, or not.

We extend the idea of within cluster resampling also to this setting to get a valid test statistic. (1) If both $N_{i1} > 0$ and $N_{i0} > 0$, group membership is resampled as G_i^* , where G_i^* takes value 0 or 1, with equal probability $\frac{1}{2}$. If $G_i^*=0$, resample X_i^* from $\underline{X}_i^{(0)}$; otherwise, if $G_i^*=1$, resample X_i^* from $\underline{X}_i^{(1)}$. (2) If $N_{i1} = 0$ and $N_{i0} = N_i > 0$, we resample X_i^* from $\underline{X}_i^{(0)}$ and have $G_i^*=0$. Here $\underline{X}_i^{(0)}$ is same as $\underline{X}_i$ as the set $\underline{X}_i^{(1)}$ is an empty set. (3) If $N_{i0} = 0$ and $N_{i1} = N_i > 0$, we resample X_i^* from $\underline{X}_i^{(1)}$ and have $G_i^*=1$. Here $\underline{X}_i^{(1)}$ is same as $\underline{X}_i$ as the set $\underline{X}_i^{(0)}$ is an empty set.

To obtain our test statistic T in this case, we proceed in the same way as in Section 2.1. With R_i^* being the rank of X_i^* among the set $\{X_j^*, 1 \leq j \leq M\}$, we obtain $S^* = \sum_{i=1}^M G_i^* R_i^*$, the Wilcoxon rank sum statistic based on the M pairs of resampled observations (X_i^*, G_i^*) . Then, our proposed test statistic T is calculated as $T = E(S^* | X, G)$. After some algebra (see Web Appendix D) we obtain T as

$$T = \sum_{i=1}^M \sum_{k=1}^{N_{i1}} \frac{1}{2N_{i1}} [(I(N_{i1} > 0, N_{i0} > 0)) + 2I(N_{i1} > 0, N_{i0} = 0) + \frac{I(N_{i1} > 0, N_{i0} > 0) + 2N_{i1} \cdot I(N_{i0} = 0)}{2} \sum_{j \neq i} \{F'_j(X_{ik}^{(1)}) + F'_j(X_{ik}^{(1)} -)\}] ,$$

where

$$F'_j(x) = \left\{ \frac{1}{2N_{j1}} \sum_{h=1}^{N_{j1}} I(X_{jh}^{(1)} \leq x) + \frac{1}{2N_{j0}} \sum_{h'=1}^{N_{j0}} I(X_{jh'}^{(0)} \leq x) \right\} I(N_{j1} > 0, N_{j0} > 0) + \{1 - I(N_{j1} > 0, N_{j0} > 0)\} \left\{ \frac{1}{N_j} \sum_{h=1}^{N_j} I(X_{jh} \leq x) \right\} .$$

The expected value of the test statistic is estimated to be

$$\hat{E}(T) = \frac{(M+1)}{4} \sum_{i=1}^M \{I(N_{i1} > 0, N_{i0} > 0) + 2I(N_{i1} > 0, N_{i0} = 0)\} .$$

Now, to find the estimated variance V of $T - \hat{E}(T)$, we use the same ‘delete-1-cluster’ jackknife approach described in Section 2.1. Finally, as in Section 2.1, we carry out the

testing using the standardized Z-statistic $Z = \{T - \hat{E}(T)\} / \{\hat{V}\}^{1/2}$, that has asymptotic $N(0, 1)$ distribution under H_0 .

3. Simulation Results

In this Section we present three simulation studies corresponding to the tests discussed in the Sections 2.1 and 2.2. In the simulation scenario 1, we consider clustered observations such that every cluster has outcomes from both the groups. In each cluster, the number of observations belonging to group 1 and the number of observations belonging to group 0s that is the two ICG sizes, are both influenced by some latent factor, that also influences the outcomes in that cluster. Also, the distributions of the two ICG sizes, within each cluster, differ between themselves. So, there is some association between the ICG sizes and outcomes in a given cluster (even after conditioning on the overall cluster size) and we can think of this as informative ICG sizes. Under this simulation scenario, we compare the performances of four tests, namely, (1) our new rank sum test developed in Section 2.1, (2) the test by Datta and Satten (2005), (3) the naive Wilcoxon rank sum test assuming all the observations as iid and ignoring their cluster membership, and (4) the signed rank test taking cluster averages for each group of observations. Further, each test was carried out under three different choices of the number of clusters (M), namely, 30, 50 and 150. In simulation scenario 2, we generate a setting that closely represents the dental setting discussed in Section 1. Basically, the idea is to have a clustered data with informative ICG sizes, where the number of units belonging to each group in a cluster cannot exceed a certain value. Under this setting we compare the four tests (1)–(4) for 50 clusters. In scenario 3, we again consider informativeness in the ICG sizes, but we do not restrict ourselves to the condition that observations from both the groups have to be present in each cluster. In other words, we include the cases of incomplete ICG structures within a cluster for which our test statistic developed in Section 2.2 looks appropriate. We investigate the performance of this new test for a simulation model with 30 clusters under scenario 4.

Additionally, in Web Appendix E we consider two more simulation scenarios (Scenario 4 and Scenario 5), where we compare the four tests (1)–(4) under situations such that either the ICG sizes or both the ICG sizes and the cluster sizes are noninformative.

Performances of all the tests are evaluated on the basis of their sizes (nominal $\alpha = 0.05$) and power values. These are estimated by the proportion of 3,000 Monte Carlo iterates in which null hypothesis is rejected.

3.1 Simulation Scenario 1

Let M be the number of clusters (fixed). For a typical cluster i , we define, N_{i1} as the number of observations from group 1 in the i^{th} cluster, N_{i0} as the number of observations from group 0 in the i^{th} cluster, a_i as the random cluster effect due to the i^{th} cluster. In the i^{th} cluster, we generate a_i from Normal(0, 0.25) distribution, N_{i1}^* from Poisson($10 + 5a_i$) distribution where $N_{i1} = N_{i1}^* + 1$, N_{i0}^* is generated from Poisson($10 + 5a_i^2$) such that $N_{i0} = N_{i0}^* + 1$. Also, we know that $N_i = N_{i1} + N_{i0}$. Let G_{ij} be the group indicator of the j^{th} observation in the i^{th} cluster. We assign $G_{ij} = 0$ for $1 \leq j \leq N_{i0}$, while $G_{ij} = 1$ for $N_{i0} + 1 \leq j \leq N_i$. We generate Y_{ij} the j^{th}

outcome in the i^{th} cluster, through a random effects model as $Y_{ij} = 0.5 + a_i + e_{ij}$, $1 \leq j \leq N_i$, $i = 1, \dots, M$, such that if $G_{ij} = 0$, then $e_{ij} \sim \text{Normal}(0, 0.3)$, while if $G_{ij} = 1$, then $e_{ij} \sim \text{Normal}(\delta, 0.3)$. Under the null model, $\delta = 0$.

Performances of the four tests (1)–(4) are summarized in Table 1 for three choices of M , namely, 30, 50 and 150.

Table 1 illustrates a number of points. Our new test closely maintains the nominal size and is sufficiently strong in terms of power even under small effect sizes. The rank sum test proposed by Datta and Satten (2005) and the standard Wilcoxon rank sum test have grossly inflated size and very low power compared to our test for all three choices of the number of clusters. The size of the cluster average signed rank test tends to be close to the nominal size under this simulation scenario. Its power is also close to our test, though a bit less in almost all cases. Although the clustered average signed rank test appears to be a good competitor of our test in this simulation scenario, one can acknowledge the fact that the distribution of the average of independent and identical random variables is not always same as that of the individual variables. Thus, it is expected that the cluster average signed rank test is not a good choice for testing the hypothesis of our interest and this fact might be evident if we have widely different within ICG sizes each cluster.

3.2 Simulation Scenario 2

This simulation setting is carried out to mimic the setting of dental study mentioned in Section 1, where the number of units (teeth) within a cluster (mouth of an individual) cannot exceed 32. This can be generalized for any study where the cluster sizes or the ICG sizes are bounded.

This simulation scenario is almost same as that described in Section 3.1, the only difference being that both the ICG sizes within each cluster are less than or equal to 16, such that the cluster size cannot exceed 32. Following the same notations for the quantities in 3.1, in the i^{th} cluster, we generate a_i from $\text{Normal}(0, 0.25)$, N_{i1}^* from $\text{Poisson}(10 + 5a_i)$ such that $N_{i1} \leq 16$, N_{i0}^* from $\text{Poisson}(10 + 5a_i^2)$ such that $N_{i0} \leq 16$. So, we have $N_i = N_{i1} + N_{i0} \leq 32$. Apart from these, G_{ij} , e_{ij} and the outcome Y_{ij} are generated in the same manner as in simulation scenario 1. Table 2 compares the four tests (1)–(4) under this simulation scenario with the number of clusters (M) as 50, and the results are similar to the results obtained from simulation scenario 1. Table 2 shows that our new test closely maintains the nominal size and has substantial power under a variety of effect sizes. The rank-sum test proposed by Datta and Satten (2005), as well as the standard Wilcoxon rank sum test, has highly inflated size. The clustered average signed rank test, just like in simulation scenario 1, apparently maintains the nominal size and has substantial power. But, as mentioned before in Section 3.1, theoretically it is not a good choice for testing the hypothesis of our interest.

3.3 Simulation Scenario 3

This simulation scenario is almost similar to that described in Section 3.1, the only difference being that the ICG sizes within each cluster are not restricted to be strictly positive always. Following the same notations for the quantities in 3.1, in the i^{th} cluster, we

generate a_i from $\text{Normal}(0, 0.25)$, N_{i1} from $\text{Poisson}(10 + 5a_i)$, N_{i0} from $\text{Poisson}(10 + 5a_i^2)$. We have $N_i = N_{i1} + N_{i0}$. Apart from these, G_{ij} , e_{ij} , and the outcome Y_{ij} are generated in the same manner as in simulation scenario 1. Evidently, observed values of any of the ICG sizes N_{i1} and N_{i0} in the i^{th} cluster can be 0, as long as $N_i > 0$.

In Table 3, we evaluate the empirical size and power of our test, developed in Section 2.2, with the choice of $M = 30$. Thus, from Table 3, we see that our test closely mimics the nominal size and has moderate to high power under different effect sizes.

4. Application to Dental Data

We consider data from the Piedmont 65+ Dental study by Beck *et al.* (1990). This study examined two older populations, urban whites and urban and rural blacks. The Piedmont Health Study of the Elderly by Blazer and George (2004), which was the parent study for this Piedmont 65 + Dental Study, was a longitudinal study of the health status of a stratified, clustered, random sample of people aged 65 and over in five contiguous North Carolina counties. The Piedmont 65+ Dental Study used the data available from the parent study while collecting additional information. For the Piedmont 65+ Dental Study, we have the gingival recession and pocket depth measures for all teeth present in the mouth, at baseline, 18, 36 and 60 months, respectively. Attachment level scores (attachment losses) were computed from the gingival recession and pocket depth measures. Also, all these clinical measures were computed for two sites, buccal and mesial, for every tooth measured. A number of additional covariates were also available which are ignored for the present marginal analyses. The number of subjects observed varied across the four data points. This may be because, being a study involving elderly population, many subjects who were reported at the beginning of the study failed to come back at later time points of the study. For our illustration, we investigate the baseline and 18 month data cross-sectionally.

Attachment loss is a common problem associated with periodontal diseases in elderly population, often indicating the severity of certain diseases. It has been suggested in some studies that the nature of attachment loss varies across the different surfaces of a tooth. Suspecting one such possibility, it may be interesting to identify whether the distributions of attachment loss scores are same for the buccal and mesial surfaces of teeth. Since the outcomes (attachment level scores) from the units (teeth surfaces) within a cluster (individual) are correlated, while that from the units between different clusters are independent, the data fall into the category of the type of clustered data we are interested in. In addition since the cluster size (number of teeth surfaces an individual has) may indicate the overall oral health, the cluster size might be associated to the outcome of interest (attachment loss score). We apply our new test and the test of Datta and Satten to investigate possible differences in the distributions of attachment loss at the buccal and mesial sites (the two groups under study) to data at baseline involving 697 subjects with at least one tooth. A significant difference was obtained for the novel test ($Z = -10.29$, $p\text{-value} = 7.92 \times 10^{-25}$) and for the Datta and Satten test ($Z = -9.40$, $p\text{-value} = 5.56 \times 10^{-21}$). So, our new test and the test by Datta and Satten lead to the same conclusion but with different p-values. We then consider the same testing problem but with the data for 18 month (with 496 available subjects) where, again, significant difference was obtained using the new test ($Z = -11.49$,

p-value = 1.48×10^{-30}) as well as the test by Datta and Satten ($Z = -9.94$, p-value = 2.72×10^{-23}). Overall, we conclude that the distribution of the attachment loss of teeth differs between the mesial and buccal sites. Also, we see that our new test gives consistent result in a situation where the test by Datta and Satten appears to be valid as well. Plots of the empirical cumulative distribution functions ($\hat{\mathcal{F}}_3(\cdot)$) of attachment scores in the two groups (buccal and mesial) are shown for both the baseline data and the 18-month data in Figure 1 and Figure 2 respectively. Some indications regarding the significant difference in the distributions of attachment scores between buccal and mesial sites can be obtained from these figures. In addition, plots of the empirical mass functions for mesial and buccal attachment loss scores at baseline study are given in Web Figure 1 (in Web Appendix F) that explain the substantial differences between the mesial and buccal attachment loss scores at the low score values of 1 and 2. Incidentally, these two scores together constitute more than half of the observed scores for the population under study. To calculate the effect size we use the following approach: if $\underline{Y}^{(0)}$ and $\underline{Y}^{(1)}$ denote the sets of mesial and buccal attachment scores, such that the test statistic $T = \mathcal{T}(\underline{Y}^{(0)}, \underline{Y}^{(1)})$, and τ be a real number such that $T = \mathcal{T}(\underline{Y}^{(0)}, \underline{Y}^{(1)} + \tau)$, then the effect size is estimated by the absolute value of τ^* , where $\tau^* = \sup \{ \tau : T - E(T) = 0 \}$. For both the baseline and 18 months data unstandardized effect size turns out to be approximately 0.5.

Another interesting question, as discussed previously in Section 1, would be whether the distributions of attachment loss scores differ between the teeth of upper and lower jaws. To investigate this fact using the same data, we have considered attachment loss at the mesial site of tooth, although one can also pose the same question with the buccal site. The null hypothesis here is that the distribution of attachment loss at the mesial site of a tooth is the same for the upper and lower jaws. Here the setting for this problem is quite similar to that of the previous problem. The difference is that in this setting the mesial site attachment loss score (outcome) of a tooth (unit) in any particular jaw (group) of an individual (cluster) may be related to the number of teeth present in that jaw of that individual. So, we may have some informativeness in the ICG size (number of teeth present in a jaw of an individual) even after conditioning on the cluster sizes. We consider the 60 month data for this analysis with 292 available subjects at that point. This data falls under the category of clustered data with some clusters having incomplete ICG structures, as described in Section 2.2, because there are a few subjects (clusters) who have teeth (units) in only one of the two jaws (groups). Our new test, developed in Section 2.2, is the only test that can be used to test the hypothesis under this setting and it gives a p-value of 4.06×10^{-5} ($Z = 4.10$). Thus, we conclude that there is a significant difference between the distributions of the attachment loss at the mesial sites of the upper and lower jaws. The estimated effect size, estimated like before, comes out to be around 3.0 units for this data. Web Figure 2 (in Web Appendix F) shows the empirical cumulative distribution functions ($\hat{\mathcal{F}}_4(\cdot)$) for the attachment loss scores of upper and lower sets of teeth.

5. Discussions

For clustered data with informative cluster sizes, the ordinary rank-sum test assuming independent observations can be biased as indicated in a simulation study in Section 3. The rank-sum test by Datta and Satten, which compares group-specific marginal distributions $\mathcal{F}_2$,

appears to be a valid test under informative cluster sizes. But when an outcome from a group d ($d = 0, 1$) in a typical cluster depends on the number of observations from the group d in that cluster, we have informativeness in the ICG sizes formed by the two groups. As discussed earlier in Section 1 and Section 4, this type of clustered data with informative ICG sizes are common in dental studies. Simulation studies from Section 3 indicate that even the rank-sum test by Datta and Satten (2005) has inflated size under this scenario of informative ICG size. There are no rank based tests in current literature that address this issue of informative ICG sizes. Thus, our main focus was to develop a rank-sum test for clustered data which works under this scenario of informative ICG sizes. This has led us to compare group-specific marginal distribution $\mathcal{F}_3$ that gives equal weights to each cluster (treating cluster as the basic sampling unit), but the weight given to an outcome from group d in a cluster depends on the number of observations from group d in that cluster. This is in contrast with $\mathcal{F}_2$ where the weight given to an outcome from a typical cluster depends on the number of outcomes in that cluster ignoring the information on the group membership of that outcome. Thus, the question of importance is which marginal distribution should be considered in testing hypothesis. It appears that comparing $\mathcal{F}_3$ may be more meaningful under informative ICG sizes and through a number of simulation settings, we have showed that our test maintains the nominal size and has substantial power in clustered data with informative ICG sizes. Even when the ICG sizes are not informative, simulation studies from Web Appendix E reveal that our test closely maintains the nominal size and has acceptable power when compared to other rank tests based on $\mathcal{F}_2$ or $\mathcal{F}_1$.

As we consider clustered data, we may, in practice, encounter a few clusters which have outcomes from only one of the two groups under study. In that case, there are two possible ways of addressing this issue. One simple way is to ignore the clusters which do not have outcomes from both the groups and carry out the test, developed in Section 2.1, based on the remaining clusters. But, oftentimes, it is suspected that the information on the outcome of interest may be different between clusters with incomplete ICG structures (i.e., clusters with observations from one of the two groups) and clusters having both groups of observations. Keeping this in mind, we extended our test, in Section 2.2, to account for the clusters with incomplete ICG structures, so that we effectively use all the information present in the data. A simulation study showed that our test has the correct size and substantial power for a model accommodating incomplete ICG structures with informative ICG sizes. But, one can expect the power of this test to be low compared to that of the test involving only clusters with complete ICG structures. Therefore, in presence of a few clusters with incomplete ICG structures among a large number of clusters, it might be important to decide beforehand whether to apply the test developed in Section 2.1 ignoring a few clusters or to use the test from Section 2.2 keeping the full data. In case of clustered data where the outcomes within the same cluster belong to the same group, our test statistic reduces to that of Datta and Satten (2005), and, thus, will have superior size and power performance than the rank-sum test by Rosner et al. (2003) when the correlation structure within a cluster depends on the group membership.

Sometimes, when testing for group effect in outcomes from clustered data, one can expect the presence of some additional covariate(s) unrelated to the grouping factor. In such cases these additional covariates (confounders) may act as nuisance factors in comparing the

group-specific marginal distributions of the outcomes. For example, suppose we have a linear regression of the form

$$Y_{ij} = \beta_1 X_{1ij} + \beta_2 X_{2ij} + \beta_3 X_{3ij} + \varepsilon_{ij}, 1 \leq i \leq M, 1 \leq j \leq N_i.$$

Here Y_{ij} is the outcome of the j^{th} observation in the i^{th} cluster, is X_1 the binary indicator variable taking value 1 or 0 according to the group membership, X_2 and X_3 are the confounders (unrelated to the group membership) and ε is the random error following some unknown distribution F_ε . To compare the group-specific marginal distributions of the outcomes, one may want to test the null hypothesis $H: \beta_1 = 0$ against the alternative hypothesis $K: \beta_1 \neq 0$. But, if the distributions (unknown) of the confounders are different from that of the random error and also among themselves, then the rank tests based on the outcome Y can be misleading. This is, in general, true for any regression model involving confounders. To overcome this, one, often, uses aligned rank tests (see, e.g., Hájek, Šidák, and Sen, 1999, Section 10.1.2). The basic idea involves estimation of the (nuisance) parameters relating to the confounders through some appropriate rank statistics, formation of aligned observations (residuals) by plugging in the estimates and then developing a rank test based on the aligned observations. In presence of informative ICG size, one can extend the resampling technique discussed in this article to formulate suitable rank based statistics for estimating the nuisance parameters and testing the appropriate (sub)hypothesis under aligned rank tests.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

This research was supported by NIH grants 1R03DE020839 and 1R03DE022538. The authors would like to thank Jim Beck and Kevin Moss in the School of Dentistry at the University of North Carolina for providing the data set on periodontal disease from the Piedmont 65+ Dental study. We also thank the editor, the associate editor and a referee for their constructive comments.

REFERENCES

- Beck JD, Koch GG, Rozier RG, Tudor GE. Prevalence and risk indicators for periodontal attachment loss in a population of older community-dwelling blacks and whites. *Journal of Periodontology*. 1990; 61:521–528. [PubMed: 2391631]
- Blazer, DG.; George, LK. ICPSR02744-v1. Vol. Inter-university Consortium for Political and Social Research [distributor]. Ann Arbor, MI: 2004. Established Populations for Epidemiologic Studies of the Elderly, 1996–1997: Piedmont Health Survey of the Elderly, Fourth In-Person Survey [Durham, Warren, Vance, Granville, and Franklin Counties, North Carolina] [Computer file].
- Datta S, Satten GA. Rank-sum tests for clustered data. *Journal of the 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, CA: Academic Press; 1999.
- Hoffman EB, Sen PK, Weinberg CR. Within-cluster resampling. *Biometrika*. 2001; 88:1121–1134.
- Huang Y, Leroux B. Informative cluster sizes for subcluster-level covariates and weighted generalized estimating equations. *Biometrics*. 2011; 67:843–851. [PubMed: 21281273]

- Rosner B, Glynn RJ, Ting Lee ML. Incorporation of clustering effects for the Wilcoxon rank sum test: a large-sample approach. *Biometrics*. 2003; 59:1089–1098. [PubMed: 14969489]
- Williamson JM, Datta S, Satten GA. Marginal analyses of clustered data when cluster size is informative. *Biometrics*. 2003; 59:36–42. [PubMed: 12762439]
- Wilcoxon F. Individual comparisons by ranking methods. *Biometrics Bulletin*. 1945; 1:80–83.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Empirical cdf plot of scores at buccal and mesial sites at baseline study

Figure 1. Empirical cdf plot of scores at buccal and mesial sites at baseline study

Plot of empirical cumulative distribution functions ($\hat{F}_3(\cdot)$) of attachment scores in buccal and mesial sites at baseline study.

Empirical cdf plot of scores at buccal and mesial sites at 18 months

Figure 2. Empirical cdf plot of scores at buccal and mesial sites at 18 months

Plot of empirical cumulative distribution functions ($\hat{F}_3(\cdot)$) of attachment scores in buccal and mesial sites at 18 months.

Table 1

Size, along with a 95% confidence interval, and power comparisons of four tests (nominal $\alpha = 0.05$) under Simulation Scenario 1. The empirical calculations are based on 3,000 replicates each.

$M = 30$ clusters				
Test	Size (CI)	Power (under effect size δ)		
		$\delta = 0.05$	$\delta = 0.10$	$\delta = 0.15$
New Test	0.060 (0.052, 0.068)	0.319	0.833	1.000
DS	0.132 (0.120, 0.144)	0.050	0.203	0.500
W	0.159, (0.146, 0.172)	0.058	0.263	0.645
CA	0.055 (0.047, 0.063)	0.296	0.814	0.985
$M = 50$ clusters				
Test	Size (CI)	Power (under effect size δ)		
		$\delta = 0.05$	$\delta = 0.10$	$\delta = 0.15$
New Test	0.053 (0.045, 0.061)	0.500	0.960	1.000
DS	0.199 (0.185, 0.213)	0.050	0.310	0.730
W	0.215 (0.200, 0.230)	0.061	0.390	0.830
CA	0.051 (0.043, 0.059)	0.460	0.950	1.000
$M = 150$ clusters				
Test	Size (CI)	Power (under effect size δ)		
		$\delta = 0.05$	$\delta = 0.10$	$\delta = 0.15$
New Test	0.055 (0.047, 0.063)	0.910	1.000	1.000
DS	0.508 (0.490, 0.526)	0.052	0.699	0.900
W	0.528 (0.510, 0.546)	0.073	0.778	0.993
CA	0.050 (0.042, 0.058)	0.896	1.000	1.000

New Test = Test developed in Section 2.1, DS= rank-sum test by Datta and Satten, W=Wilcoxon rank-sum test, CA=signed rank test with cluster averages

Table 2

Size, along with a 95% confidence interval, and power comparisons of four tests (nominal $\alpha = 0.05$) under Simulation Scenario 2. The number of clusters, M , equals 50. The empirical calculations are based on 3,000 replicates each.

Test	Size (CI)	Power (under effect size δ)		
		$\delta = 0.05$	$\delta = 0.10$	$\delta = 0.15$
New Test	0.054 (0.046, 0.062)	0.465	0.960	1.000
DS	0.146 (0.133, 0.159)	0.071	0.442	0.845
W	0.136 (0.124, 0.148)	0.073	0.501	0.916
CA	0.047 (0.039, 0.055)	0.445	0.960	1.000

New Test = Test developed in Section 2.1, DS= rank-sum test by Datta and Satten, W=Wilcoxon rank-sum test, CA=signed rank test with cluster averages

Table 3

Size, along with a 95% confidence interval, and power calculations (nominal $\alpha = 0.05$) of the new test developed in Section 2.2 under Simulation Scenario 3. Note that the CA test statistic is not computable in this situation. The number of clusters, M , equals 30. The empirical calculations are based on 3,000 replicates each.

Size	Power (under effect size δ)		
	$\delta = 0.05$	$\delta = 0.10$	$\delta = 0.15$
0.053 (0.045, 0.061)	0.275	0.743	0.964