

R package Clustered Wilcoxon Rank Sum Test and Signed Rank Test

Yujing Jiang

May 8, 2015

Abstract

The Wilcoxon rank sum test and Wilcoxon signed rank test are frequently used nonparametric tests for comparing paired data and two independent samples respectively. These two tests cannot be used for clustered data (e.g., if paired comparisons are available for each of two eyes of an individual). To incorporate clustering, a generalization of the randomization test formulation for each test is proposed by [1] and [2], where the unit of randomization is at the cluster level (e.g., person), while the individual units of analysis are at the subunit within cluster level (e.g., eye within person). The authors proposed adjusted variance estimates of these two test statistics which can be used for both balanced (same number of subunits per cluster) and unbalanced data (different number of subunits per cluster), in the presence or absence of ties. This R package is a realization of these two non-parametric tests adjusted for clustered data.

Keywords: Clustered data; Nonparametric tests; Wilcoxon rank sum test; Wilcoxon signed rank test; R package.

1 Introduction

Clustered data are often found in studies of eyes, ears, knees as well as in family studies. Regardless of the type of data, responses for individual eyes of the same subject are usually highly correlated. If the eye is the unit of analysis, then standard test procedures are not appropriate because they assume independence of sampling units, in general, in the presence of clustering, true p-values will be underestimated and confidence interval width will be too narrow.

An introduction of large sample Wilcoxon rank sum test and signed rank test will be given. We will also present the usage of the package **clusrank**.

2 Methods

2.1 Wilcoxon rank sum test for clustered data

Assume the data come in clusters where X_{ij} denotes the score for the j th subunit from the i th cluster in the first group, $i = 1, \dots, m; j = 1, \dots, g_i$ and Y_{kl} denotes the score for the l th

subunit from the k th cluster in the second group, $k = 1, \dots, n; l = 1, \dots, h_k$. The clustered Wilcoxon rank sum statistic $W_{c,obs}$ is defined as

$$W_{c,obs} = \sum_{i=1}^m \sum_{j=1}^{g_i} \text{Rank}(X_{ij}) \quad (1)$$

where ranks are determined based on the combined sample of all subunits over the X and Y clusters combined. It is assumed that the subunits for a given cluster are exchangeable. Hypothesis being tested is

$$H_0 : Pr\{U(X_{ij} - Y_{kl}) = 1\} = Pr\{U(X_{ij} - Y_{kl}) = 0\}, \text{ for any } i, j, k, l,$$

vs.

$$H_1 : Pr\{U(X_{ij} - Y_{kl}) = 1\} \neq Pr\{U(X_{ij} - Y_{kl}) = 0\}, \text{ for any } i, j, k, l,$$

where $U(a) = 1$ if $a > 0$, $U(a) = 1/2$ if $a = 0$ and $U(a) = 0$ if $a < 0$. In words, the test is based on the probability that the score from a random subunit from the X sample is greater than the score from a random subunit from the Y sample.

2.1.1 Balanced designs

Under H_0 , scores for clusters assigned to the X and Y treatments are identically distributed, so we can pool X and Y clusters together and refer to a combined set of Z clusters, where Z_{ij} = score for the j th subunit of the i th cluster, $j = 1, \dots, g, i = 1, \dots, m + n = N$. Suppose that m of the N clusters are assigned at random to the X treatment and the remaining n clusters to the Y treatment. Let $\delta_i = 1$ if $i \in I$, and $\delta_i = 0$ if $i \notin I$ denote the indicator function of the m unique values out of $\{1, \dots, N\}$ randomly assigned to the X group. The distribution of the clustered rank sum statistic $W_{c,obs}$ is

$$W_c = \sum_{i=1}^N \delta_i R_{i+} \quad \text{where } R_{i+} = \sum_{j=1}^g R_{ij} \quad (2)$$

and R_{ij} = rank of the j th subunit in the i th cluster among all gN subunits over all Z clusters. It is shown that under H_0 ,

$$E(W_c) = gm(gN + 1)/2 \quad (3)$$

and

$$\text{Var}(W_c) = [mn/\{N(N - 1)\}] \sum_{i=1}^N \{R_{i+} - g(1 + gN)/2\}^2 \quad (4)$$

So a natural large sample test statistic based on (2), (3), and (4) is

$$Z_c \{W_c - E(W_c)\} / \{\text{Var}(W_c)\}^{1/2} \quad (5)$$

Z_c is asymptotically normal if both $m \rightarrow \infty$ and $n \rightarrow \infty$

2.1.2 Unbalanced designs

For unbalanced designs, let (m_g, n_g) = number of clusters of size g assigned to the X and Y treatment. Denote $N_g = m_g + n_g$ for $g = 1, \dots, g_{max}$ and $N = \sum_{g=1}^{g_{max}} N_g$. Let $R_{ij,g}$ = rank for the j th subunit in the i th cluster of size g , $g = 1, \dots, g_{max}$; $i = 1, \dots, N_g$, $j = 1, \dots, g$, where ranks are computed based on the total study population. The construction of statistic $W_{c,obs}$ and its distribution W_c is the same as in the balanced case, i.e., $W_{c,obs}$ is the summation of ranks in the sample that is assigned to X clusters and W_c is the summation of ranks that is **randomly** assigned to X clusters. But $E(W_c)$ and $\text{Var}(W_c)$ is slightly different from those in balanced design.

$$E(W_c) = \sum_{g=1}^{g_{max}} m_g (R_{++g}/N_g) \quad (6)$$

$$\text{Var}(W_c) = \sum_{g=1}^{g_{max}} [m_g n_g / \{N_g(N_g - 1)\}] \sum_{i=1}^{N_g} (R_{i+,g} - R_{++g}/N_g)^2 \quad (7)$$

A large-sample test statistics based on W_c , (6), and (7) is

$$Z_c = \{W_c - E(W_c)\} / \{\text{Var}(W_c)\}^{1/2} \quad (8)$$

2.1.3 Stratification

Sometimes stratification needs to be considered, e.g., in multiple clinical trials, stratification by center is common, here center can be treated as a confounding variable. A modification of W_c to control confounding variable is as follows. Suppose the set of relevant confounding variables can be summarized in terms of V strata. Let $(m_{g,v}, n_{g,v})$ = number of clusters of size g in stratum v assigned to the X and Y clusters of size g in stratum v , $g = 1, \dots, g_{max}$, $v = 1, \dots, V$. Again $W_{c,obs}$ and W_c is constructed in the same way as in sections 2.1.1 and 2.1.2, whereas $E(W_c)$ and $\text{Var}(W_c)$ is adjusted accordingly as follows:

$$\begin{aligned} E(W_c) &= \sum_{g=1}^{g_{max}} \sum_{v=1}^V m_{g,v} R_{++g,v} / N_{g,v} \\ \text{Var}(W_c) &= \sum_{g=1}^{g_{max}} \sum_{v=1}^V [m_{g,v} n_{g,v} / \{N_{g,v}(N_{g,v} - 1)\}] \\ &\quad \times \sum_{i=1}^{N_{g,v}} (R_{i+,g,v} - R_{++g,v} / N_{g,v})^2 \end{aligned}$$

And the test statistic is

$$Z_c = \{W_c - E(W_c)\} / \{\text{Var}(W_c)\}^{1/2} \quad (9)$$

2.2 Wilcoxon signed rank test for clustered data

Let $X_{ij}(Y_{ij})$ denotes the baseline (follow-up) score for the j th subunit in the i th cluster (subject) and define $Z_{ij} = Y_{ij} - X_{ij}$, $j = 1, \dots, g_i$; $i = 1, \dots, m$. Also within each cluster, the

difference scores are assumed to be independent and identically distributed. Hypothesis being tested is

$$H_0 : \text{the difference score } Z \text{ is symmetric about } 0$$

vs

$$H_1 : Z \text{ is symmetric about } \gamma, \gamma \neq 0.$$

Rank $|Z_{ij}|$ over the total of $G = \sum_{i=1}^m g_i$ subunits from the m clusters and let $S_{ij} = R_{ij}V_{ij}$, where R_{ij} = rank of $|Z_{ij}|$ within the total data set of G subunits over m clusters, and $V_{ij} = \text{sign}(Z_{ij})$.

2.2.1 Balanced Design

In a balanced design (i.e., each cluster has the same number of subunits, g), the clustered Wilcoxon signed rank statistic is defined by

$$T_c^{(obs)} = \sum_{i=1}^m S_{i+} \equiv \sum_{i=1}^m \sum_{j=1}^g R_{ij}V_{ij}, \quad (10)$$

where $S_{i+} = \sum_{j=1}^g S_{ij}$ and only consider nonzero Z_{ij} in the computation of signed ranks. When considering the randomization distribution corresponding to T_c , the unit of randomization is the cluster while the unit of analysis is the subunit. Let $\delta_1, \dots, \delta_m$ be i.i.d. random variables each taking on the values +1 and -1 with probability 1/2, then

$$T_c = \sum_{i=1}^m \delta_i S_{i+}. \quad (11)$$

It is shown that under H_0 , $E(T_c) = 0$ and $\text{Var}(T_c) = \sum_{i=1}^m S_{i+}^2$. And the large sample test statistic is

$$Z_c = T_c / \left(\sum_{i=1}^m S_{i+}^2 \right)^{1/2} \sim N(0, 1) \quad \text{under } H_0. \quad (12)$$

2.2.2 Unbalanced Designs

In the case of an unbalanced design, the signed rank statistic is

$$T_c^{(obs)} = \sum_{i=1}^m w_i \bar{S}_i \quad (13)$$

where $\bar{S}_i = S_{i+}/g_i$, $w_i = 1/\text{Var}(\bar{S}_i)$ under H_0 . The randomization distribution corresponding to $T_{c,s}^{obs} = \sum_{i=1}^m \delta_i w_i \bar{S}_i$ is defined as in (11). And the test statistic is defined as

$$Z_{c,s} = T_{c,s} / \left(\sum_{i=1}^m \hat{w}_i^2 \bar{S}_i^2 \right)^{1/2} \sim N(0, 1) \text{ under } H_0, \quad (14)$$

where $\hat{w}_i = g_i / [\widehat{\text{Var}}(S_{ij})\{1 + (g_i - 1)\hat{\rho}_{s,cor}\}]$, $\hat{\rho}_{s,cor} = \hat{\rho}_s \left(1 + \frac{1 - \hat{\rho}_s^2}{m - 5/2}\right)$, $\hat{\rho}_s = \max[\hat{\sigma}_A^2 / (\hat{\sigma}_A^2 + \hat{\sigma}^2), 0]$, $\hat{\sigma}^2 = \sum_{i=1}^m \sum_{j=1}^{g_i} (S_{ij} - \bar{S}_i)^2 / (G - m)$, $\hat{\sigma}_A^2 = \max[\{\sum_{i=1}^m g_i (\bar{S}_i - \bar{\bar{S}})^2 / (m - 1) - \hat{\sigma}^2\} / g_0, 0]$, $g_0 = [\sum_{i=1}^m g_i - \sum_{i=1}^m g_i^2 / \sum_{i=1}^m g_i] / (m - 1)$ and $\widehat{\text{Var}}(S_{ij}) = \sum_{i=1}^m \sum_{j=1}^{g_i} (S_{ij} - \bar{\bar{S}})^2 / (G - 1)$.

3 R package clusrank

The R package is a realization of all the tested we presented in the previous section. It contains two functions, The package *clsrank* contains two functions, `cluswilcox` and `clusignrank`. The `cluswilcox` is a function for clustered Wilcoxon ranksum test and `clusignrank` is a function for clustered Wilcoxon signed rank test. Also there are four test datasets in this package, data *crd* is clustered non-stratified data set comparing effects of two treatments, data *crd.str* is clustered stratified data set comparing effects of two treatments; data *crsd* is clustered balanced data set comparing the difference between pre and post treatment effect, data *crsd.unb* is clustered unbalanced data set comparing the difference between pre and post treatment effect.

3.1 cluswilcox

The *cluswilcox* function takes in a formula $score \sim group$ to indicate which variable in the dataframe is the observed score of the subunit and which is the label of treatment. Also a vector of *id* is required to label the cluster each observation belongs to. *id* and *group* have to be numerical. If the data is stratified, then a vector of *stratum* is also needed to label which stratum each observation comes from. There are also some currently unused options which is may be extended later. A usage example is as follows:

```
library(clsrank)
data(crd)
cluswilcox(z ~ group, data = crd, id = id)
data(crd.str)
cluswilcox(z ~ group, data = crd.str, id = id, stratum = stratum)
```

3.2 clusignrank

The *clusignrank* function is used to carry out clustered Wilcoxon signed rank test for paired comparisons. It takes in three arguments: *z* is observed score, *id* is the cluster id of each observation. A usage example is as follows:

```
data(crsd)
clusignrank(z, id, data = crsd)
data(crsd.unb)
clusignrank(z, id, data = crsd.unb)
```

4 Summary

In this article, clustered Wilcoxon rank sum test and Wilcoxon signed rank are briefly introduced, for both balanced and unbalanced designs, also, an extra situation considering stratified data is mentioned for Wilcoxon rank sum test. Also a R **clsrank** package with

functions for each test is built. This R package is a counterpart of the existing SAS macros for these two tests.

References

- [1] B. Rosner, R. J. Glynn, and M.-L. Ting Lee, “Incorporation of clustering effects for the wilcoxon rank sum test: A large-sample approach,” *Biometrics*, vol. 59, no. 4, pp. 1089–1098, 2003.
- [2] B. Rosner, R. J. Glynn, and M.-L. T. Lee, “The wilcoxon signed rank test for paired comparisons of clustered data,” *Biometrics*, vol. 62, no. 1, pp. 185–192, 2006.