science, engineering and health studies

Research Article

Science, Engineering and Health Studies
https://lio1.tci-thaijo.org/index.php/sehs
ISSN (Online): 2630-0087

Adjusted rank tests for clustered data in balanced design

Prayad Sangngam and Wipawan Laoarun[*]

Department of Statistics, Faculty of Science, Silpakorn University, Nakhon Pathom, 73000, Thailand

ABSTRACT

The Wilcoxon test is commonly used to test whether two independent samples are drawn from the same population distributions. In many practical situations, the data in each sample are clustered. The clustered rank sum test was developed for testing the differences of location parameters from two samples with clustered data. However, the critical value of the clustered rank sum test for a data set depends on the sums of observation ranks within clusters. In a balanced design, the data sets with same numbers of clusters in two samples may use different critical values. This study proposed adjusted rank test (T) that makes adjustments to sums of observation ranks in two independent samples. This test used the same critical values for data sets with equal numbers of clusters. Two tests for the equivalence distributions of three or more populations were also considered using samples with clustered data and referred to as modified rank test (T_1) and adjusted rank test (T_2). The simulation study showed that the adjusted rank tests can maintain the sizes of the tests for all situations. For a small number of clusters and correlation coefficients between observations in a cluster, the T was the best choice. The empirical power of the T_2 was higher than that of the Kruskal-Wallis test, based on a mean cluster. The powers of the adjusted tests increased when the number of clusters, number of observations per cluster, and effect size increased. However, the powers of these tests decreased when the size of the correlation coefficients between observations in a cluster increased.

Keywords: clustered data; independent samples; ranks sum test; power of the test

[*]***Corresponding author***:
Wipawan Laoarun
laoarun_w@silpakorn.edu

Received: *17 March 2020*
Revised: *21 August 2020*
Accepted: *3 September 2020*
Published: *16 February 2021*

Citation:
Sangngam, P., and Laoarun, W. (2021). Adjusted rank tests for clustered data in balanced design. Science, Engineering and Health Studies, 15, 21020002.

1. INTRODUCTION

In certain situations, researchers are interested in testing the null hypothesis that two or more independent samples are drawn from the same population. In parametric statistics, the independent samples t-test is widely used to test the differences between two independent populations. The one-way analysis of variance F-test is also used to test the null hypothesis that three or more samples are drawn from populations with equal means. These tests require random samples to be drawn from normal distributed populations. If the assumption of these tests cannot be met, the Wilcoxon test (Wilcoxon, 1945), which is a nonparametric procedure, can be used for two independent samples. Mood (1954) showed that for very large samples, the power efficiency of the Wilcoxon test relative to t-tests approaches 95.5%. The Kruskal-Wallis test is a useful nonparametric tool for comparing three or more independent samples (Kruskal and Wallis, 1952). Hodges and Lehmann (1956) showed that under certain conditions, the relative efficiency of the Kruskal-Wallis test relative to the usual parametric F-test may be greater than 1. A common assumption of the t-test, F-test, Wilcoxon test, and Kruskal-Wallis test is that all observations are independent.

In numerous studies, a sample may have clusters of correlated observations. Examples of clustered data include the repeated measure of blood pressure of a single unit, the socioeconomic characteristics of households in a block, and the mass index of siblings. The test statistic used to analyze correlated data as independent data is known to have an inflated probability of making a Type I error. In parametric approaches, different procedures are used to test hypotheses with correlated clustered data. Most theoretical research for clustered data assumes a parametric model. The adjusted F-test statistic was developed by using intracorrelation so that the adjusted statistic has approximately the F distribution with the same degrees of freedom as those of the F-test statistic (Wu et al., 1988). A two-stage general least squares test was proposed by transforming observations into uncorrelated ones (Rao et al., 1993). The two aforementioned tests depend on the unknown intracorrelation. Lahiri and Yan (2009) proposed an alternative test that does not require the estimation of intracorrelation. As for nonparametric approaches, the relevant literature that incorporates clustered data is limited. Rosner and Grove (1999) considered the combination of clustered data in the Mann-Whitney U test. The estimates of correlation parameters have been used to correct the estimated variance of the test statistic. The simulation results showed that the test has an appropriate Type I error rate in a balanced design with as few as 20 clusters per sample. However, the study did not consider large sample theory. A large sample randomization test for clustered data were introduced by applying the approach of the Wilcoxon test. This refers to as the Rosner, Glynn, and Lee (RGL) test (Rosner et al., 2003). The signed rank test was also developed to compare the parameters under clustered data settings (Rosner et al., 2006). However, under different data sets with equal numbers of clusters in samples for a balanced design, the critical values of the RGL test may be different. In practice, the RGL test requires researchers to find a critical value for one data set.

To use the same critical value for different data sets, the RGL test for two independent samples in a balanced design was adjusted. Under such balanced design, two tests for testing the differences among three or more independent samples with clustered data were also considered. Through a simulation study, the performance of the two tests was studied in terms of their power and their capability of controlling the probability of Type I errors. The efficiency of the proposed tests was compared with that of alternative tests.

2. MATERIALS AND METHODS

2.1 RGL test for two samples with clustered data

Let X_{ijk} be the k-th observation in the j-th cluster of the i-th sample for i = 1,2, j = 1,2,...,n_i, and k = 1,2,...,m_{ij}, where n_i is the number of clusters in the i-th sample and m_{ij} is the cluster size of the j-th cluster in the i-th sample. The indicator δ_{ijk} denotes the group of the samples; $\delta_{ijk} = 1$ if X_{ijk} belongs to the first sample, and $\delta_{ijk} = 0$ if X_{ijk} belongs to the second sample. The data are presented in the form of $(\mathrm{X}, \delta) = \{(X_{ijk}, \delta_{ijk}): k = 1,2,\dots,m_{ij}, j = 1,2,\dots,n_i, i = 1,2\}$

We assumed that clusters are independent and that the observations within clusters are not. The hypothesis to be tested herein is that no difference exists between the location parameters of two populations. In this work, we considered a case of balanced design, i.e., the same number of clustered sizes $(m_{ij} = m)$ for all i = 1,2 and j = 1,2,...,n_i.

Rosner, Glynn, and Lee (2003) proposed the RGL Wilcoxon rank sum test for clustered data. Let R_{ijk} be the rank of X_{ijk} based on the combined samples of all observations. The sum of ranks from the first sample is assigned as the test statistic. Let $\delta_{ij} = \delta_{ijk}$ for all j = 1,2,...,m and j = 1,2,...,n_i. The RGL test statistic can be defined as:

$$W_c = \sum_{i=1}^{2} \sum_{j=1}^{n_1+n_2} \delta_{ij} R_{ij+},$$

where $R_{ij+} = \sum_{k=1}^{m} R_{ijk}$ is the sum of observation ranks in the j-th cluster of the i-th sample.

The RGL method assumes that the observations in a given cluster are exchangeable. The exact distribution of W_c is considered on the basis of random permutation conditioning on the sum of observation ranks in the j-th cluster of the i-th sample, R_{ij+}. In deriving the distribution of the test under a null hypothesis, if $n_1 + n_2$ is small, then the distribution of W_c conditioning on R_{ij+} can be generated by combining all possible permutations of R_{ij+} between two samples. The total number of permutations is $\binom{n_1+n_2}{n_1}$. If $n_1 + n_2$ is large, then the computation is intensive. The RGL asymptotic test statistic is:

$$z = \frac{W_c - mn_1/(mN-1)}{\sqrt{Var(W_c)}} \sim N(0,1),$$

where $Var(W_c) = \frac{n_1 n_2}{N(N-1)} \sum_{i=1}^{2} \sum_{j=1}^{n_i} \left[R_{ij+} - \frac{m(1+nM)}{2}\right]^2$ and $N = n_1 + n_2$.

Under mild conditions, the test statistic Z has an asymptotic standard normal distribution. For unequal numbers of clusters between two samples $(n_1 \neq n_2)$, Rosner et al. (2003) showed that the test statistic may result in low efficiency.

Note that in the case of different data sets with equal numbers of clusters in two samples, the set of R_{ij+} in each data set may be different, although $\sum_{i=1}^{2} \sum_{j=1}^{n_i} R_{ij+}$ of these data sets are equal. Therefore, if $n_1 + n_2$ is small, then the critical values of statistic W_c for these data sets are different. To use the same critical values in Section 2.2, one adjusts the Wilcoxon test to the sums of observation ranks from clusters as the raw data.

2.2 Adjusted rank test for two samples with clustered data

Let X_{ijk} and R_{ijk} be the same as those defined in Section 2.1. The balanced design is also considered in this section. After assigning the ranks to the observations, let $R_{ij+} = \sum_{k=1}^{m} R_{ijk}$ be the sum of observation ranks in the j-th cluster of the i-th sample. To obtain the same critical values of the test statistic for different data sets with the same numbers of clusters, we proposed an adjusted test statistic.

To compute the observed value of the adjusted test statistic, we combined the two samples and assigned new ranks to all R_{ij+} in ascending order. The mean of new ranks is given to the sums of observations with ties. Let Z_{ij}

be the new rank of R_{ij+}. Let $\delta_i = 1$ if R_{ij+} belongs to the first sample and $\delta_i = 0$ if R_{ij+} belongs to the second sample. The adjusted rank test-T can be defined as

$$T = S_R - \frac{n_1(n_1+1)}{2},$$

where $S_R = \sum_{i=1}^{2}\sum_{j=1}^{n_i+n_2} \delta_i Z_{ij}$ is the sum of the new ranks from the first sample. When the hypothesis about the difference between two location parameters is tested, the null hypothesis is rejected for either a sufficiently small or a sufficiently large value of T. Therefore, we rejected the null hypothesis at the significance level of α if the computed value of T is less than or equal to the critical value of $w_{\alpha/2}$ or greater than or equal to the critical value of $w_{1-\alpha/2}$. When the null hypothesis is true, Theorem 1 shows that the sampling distribution for each of $\binom{n_1+n_2}{n_1}$ permutation of the observation $(z_{11}, \dots, z_{1n_1})$ is equally likely. It should be noted that, under the null hypothesis, the random vectors $(\mathrm{X}_{11}, \dots, \mathrm{X}_{1n_1})$ and $(\mathrm{X}_{21}, \dots, \mathrm{X}_{2n_2})$ have the same distribution, with $\mathrm{X}_{ij} = (X_{ijl}, \dots, X_{ijm})$ consisting of m exchangeable random variables.

Theorem 1: Let $\mathrm{X}_{11}, \dots, \mathrm{X}_{1n_1}; \mathrm{X}_{21}, \dots, \mathrm{X}_{2n_2}$ be independently distributed according to a common continuous distribution. Let $\mathrm{Z}_{11}, \dots, \mathrm{Z}_{1n_1}$ denote the ranks of $\mathrm{R}_{11}, \dots, \mathrm{R}_{1n_1+}$ in the combined ranking of all $n_1 + n_2$ of R_{ij+}. Then, the probability of observing $(\mathrm{Z}_{11}, \dots, \mathrm{Z}_{1n_1})$ is equal to

$$\mathrm{P}(\mathrm{Z}_{11} = \mathrm{z}_{11}, \dots, \mathrm{Z}_{1n_1} = \mathrm{z}_{1n_1}) = \frac{1}{\binom{n_1+n_2}{n_1}}$$

Proof: Rosner et al. (2003) considered the set of R_{ij+} belonging to the first sample as a random sample of size n_1 from the population $\{R_{ij+}: i = 1,2; j = 1,2, \dots, n_i\}$. Thus, the probability of obtaining $R_{11+} = r_{11}, \dots, R_{1n_1+} = r_{1n_1}$ is equally likely. The new rank Z_{ij} of R_{ij+} is the rank transformation in which the number of distinct values and the number of replicated values of Z_{ij} are equivalent to that of R_{ij+}. Therefore, an observation $(r_{11}, \dots, r_{1n_1})$ leads to the observation $(z_{11}, \dots, z_{1n_1})$. The probability of obtaining $(z_{11}, \dots, z_{1n_1})$ is also equal to $\frac{1}{\binom{n_1+n_2}{n_1}}$.

The adjusted rank test statistic (T) is a mapping from the original space of W_c to a new space of T. As W_c is a discrete random variable, the probability mass function for the T is

$$\mathrm{P}(T = t) = \sum_{w_c \in f^{-1}(t)} P(W_c = w_c),$$

where f^{-1} is an inverse mapping from subsets of the spaces of T to subsets of the space of W_c.

The procedure for deriving the T is equivalent to that for the Wilcoxon test statistic and that their distributions are identical. The critical values of the proposed test statistic can be easily found in the table of critical values of the Wilcoxon test statistic. For the T, the critical value is the same for all data sets when many data sets have equal numbers of clusters.

2.3 Rank tests for three or more samples with clustered data

In this section, we proposed the procedure for the case in which observations are collected from three or more independent samples with clustered data. It is assumed that the observations consist of $p \geq 3$ samples. We are interested in testing the null hypothesis that several samples are drawn from the same population. Let X_{ijk} be the k-th observation in cluster j of the i-th sample for $i = 1,2, \dots, p$, $j = 1,2, \dots, n_i,$, and $k = 1,2, \dots, m_{ij}$, where n_i is the number of clusters in the i-th sample and m_{ij} is the cluster size of the j-th cluster in the i-th sample. The case of balanced design was also considered, i.e., the same number of clustered sizes $(m_{ij} = m,)$ for all $i = 1,2, \dots, p$ and $j = 1,2, \dots, n_i$.

To compute the proposed test statistics, we replaced each observation X_{ijk} with its rank R_{ijk} relative to all observations in p samples. We assigned rank 1 to the smallest observation, rank 2 to the next higher observation, and so on. In the case of ties, the average of the ranks was assigned. After assigning the ranks to the observations, let $R_{ij+} = \sum_{k=1}^{m} R_{ijk}$ be the sum of observation ranks in the j-th cluster of the i-th sample. Corresponding to sample i, $i = 1,2, \dots, p$ the sum of ranks in a sample is computed by $W_i = \sum_{j=1}^{n_i} R_{ij+}$. The expectation of the sum of ranks in a sample is equal to $E(W_i) = \frac{n_i}{N}\sum_{i=1}^{p}\sum_{j=1}^{n_i} R_{ij+} = \frac{n_i m(mN+1)}{2}$, where $N = \sum_{i=1}^{p} n_i$. The modified test- T_1 is given by

$$T_1 = \frac{12}{N(N+1)}\sum_{i=1}^{p}\frac{1}{n_i}\left[W_i - \frac{n_i m(mN+1)}{2}\right]^2.$$

We conducted an α-level test of the null hypothesis that the p samples are drawn from the same population. The statistical value of T_1 can be compared with the $(1-\alpha)100\%$-th percentile of T_1 under H_0, so that the null hypothesis is rejected if the observed value of T_1 is greater than or equal to this percentile. The exact distribution of T_1 is considered on the basis of the permutation procedure conditioning on the sum of observation ranks in the j-th cluster of the i-th sample, R_{ij+}. To derive the distribution of T_1, we assumed that the observations in a cluster are exchangeable. Under the null hypothesis, if $N = \sum_{i=1}^{p} n_i$ is small, then the distribution of T_1 conditioning on R_{ij+} can be generated by combining all possible permutations of R_{ij+}between p samples. An exhaustive permutation number of R_{ij+} is $\frac{N!}{\prod_{i=1}^{p} n_i!}$. Many data sets have the same numbers of clusters for all p samples, the percentiles of T_1 may differ between data sets. To obtain the same percentile of a test statistic, we proposed an adjusted test- T_2.

From the sum of the observation ranks of the j-th cluster in the i-th sample R_{ij+}, let Z_{ij} be the rank of R_{ij+} among $R_{11+}, \dots, R_{1n_1+}, . R_{21+}, \dots, R_{22+}, \dots, . R_{p1+}, \dots, R_{\mathrm{p}n_p+}$. That is, let Z_{ij} be the rank of the pooled samples of $N = \sum_{i=1}^{p} n_i$. Let $Z_i = \sum_{j=1}^{n_i} Z_{ij}$ be the rank of the sums associated with the i-th sample for $i = 1,2, \dots, p$. The T_2 is given by the following statistic:

$$T_2 = \frac{12}{N(N+1)}\sum_{i=1}^{p}\frac{Z_i^2}{n_i} - 3(N+1).$$

The critical value corresponding to the α-level of statistic T_2 is denoted by t_α and is the upper $(1-\alpha)100\%$-th percentile under the null hypothesis. Therefore, the null hypothesis is rejected when the computed value of T_2 is greater than or equal to t_α. Similar to the idea of Theorem 1, the probability for each of $\frac{N!}{\prod_{i=1}^{p} n_i!}$ permutations of Z_{ij} is equally likely.

The procedure for constructing the T_2 statistic is equivalent to the Kruskal-Wallis test statistic. Thus, the critical value of the T_2 is obtained by using the critical values for the Kruskal-Wallis test statistic. In addition, the same critical values are adopted in using the T_2 for the different data sets with the same numbers of clusters in p samples.

2.4 Simulation study

A simulation study was conducted to analyze the properties of the T and T_2, i.e., robustness and power of the test. In this study, the robustness evaluation was based on Bradley's criterion (0.0250, 0.0750) for the significance level of 0.05 (Bradley, 1978). We considered the situation in which all the observations of a cluster belong to only one sample. The study was constructed under two and three samples. Let n_1, n_2, and n_3 denote the number of clusters from the first, second, and third samples, respectively. When the sample sizes in each sample are equal, the power of the test is usually high. Thus, we set $n_1 = n_2$, in the case of two samples and $n_1 = n_2 = n_3$ in the case of three samples.

We generated data $X_{ijk} = \exp(Y_{ijk}) + (i-1)d$, where $\boldsymbol{Y}_{ij} = (Y_{ij1}, Y_{ij2}, .., Y_{ijm})$ is independent multivariate normal with mean vector $\mathbf{0}$ and exchangeable covariance matrix $\sum = (1-\rho)\mathbf{I} + \rho\mathbf{1}$, where $\mathbf{I}$ is the identity matrix of size m x m and $\mathbf{1}$ is the m x m matrix of all elements equal to 1. This procedure creates log-normal distribution data with usually skewed distributions, for which rank procedures are often used.

In the case of two samples, the numbers of clusters in samples (n_1,n_2) are equal to (5,5), (7,7), and (10,10) with cluster sizes m of 2, 4, and 6, respectively. In addition, we computed the ordinary Wilcoxon rank sum test based on the cluster means of observations. This test is called the cluster mean test and denoted it as T_M.

In the case of three samples, the numbers of clusters in samples (n_1,n_2,n_3) are equal to (4,4,4), (6,6,6), and (8,8,8) with cluster sizes m of 2, 4, and 6, respectively. In this case, the Kruskal-Wallis test based on mean cluster observations was also computed and denoted as T_{KW}.

For each case involving two and three samples, the coefficient of correlation between observations in a cluster (ρ) was set to be 0.1, 0.3, 0.5, 0.7, and 0.9. The effect size (d) is equal to 0.0, 0.3, and 0.5. The significance level of 0.05 was used for all tests. For each situation, the rejection rate was obtained from 10,000 replicates.

3. RESULTS AND DISCUSSION

3.1 Simulation study results

Tables 1 and 2 respectively present the results of the comparison of the estimated probabilities of Type I errors and power among T, W_c, and T_M for testing the differences in the location parameters of two populations at two different effect sizes at the nominal significance level of 0.05.

When the numbers of clusters in each sample are equal to (5, 5), (7, 7), and (10, 10), the estimated probabilities of Type I errors for all three tests are close to the significance level of 0.05 and lie in Bradley's criterion for all situations (Table 1). Hence, the T, W_c, and T_M can control the probabilities of Type I errors for all situations in this study.

Table 1. Estimated probabilities of Type I errors of adjusted rank test (T), RGL test (W_c), and cluster mean test (T_M) at nominal alpha of 0.05

m	ρ	(n_1,n_2)=(5,5)			(n_1,n_2)=(7,7)			(n_1,n_2)=(10,10)		
		T	W_c	T_M	T	W_c	T_M	T	W_c	T_M
2	0.1	0.0479	0.0411	0.0559	0.0466	0.0459	0.0532	0.0481	0.0453	0.0524
	0.3	0.0496	0.0442	0.0568	0.0461	0.0455	0.0512	0.0463	0.0458	0.0515
	0.5	0.0473	0.0437	0.0542	0.0497	0.0479	0.0531	0.0496	0.0478	0.0532
	0.7	0.0527	0.0486	0.0574	0.0519	0.0498	0.0575	0.0521	0.0488	0.0515
	0.9	0.0532	0.0479	0.0576	0.0467	0.0463	0.0495	0.0521	0.0503	0.0546
4	0.1	0.0530	0.0431	0.0556	0.0486	0.0435	0.0525	0.0466	0.0426	0.0501
	0.3	0.0554	0.0460	0.0568	0.0497	0.0465	0.0532	0.0495	0.0471	0.0510
	0.5	0.0525	0.0439	0.0549	0.0519	0.0478	0.0529	0.0499	0.0470	0.0524
	0.7	0.0525	0.0447	0.0535	0.0521	0.0474	0.0550	0.0500	0.0469	0.0509
	0.9	0.0537	0.0468	0.0555	0.0514	0.0472	0.0525	0.0518	0.0471	0.0521
6	0.1	0.0521	0.0410	0.0554	0.0543	0.0477	0.0540	0.0511	0.0477	0.0512
	0.3	0.0552	0.0435	0.0572	0.0560	0.0502	0.0572	0.0488	0.0444	0.0512
	0.5	0.0586	0.0482	0.0587	0.0556	0.0497	0.0565	0.0548	0.0496	0.0531
	0.7	0.0553	0.0454	0.0558	0.0553	0.0495	0.0555	0.0508	0.0459	0.0530
	0.9	0.0543	0.0453	0.0547	0.0492	0.0449	0.0512	0.0527	0.0482	0.0521

Table 2. Empirical powers of adjusted rank test (T), RGL test (W_c), and cluster mean test (T_M) at effect size (d) of 0.3, 0.5 and nominal alpha of 0.05

d	m	ρ	(n_1,n_2)=(5,5)			(n_1,n_2)=(7,7)			(n_1,n_2)=(10,10)		
			T	W_c	T_M	T	W_c	T_M	T	W_c	T_M
0.3	2	0.1	0.1046	0.0950	0.0883	0.1280	0.1258	0.0997	0.1676	0.1696	0.1210
		0.3	0.0934	0.0893	0.0874	0.1205	0.1217	0.1022	0.1467	0.1525	0.1203
		0.5	0.0894	0.0842	0.0822	0.1114	0.1076	0.1002	0.1373	0.1397	0.1193
		0.7	0.0880	0.0820	0.0888	0.1041	0.1029	0.1018	0.1285	0.1297	0.1196
		0.9	0.0896	0.0810	0.0933	0.0973	0.0929	0.0987	0.1226	0.1222	0.1247
	4	0.1	0.1455	0.1263	0.0911	0.1946	0.1894	0.1091	0.2618	0.2647	0.1338
		0.3	0.1184	0.1060	0.0900	0.1442	0.1418	0.0990	0.1921	0.1935	0.1239
		0.5	0.1015	0.0895	0.0854	0.1244	0.1163	0.0960	0.1665	0.1675	0.1253
		0.7	0.0909	0.0797	0.0815	0.1106	0.1059	0.0993	0.1363	0.1341	0.1175
		0.9	0.0845	0.0727	0.0821	0.1061	0.0996	0.1037	0.1261	0.1212	0.1210
	6	0.1	0.1785	0.1515	0.0965	0.2389	0.2308	0.1171	0.3217	0.3210	0.1426
		0.3	0.1267	0.1066	0.0857	0.1617	0.1556	0.1064	0.2155	0.2161	0.1256
		0.5	0.1122	0.0974	0.0927	0.1361	0.1311	0.1014	0.1709	0.1745	0.1210
		0.7	0.0960	0.0856	0.0879	0.1137	0.1056	0.0974	0.1413	0.1410	0.1123
		0.9	0.0875	0.0768	0.0842	0.1025	0.0943	0.0977	0.1267	0.1207	0.1188
0.5	2	0.1	0.1878	0.1772	0.1413	0.2584	0.2565	0.1810	0.3495	0.3594	0.2371
		0.3	0.1647	0.1582	0.1374	0.2239	0.2285	0.1781	0.3123	0.3223	0.2395
		0.5	0.1502	0.1452	0.1317	0.1980	0.2008	0.1738	0.2712	0.2781	0.2239
		0.7	0.1416	0.1366	0.1392	0.1886	0.1869	0.1758	0.2511	0.2579	0.2298
		0.9	0.1363	0.1297	0.1450	0.1726	0.1677	0.1751	0.2350	0.2365	0.2327
	4	0.1	0.2881	0.2616	0.1559	0.3962	0.3939	0.1998	0.5420	0.5550	0.2696
		0.3	0.2162	0.1978	0.1417	0.2892	0.2871	0.1818	0.4034	0.4133	0.2464
		0.5	0.1835	0.1648	0.1355	0.2347	0.2310	0.1732	0.3308	0.3404	0.2390
		0.7	0.1537	0.1414	0.1309	0.2069	0.2008	0.1751	0.2743	0.2762	0.2282
		0.9	0.1366	0.1234	0.1332	0.1811	0.1751	0.1729	0.2400	0.2367	0.2274
	6	0.1	0.3588	0.3276	0.1692	0.4963	0.4923	0.2249	0.6453	0.6582	0.2975
		0.3	0.2393	0.2183	0.1420	0.3359	0.3272	0.1915	0.4504	0.4641	0.2520
		0.5	0.1988	0.1800	0.1447	0.2578	0.2518	0.1782	0.3536	0.3600	0.2402
		0.7	0.1631	0.1476	0.1347	0.2075	0.2014	0.1692	0.2844	0.2846	0.2273
		0.9	0.1398	0.1254	0.1326	0.1796	0.1702	0.1698	0.2452	0.2411	0.2283

As shown in Table 2, when (n_1, n_2) = (5,5) and (7,7), with the correlation coefficient being less than 0.7, the T achieves the highest empirical power, followed by the W_c. However, for (n_1, n_2) = (10,10), the W_c has the highest empirical power, which is slightly higher than that of the T.

For a given number of clusters (n_1, n_2), the number of observations in each cluster (m), the effect size (d), and the estimated powers of the T and W_c decrease as the correlation coefficients (ρ) of the observations increase. However, for a fixed number of clusters, number of observations in a cluster, and correlation coefficient, the powers of all three tests increase when the effect size increases. For a fixed number of observations, coefficient of correlation, and effect size, the estimated powers of all three tests increase as the number of cluster increases. When the number of clusters, coefficient of correlation, and effect size are given, the estimated power of the T increases as the number of observations increases.

Tables 3 and 4 respectively present the results of the comparison of the estimated probabilities of Type I errors and powers of the T_2 and Kruskal-Wallis test based on mean cluster (T_{KW}) for comparing the location parameters of three or more populations at two different effect sizes at the nominal significance level of 0.05.

Table 3 shows that when the numbers of clusters in each sample are equal to (4,4,4), (6,6,6), and (8,8,8), the estimated probabilities of Type I errors of both tests are close to the significance level of 0.05 and lie in Bradley's criterion for all situations. Hence, the T_2 and T_{KW} can control the probabilities of Type I errors for all situations in this study.

Table 4 shows that, for a given number of clusters (n_1, n_2, n_3), number of observations in each m, and d, the empirical powers of the T_2 and T_{KW} tend to decrease to their corresponding limiting values as ρ increase. The power of each test increases with an increase in the number of observations, effect size, and number of clusters. For all situations, the T_2 is significantly more powerful than T_{KW}, and their powers tend to one for a large number of clusters with a large number of observations and low correlation.

Table 3. Estimated probabilities of Type I errors of adjusted rank test (T_2) and Kruskal-Wallis test based on the mean cluster (T_{KW}) at nominal alpha of 0.05

m	ρ	$(n_1,n_2,n_3)=(4,4,4)$		$(n_1,n_2,n_3)=(6,6,6)$		$(n_1,n_2,n_3)=(8,8,8)$	
		T_2	T_{KW}	T_2	T_{KW}	T_2	T_{KW}
2	0.1	0.0459	0.0480	0.0482	0.0445	0.0505	0.0491
	0.3	0.0499	0.0495	0.0488	0.0468	0.0533	0.0527
	0.5	0.0480	0.0480	0.0517	0.0494	0.0485	0.0485
	0.7	0.0508	0.0489	0.0489	0.0455	0.0536	0.0542
	0.9	0.0471	0.0482	0.0569	0.0552	0.0497	0.0497
4	0.1	0.0486	0.0485	0.0502	0.0505	0.0523	0.0495
	0.3	0.0509	0.0464	0.0479	0.0474	0.0506	0.0508
	0.5	0.0472	0.0483	0.0481	0.0487	0.0476	0.0497
	0.7	0.0496	0.0497	0.0546	0.0520	0.0514	0.0522
	0.9	0.0446	0.0458	0.0515	0.0514	0.0485	0.0491
6	0.1	0.0455	0.0469	0.0493	0.0495	0.0536	0.0520
	0.3	0.0462	0.0486	0.0441	0.0452	0.0541	0.0559
	0.5	0.0460	0.0466	0.0463	0.0467	0.0472	0.0468
	0.7	0.0495	0.0496	0.0506	0.0515	0.0529	0.0514
	0.9	0.0501	0.0517	0.0468	0.0466	0.0494	0.0502

Table 4. Empirical powers of adjusted rank test (T_2) and Kruskal-Wallis test based on mean cluster (T_{KW}) at effect size (d) of 0.3, 0.5 and nominal alpha = 0.05

d	m	ρ	$(n_1,n_2,n_3)=(4,4,4)$		$(n_1,n_2,n_3)=(6,6,6)$		$(n_1,n_2,n_3)=(8,8,8)$	
			T_2	T_{KW}	T_2	T_{KW}	T_2	T_{KW}
0.3	2	0.1	0.1409	0.1051	0.2224	0.1433	0.3001	0.1938
		0.3	0.1292	0.1021	0.1975	0.1439	0.2689	0.1968
		0.5	0.1235	0.1056	0.1836	0.1485	0.2331	0.1851
		0.7	0.1175	0.1097	0.1602	0.1395	0.2099	0.1882
		0.9	0.1078	0.1088	0.1552	0.1455	0.1987	0.1872
	4	0.1	0.2170	0.1206	0.3543	0.1702	0.4738	0.2233
		0.3	0.1630	0.1107	0.2478	0.1477	0.3475	0.1961
		0.5	0.1335	0.1019	0.1984	0.1390	0.2786	0.1866
		0.7	0.1198	0.1064	0.1793	0.1512	0.2311	0.1874
		0.9	0.1099	0.1061	0.1546	0.1454	0.1968	0.1849
	6	0.1	0.2756	0.1253	0.4462	0.1766	0.6000	0.2500
		0.3	0.1846	0.1130	0.2869	0.1553	0.3960	0.2080
		0.5	0.1421	0.1039	0.2124	0.1439	0.2963	0.1902
		0.7	0.1234	0.1065	0.1850	0.1512	0.2401	0.1872
		0.9	0.1151	0.1110	0.1585	0.1493	0.2058	0.1902
0.5	2	0.1	0.2899	0.1994	0.4674	0.3018	0.6039	0.4170
		0.3	0.2526	0.1935	0.4096	0.2998	0.5425	0.4057
		0.5	0.2343	0.1953	0.3692	0.2967	0.4836	0.3898
		0.7	0.2149	0.1971	0.3227	0.2802	0.4372	0.3849
		0.9	0.2041	0.1992	0.3071	0.2873	0.4080	0.3912
	4	0.1	0.4617	0.2342	0.6985	0.3667	0.8466	0.4962
		0.3	0.3319	0.2092	0.5360	0.3182	0.6859	0.4328
		0.5	0.2584	0.1915	0.4280	0.2885	0.5636	0.4041
		0.7	0.2245	0.1923	0.3554	0.2864	0.4777	0.3940
		0.9	0.2047	0.1943	0.3071	0.2847	0.4155	0.3902
	6	0.1	0.5696	0.2570	0.8121	0.4100	0.9285	0.5509
		0.3	0.3773	0.2184	0.5947	0.3310	0.7484	0.4643
		0.5	0.2837	0.2002	0.4406	0.2947	0.6075	0.4199
		0.7	0.2359	0.1978	0.3720	0.2950	0.5012	0.3985
		0.9	0.2089	0.1956	0.3117	0.2899	0.4209	0.3931

3.2 Application to real data

In this section, the T_2 was applied to the data set from Crowder and Hand (1989). This data set was used to study the effect of a vitamin E diet supplement on the growth of guinea pigs. For each animal, the body weight (in grams) was recorded at the end of weeks 1, 3, 4, 5, 6, and 7. All animals were given a growth-inhibiting substance during week 1 and vitamin E therapy at the beginning of week 5. Three groups comprising five animals each received zero, low, and high doses of vitamin E. The main issue was the possible difference in the growth profiles of the groups.

The sum of the observation ranks in each cluster and the cluster mean of observations are given in Table 5. Using the T_2, we found that the testing statistic was equal to 0.740 and that the exact p-value was equal to 0.725. The T_{KW} was equal to 1.340 with an exact p-value equal to 0.538. The results of the two tests led to no significant difference in the growth profiles of the three groups.

Table 5. Sum of observation ranks in each cluster and cluster mean of observations

Group	Animal (Cluster)	Cluster mean	Cluster rank sum
1	1	471.83	488
	2	561.17	255
	3	558.83	253
	4	571.17	221
	5	532.67	357
2	6	552.00	303
	7	512.67	395
	8	578.67	224
	9	621.67	123
	10	596.33	178
3	11	600.33	162
	12	560.17	272
	13	550.17	296
	14	595.50	171
	15	524.83	374

4. CONCLUSION

Clustered data are common in scientific research. In this study, we applied the procedure of the Wilcoxon test to construct the adjusted rank test for the observation ranks of clustered data of two independent samples. For a balanced design, this test uses the same critical values that benefit researchers in testing the differences between two central tendency populations. We also considered the clustered rank tests for three or more populations with clustered data. In this case, the T_1 was modified by using the sums of the observation ranks in a cluster to compute the test statistic. In addition, the procedure of the Kruskal-Wallis test was applied to adjust the sums of the observation ranks with clustered data as the raw data, i.e., the T_2. For a balanced design, under data sets with the same numbers of clusters in samples, the T_2 also uses the same critical values. The simulation study showed that the adjusted rank tests could maintain the probabilities of Type I errors for all situations. Given a small number of clusters (n_1, n_2) = (5,5) and (7,7) and correlation coefficients ($\rho < 0.7$), the T has the highest empirical power. The T_2, has higher empirical power than the Kruskal-Wallis based on mean cluster for all situations. The powers of the two adjusted rank tests increase when the effect size, number of clusters, and number of observations increase. However, the powers of the adjusted rank tests decrease when the correlation coefficients between observations in clusters increase.

ACKNOWLEDGMENT

This work was supported by Silpakorn University Research, Innovation and Creativity Administration Office, Thailand. Special thanks are extended to the referees for their valuable suggestions.

REFERENCES

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31(2), 144-152.

Crowder, M. J., and Hand, D. J. (1990). *Analysis of repeated measures*, Chapman & Hall. pp. 27-28.

Hodges, J. L., and Lehmann, E. L. (1956). The efficiency of some nonparametric competitors of the *t*-test. *Annals of Mathematical Statistics*, 27, 324-335.

Kruskal, W. H., and Wallis, W. A. (1952). Use of ranks in one criterion variance analysis. *Journal of the American Statistical Association*, 47(260), 583-621.

Lahiri, P., and Yan, L. (2009). A new alternative to the standard *F* test for clustered data. *Journal of Statistical Planning and Inference,* 139(10), 3430-3441.

Mood, A. M. (1954). On the asymptotic efficiency of certain nonparametric two-sample tests. *The Annals of Mathematical Statistics*, 25(3), 514-522.

Rao, J. N. K., Sutradhar, B. C., and Yue, K. (1993). Generalized least squares *F* test in regression analysis with two-stage cluster samples. *Journal of the American Statistical Association,* 88(424), 1388-1391.

Rosner, B., and Grove, D. (1999). Use of the Mann-Whitney *U*-test for clustered data. *Statistics in Medicine*, 18(11), 1387-1400.

Rosner, B., Glynn, R. J., and Lee, M. L. T. (2003). Incorporation of clustering effects for the Wilcoxon rank sum test: a large-sample approach. *Biometrics*, 59(4), 1089-1098.

Rosner, B., Glynn, R. J., and Lee, M. L. T. (2006). The Wilcoxon signed rank test for paired comparisons of clustered data. *Biometrics*, 62(1), 185-192.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6), 80-83.

Wu, C. F. J., Holt, D., and Holmes, D. J. (1988). The effect of two-stage sampling of the *F* statistic. *Journal of the American Statistical Association*, 83(401),150-159.