

## A Modified Jonckheere Test Statistic for Ordered Alternatives in Repeated Measures Design

Hatice Tül Kübra AKDUR<sup>1,\*</sup>, Fikri GÖKPINAR<sup>1</sup>, Hülya BAYRAK<sup>1</sup>, Esra GÖKPINAR<sup>1</sup>

<sup>1</sup>Gazi Üniversitesi, Fen Fakültesi, İstatistik Bölümü, 06500, Ankara

(Alınış / Received: 06.09.2015, Kabul / Accepted: 25.07.2016, Online Yayınlanma / Published Online: 08.09.2016)

### Keywords

Repeated measures,  
Jonckheere test,  
Circular bootstrap,  
Type 1 error

**Abstract:** In this article, a new test based on Jonckheere test [1] for randomized blocks which have dependent observations within block is presented. A weighted sum for each block statistic rather than the unweighted sum proposed by Jonckheere is included. For Jonckheere type statistics, the main assumption is independency of observations within block. In the case of repeated measures design, the assumption of independence is violated. The weighted Jonckheere type statistic for the situation of dependence for different variance-covariance structure and the situation based on ordered alternative hypothesis structure of each block on the design is used. Also, the proposed statistic is compared to the existing test based on Jonckheere in terms of type I error rates by performing Monte Carlo simulation. For the strong correlations, circular bootstrap version of the proposed Jonckheere test provides lower rates of type I error.

## Tekrarlı Ölçümler Tasarımında Sıralı Alternatifler için Uyarlanmış Jonckheere Test İstatistiği

### Anahtar Kelimeler

Tekrarlı ölçümler,  
Jonckheere test,  
Döngüsel bootstrap,  
1. Tip hata

**Özet:** Bu makalede, blok içinde bağımlı ölçümler içeren rasgele bloklar için Jonckheere testine dayalı yeni bir test sunulmuştur. Jonckheere tarafından önerilen ağırlıksız toplam yerine her blok istatistiği için ağırlıklı toplam dahil edilmiştir. Jonckheere tipi istatistikler için ana varsayım blok içi gözlemlerin bağımsız olmasıdır. Tekrarlı ölçümler tasarımı durumunda, bağımsızlık varsayımı bozulur. Tasarımdaki her bloğun sıralı alternatif hipotez yapısına dayalı durumu ve farklı varyans-kovaryans yapılarının bağımlılık durumları için ağırlıklı Jonckheere tipi test istatistiği kullanılmıştır. Ayrıca, önerilen istatistik Jonckheere testine dayalı var olan test ile Monte Carlo simülasyonu yoluyla 1. Tip hata oranları bakımından karşılaştırılmıştır. Güçlü korelasyonlarda, önerilen Jonckheere testinin döngüsel bootstrap versiyonu daha düşük 1. Tip hata oranı sağlamıştır.

### 1. Introduction

In medical research, testing the direction change or trend of response levels over time/treatment is very considerable problem. For ordered alternatives in a randomized block design, Jonckheere and Page tests are well-known distribution-free tests [1,2]. Zhang and Cabilio [3] developed a generalized Jonckheere test against ordered alternatives for repeated measures in a randomized block design. Repeated measures designs are often just special cases of randomized block designs. This is also the case when a particular sequence is of interest and is used by all subjects- as is the case in many learning experiments. Agresti and Pendergast [4] proposed a rank test to detect treatment effects in repeated measures

designs where the observations within blocks are assumed to be correlated. Also, Kepner and Robinson suggest a RT-type statistics similar to the one proposed by Page [2]. The RT technique was described in detail by Conover and Iman [5] and compared efficiency of RT and F (ANOVA) statistics based on their asymptotic relative efficiency for the same problem. Repeated measures design tends to be serially correlated, and hence the usual assumptions of analysis of variance or linear regression analysis cannot be applied. The main difference between completely randomized and repeated measures designs is that in the former it is often reasonable to assume independence across all observations, while in the latter case observations are likely to be dependent. If this dependence is not taken into

account then inference about treatment effects may be biased or may fail to reveal the true significance of treatment contrasts. Each experimental unit is considered as a block of the design, and the model is shown as in Equation (1):

$$X_{ij} = \beta_i + \tau_j + e_{i,j} \quad i = 1, \dots, b; j = 1, \dots, n \quad (1)$$

$$e_{i,j} = \rho e_{i,j-1} + \varepsilon_{i,j}; \quad \varepsilon_{i,j} \sim N(0,1), \quad (2)$$

where  $n$  is the length of time period of repeated measures of the  $i^{th}$  block or experimental unit,  $b$  is the number of experimental unit,  $\tau_j$  is the time effect of the  $j^{th}$  time period,  $\beta_i$  is the effect of  $i^{th}$  block and finally  $\varepsilon_{i,j}$ 's are independent white noise. Since the observations in each block are dependent, error terms  $\{e_{i,j}\}$  are also assumed dependent and they have zero mean. This dependency weakly forms dependent stationary process, autoregressive  $AR(1)$  model.  $AR(1)$  model assumes that repeated measurements have a first-order autoregressive relationship. The correlation between any two elements is equal to  $\rho$  for adjacent elements,  $\rho^2$  for elements that are separated by a third, and so on.  $\rho$  is constrained so that  $-1 \leq \rho \leq 1$ . As the time interval increases between experimental units, correlation structure of error terms weakens [3]. Since parametric repeated measures analysis has commonly stationary autoregressive model, we will focus on  $AR(1)$  structure in our study. A very common problem in applied statistics is to test the differences between the effects of the treatments. The null and alternative hypothesis are as seen below:

$$H_0: \tau_1 = \tau_2 = \dots = \tau_n; \quad H_1: \tau_i \neq \tau_j, \exists i \neq j. \quad (3)$$

In this alternative hypothesis, it is assumed that at least one treatment is different from the others. However, sometimes we have prior knowledge of these treatments that increase or decrease. For example, dose-response studies or psychological experiments are frequently suitable to use ordered alternative hypothesis. When the repeated measures design has ordered alternative hypothesis, it is usually shown as a longitudinal design. In longitudinal designs, the same experiment subjects are measured at different time intervals. The main focus of this kind of experiment is the study of change over time or development or growth. For instance, aggression in students who were transitioning from primary school into middle school was studied by Pellegrini and Long [6]. They elicited aggression ratings for each student from their peers and teachers at four different time points: the beginning and end of the last year of primary school, and the beginning and end of the first year of secondary school. The analysis focused on the mean change across time. From primary to secondary school, the major hypothesis was confirmed as mean aggression increased. This was attributed to students establishing (or reestablishing) the social dominance

hierarchy through aggression after a major environmental change [6]. In this paper, our main attention is to examine ordered alternatives studies. Null and alternative hypothesis of ordered alternatives are given as:

$$H_0: \tau_1 = \tau_2 = \dots = \tau_n; \quad H_1: \tau_1 \leq \tau_2 \leq \dots \leq \tau_n, \quad (4)$$

where at least one of the inequalities is strict.

In testing ordered hypotheses, Page [2] and Jonckheere [1] develop test statistics for randomized blocks. Mostly, tests presenting for randomized block require the observations in the model to be independent and identically distributed. However, the data collected from the real-world usually does not fit these assumptions. For instance, each experimental subject is measured more than once to a treatment over a time period. Repeated measurements designs, crossover or changeover designs, time series designs and before-after designs are used in this kind of situations. Both Jonckheere and Page tests are based on rank correlations between the within-block rankings and the criterion alternative ranking. For the Jonckheere test, it is Kendall's tau correlation, while for the Page test, it is Spearman's rank correlation. Mann-Kendall (MK) test that is derived from a rank correlation test (Kendall's Tau) is the widely known as nonparametric method for trend detection [7]. It is widely known the rank correlation test for two sets of observations. Although MK test is a distribution free test, it has such an assumption that observations must be serially independent. Von Storch [8] states that "positive correlation increases the possibility of the null hypothesis of no trend to be rejected when it is correct, with a probability larger than the assigned (nominal) level of significance". This situation leads to inflation of the variance of the MK test by serial correlation. Many authors have attempted to modify the MK test for serially dependent data to decrease the rejection rate when there is no trend to the nominal significance level [9]. Cabilio and Zhang [3] used a generalized version of Jonckheere statistic based on MK statistics for testing the ordered alternative hypothesis. They also obtained the asymptotic distribution of this statistic and used different dependent structure for this testing procedure. As seen from their simulation study, the obtained empirical type I error rates are not close to their nominal levels sufficiently. For this reason, we modified this test statistic and, investigated circular block bootstrap and independent and identically bootstrap methods based on Jonckheere test and modified Jonckheere test. This article is organized as follows. The Jonckheere test and modified version of Jonckheere test are introduced briefly. Also bootstrap methods are examined for the cases dependent series and independent series of a dataset. After this section, we compare the performance of these tests with simulation study in terms of the type 1 error

rates of tests. Eventually, we provide a summary for our findings.

**2. Tests for Ordered Alternatives in Repeated Measures of Randomized Blocks**

In this section, we examine the Jonckheere test and modified version of Jonckheere test given as follows.

**2.1. Jonckheere test**

The Jonckheere test is based on Mann-Kendall statistic [7]. For this reason, we initially give the Mann-Kendall statistic. The Mann-Kendall statistic can be formulated as follows:

$$A = \sum_{i < j} a_{ij} b_{ij} = \sum_{i < j} sgn(X_j - X_i) sgn(Y_j - Y_i) \quad (5)$$

where

$$a_{ij} = \begin{cases} 1, & X_i < X_j \\ 0, & X_i = X_j \\ -1, & X_i > X_j \end{cases} \quad (6)$$

and  $b_{ij}$  can be similarly defined by observations of  $Y$ . Under the null hypothesis that  $X$  and  $Y$  are independent and randomly ordered, the statistic  $A$  tends to normal distribution for large  $n$ , with mean and variance as follows [9]:

$$E(A) = 0 \text{ and } Var(A) = n(n - 1)(2n + 5)/18.$$

If the values in  $Y$  are replaced with the time order of a time series, *i.e.*,  $1, 2, \dots, n$  as in repeated measures design, the test can be used as a trend test [7]. In this case, the statistic  $A$  reduces to the following equation,

$A = \sum_{i < j} a_{ij} = \sum_{i < j} sgn(X_j - X_i)$ . The derivation of the mean and variance of  $A$  is given in detail by Kendall [7].

Based on Mann-Kendal statistic, the Jonckheere test [10] is defined as follow:

$$J = \frac{1}{b} \sum_{i=1}^b T_k(i) \quad (7)$$

In the Jonckheere test formula,  $T_k(i) = \binom{n}{2}^{-1} A_k(i)$  is Mann-Kendal statistic for testing trend in a series of observations and  $A_k(i) = \sum_{l < m} sgn(\mu_i(m) - \mu_i(l))$  is the non-standardized Kendall's tau correlation between a subject's repeated responses and the alternative ordering where each subject is ranked within itself over time. Also,  $\mu_i(j)$  is the rank of the  $i^{th}$  subject at time  $j$ , and  $sgn(\mu_i(m) - \mu_i(l))$  is either 1 or -1, depending on whether  $\mu_i(m) > \mu_i(l)$  or  $\mu_i(m) < \mu_i(l)$ .

**2.2. Modified Jonckheere test**

In order to test  $H_0$  versus  $H_1$  in Eq. (4), Skillings and Wolfe [10] used weighted Jonckheere test statistics in a randomized block design [10]. In this paper, we modified the Jonckheere test statistics for repeated measures design by using the weighting given in Skillings and Wolfe (1978). Modified Jonckheere Test statistic is proposed for repeated measures designs as follow:

$$MJ = \frac{1}{b} \sum_{i=1}^b T_k(i). \quad (8)$$

In the modified Jonckheere test formula,  $T_k(i) = \binom{n}{2}^{-1} A_k(i)$  is Mann-Kendal statistic for testing trend in a series of observations and  $A_k(i) = \sum_{l < m} (m - l) sgn(\mu_i(m) - \mu_i(l))$ .

**3. Independent and Identically Distributed and Circular Bootstrap Method**

One of the most powerful nonparametric tools is bootstrap methods for estimating certain statistical properties of interest. In the bootstrap methods, to get new samples with the same number of samples from the original dataset, the core dataset are shuffled many times with replacement. The test statistic is obtained from each bootstrap sample. Since the shuffling of the core data will exterminate trend, distribution of test statistics become trend-free. By comparing the original test statistic with the new trend-free test statistic, type1 error rates can be estimated. Resampling with or without replacement from a given dataset  $X_1, X_2, \dots, X_n$  to give a distribution of estimator is the main function of the bootstrap method. One of the bootstrap methods is the independent and identically distributed (IID) bootstrap method by performing picking from  $\{1, \dots, n\}$  with equal probability with replacement. The IID bootstrap method is based on the notion that the data  $X_1, X_2, \dots, X_n$  only represents a single realization of all possible combinations of a sample. Therefore, the distribution of an estimate  $\hat{\theta}$  obtained by IID bootstrap method is an estimate of  $\hat{\theta}$  when  $X_1, X_2, \dots, X_n$  derived from a common distribution function  $P(X \leq x)$ . One of the other bootstrap methods is the block bootstrap method proposed by Kunsch [11] for a stationary dependent data. One of the most commonly used block bootstrap methods is the circular block bootstrap introduced by Politis and Romano [12]. When the dataset is serially dependent, bootstrap method is performed in circular blocks. Hence, the correlation of the original dataset pattern is saved. Dehling and Wendler [13] gave the properties of circular block bootstrap method for U-statistics when the observations are dependent. Let  $X_{i1}, X_{i2}, \dots, X_{in}$  be a sample in the  $i^{th}$  block for  $n^{th}$  time interval. It is shown that circular block length as  $cl$ , and the number of circular block as  $cb = \lceil n/cl \rceil$  from now on, brackets mean the largest integer

which is equal to a smaller than the  $n/cl$ . In the simulation study, we choice  $cl = \sqrt{n}$ , it balances variance ratio for positive and negative correlation terms of model. Thus, we used it as the optimal block length for the following simulation study. Each sampled circular block include consecutive observations of length  $cl$ . The  $cl$  consecutive observations are taken from  $X_{i1}, X_{i2}, \dots, X_{in}$  depending upon starting point picked randomly from  $\{1, \dots, n\}$ .

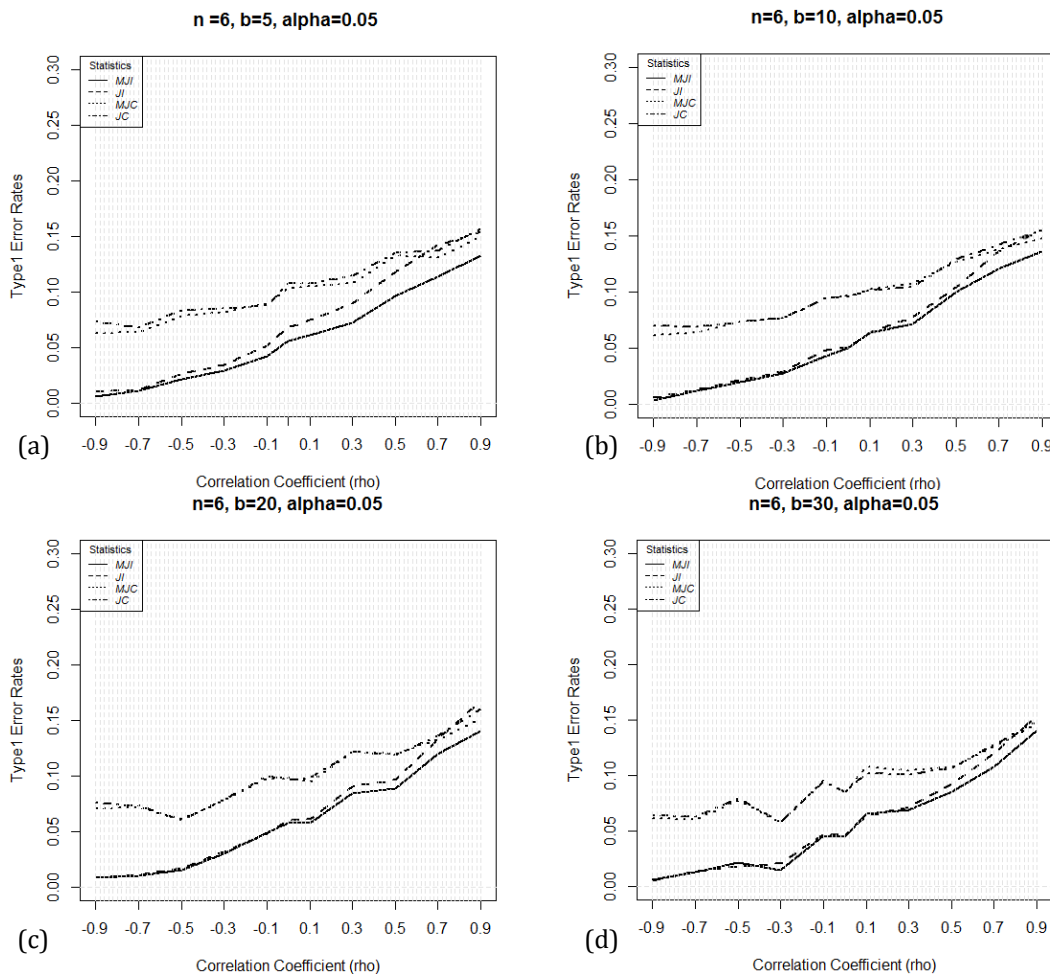
In the article, the circular block bootstrap and IID bootstrap methods are used to generate observations of repeated measures. Bootstrap samples based on two methods are used to generate Jonckheere-type statistics.

**4. Simulation Study**

In this section, the performances of Jonckheere test statistic and modified Jonckheere test statistic based on two different bootstrap methods are investigated in terms of type 1 error rates. Jonckheere test statistic based on IID bootstrap (JI), Jonckheere test statistic based on circular bootstrap (JC), modified Jonckheere test statistic based on IID bootstrap (MJI) and

modified Jonckheere test statistic based on circular bootstrap (MJC) are compared in the Monte Carlo simulation study in terms of type 1 error rates. In order to investigate effect of the correlation term on the Jonckheere based statistics, we perform R 2.9.2 statistical software program to simulate of MJI, JI, MJC, JC for time series whose errors are modeled by AR(1) variance-covariance structures.

The size of each bootstrapping sample is 3000 to estimate type1 error rates of the four statistics, and nominal level of significance is chosen as 0.05. We have conducted simulations with the values of  $b$  (5, 10, 20, 30) for the each values of  $n$  (6, 9, 12, 24) with different correlation term of AR(1) model ( $\rho = -0.9, -0.7, -0.5, -0.3, -0.1$ ) ( $\rho = 0, 0.1, 0.3, 0.5, 0.7, 0.9$ ) and standard normal white noise. The simulation study also provides opportunity to investigate effect of negative, positive correlation terms and independent case ( $\rho = 0$ ). Figures 1,2,3,4 display type1 error rates of all tests corresponding to nominal significance level of 0.05 based on IID bootstrap and circular bootstrap methods.



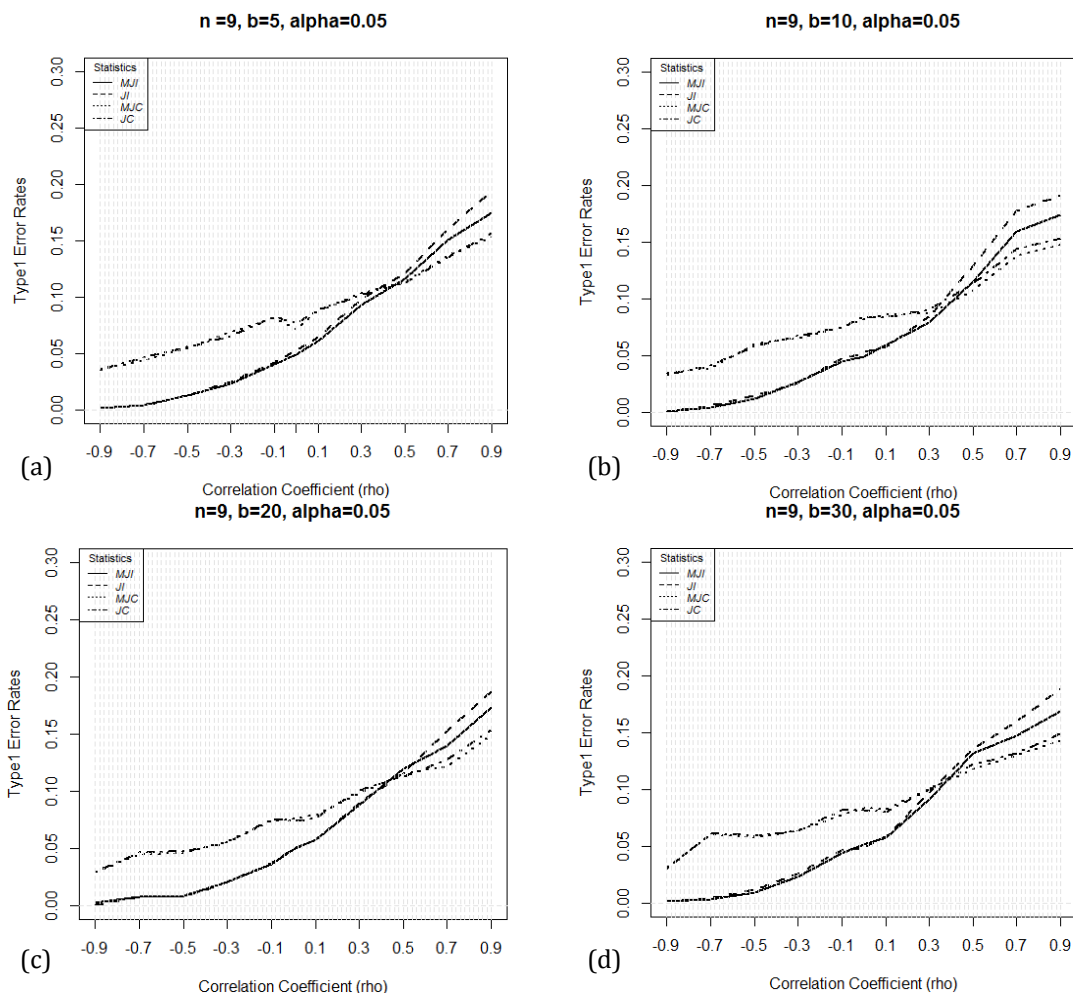
**Figure 1(a).**Type 1 Error Rates of  $n = 6$  and  $b = 5$  for different correlation coefficients of AR(1) model, **(b).**Type 1 Error Rates of  $n = 6$  and  $b = 10$  for different correlation coefficients of AR(1) model, **(c).**Type 1 Error Rates of  $n = 6$  and  $b = 20$  for different correlation coefficients of AR(1) model, **(d).**Type 1 Error Rates of  $n = 6$  and  $b = 30$  for different correlation coefficients of AR(1) model.

As seen in the Figure 1, the estimated type 1 error rates of four tests for negative values of  $\rho$  are lower than the test result of positive values of  $\rho$ . When the block size is 5, for negative values of  $\rho$  MJI generates the lowest type1 error rates among four of them. For the independent case, MJI generates the closest value to the nominal level. For the positive values of  $\rho$ , MJI also provides closer values to nominal level. When block size is increased to 10, 20 and 30, MJI gives the lowest values of type1 error rates even for  $\rho = 0.9$ . As seen in the Figure 1, there is slight difference between bootstrap methods for the combination  $n$  and  $b$  except for  $\rho = 0.9$ . Furthermore, it seems that when block size is increased, estimated type1 error rates are decreased for all test statistics.

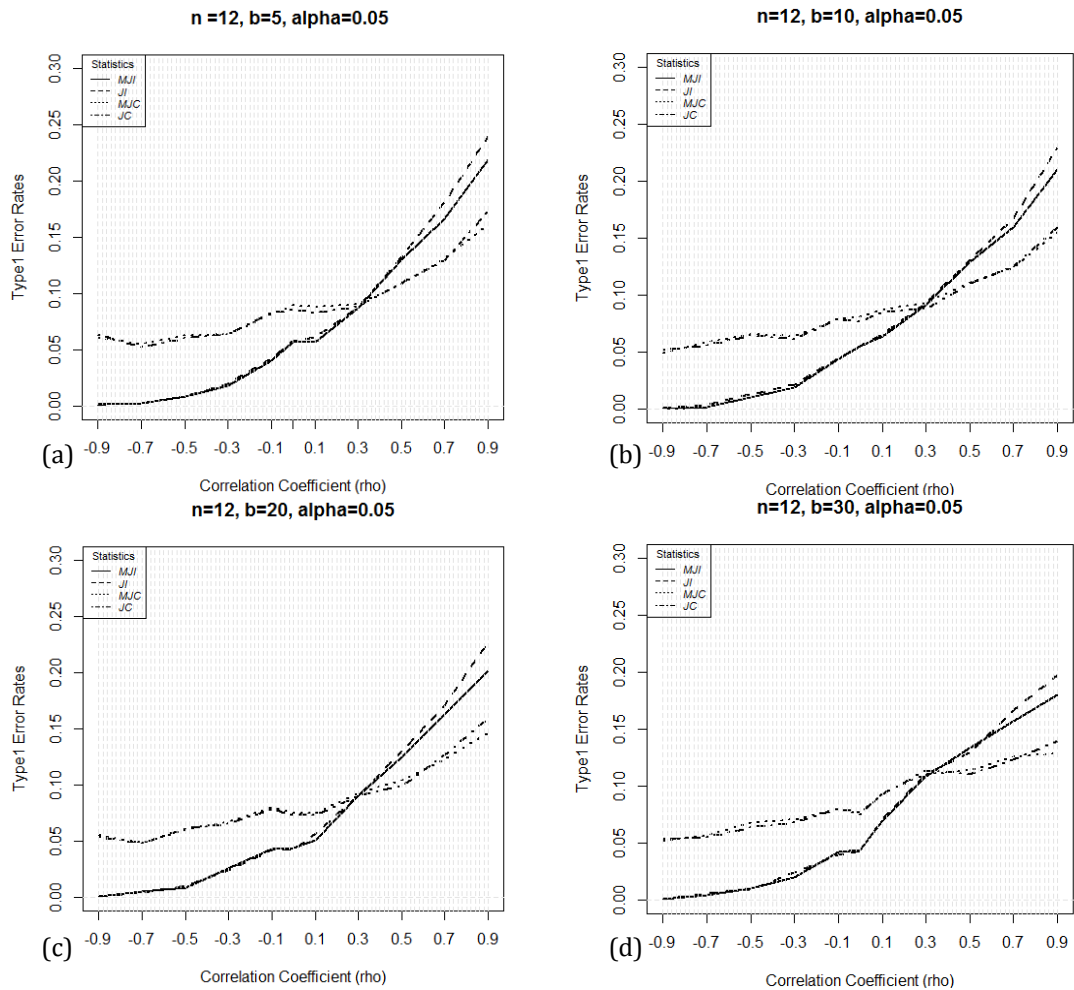
The lowest results of type 1 error are generated by MJI for the block size 5 for the negative values of  $\rho$  and weak correlation values ( $\rho = 0.1, 0.3$ ). For the higher positive correlation values ( $\rho = 0.5, 0.7, 0.9$ ) MJC gives the higher value of type 1 error for the block size 5. For the block sizes  $b = 10, 20, 30$ , simulation results show similar trend with the block size 5. The test statistics based on circular bootstrap

(JC and MJC) provides closer results to each other in terms of type 1 error rates. After moderate positive values of  $\rho$ , MJC and JC generates lower type 1 error rates.

As shown on Figure 3 (a), the lowest type1 error rates are given by MJI for the negative correlation term for the smallest block size. Also, it is seen that MJC and JC generate very close type1 error rates to the nominal level for the negative values of  $\rho$ . For the lower positive values of  $\rho$ , MJI has lower results of type1 error rates. For the higher positive values of  $\rho$ , MJC generates better type1 error rates than JC does. For the block sizes  $b = 10, 20, 30$ , the trend of four tests stays same with  $b = 5$  yet, type1 error rates are decreased. For example, for the block size 5 and  $\rho = 0.9$ , MJC gives type 1 error rate as 0.1615. As block size is increased to 30 for  $\rho = 0.9$  type 1 error rate is decreased to 0.1284 for MJC.  $\rho = 0.3$  is intersection point for all test of  $n = 12$ , circular bootstrap-based tests provide lower type1 error rates for  $\rho = 0.5, 0.7, 0.9$  as seen in Figure 3. As positive correlation terms are increased, MJC and JC give lower type1 error rates than MJI and JI.



**Figure 2(a).**Type 1 Error Rates of  $n = 9$  and  $b = 5$  for different correlation coefficients of AR(1) model, **(b).**Type 1 Error Rates of  $n = 9$  and  $b = 10$  for different correlation coefficients of AR(1) model, **(c).**Type 1 Error Rates of  $n = 9$  and  $b = 20$  for different correlation coefficients of AR(1) model, **(d).**Type 1 Error Rates of  $n = 9$  and  $b = 30$  for different correlation coefficients of AR(1) model.



**Figure 3(a).**Type 1 Error Rates of  $n = 12$  and  $b = 5$  for different correlation coefficients of  $AR(1)$  model, **(b).**Type 1 Error Rates of  $n = 12$  and  $b = 10$  for different correlation coefficients of  $AR(1)$  model, **(c).**Type 1 Error Rates of  $n = 12$  and  $b = 20$  for different correlation coefficients of  $AR(1)$  model, **(d).**Type 1 Error Rates of  $n = 12$  and  $b = 30$  for different correlation coefficients of  $AR(1)$  model.

As given in Figure 4, the lowest type1 error rates are generated by MJI when correlation terms are negative. It can be seen that that MJC generates very close type 1 error rates to the nominal level for the negative values. When the observation terms are independent, MJI gives the closest type 1 error rates to the nominal alpha. MJI gives lower actual type1 error rates when correlation terms are moderate ( $\rho = 0.3, 0.1$ ). For the higher positive ( $\rho = 0.5, 0.7, 0.9$ ) correlation terms, MJC generates more reasonable type 1 error rates. The results above stays almost similar with  $b = 5$  when the block size is increased to 10, 20 and 30. However, type 1 error rates are decreased about %2. For example, when the block size is and  $\rho = 0.9$ , MJC gives type 1 error rate as 0.1585. As block size is increased to 30 for  $\rho = 0.9$  type 1 error rate is decreased to 0.1375 for MJC. The intersection point is  $\rho = 0.1$  for all test statistics of  $n = 24$  that means it is hard to recognize difference of four tests on this point. Also, circular bootstrap-based tests indicate lower type1 error rates for almost all positive correlation terms except for  $\rho = 0.1$  for each block size.

### 5. Numerical Example

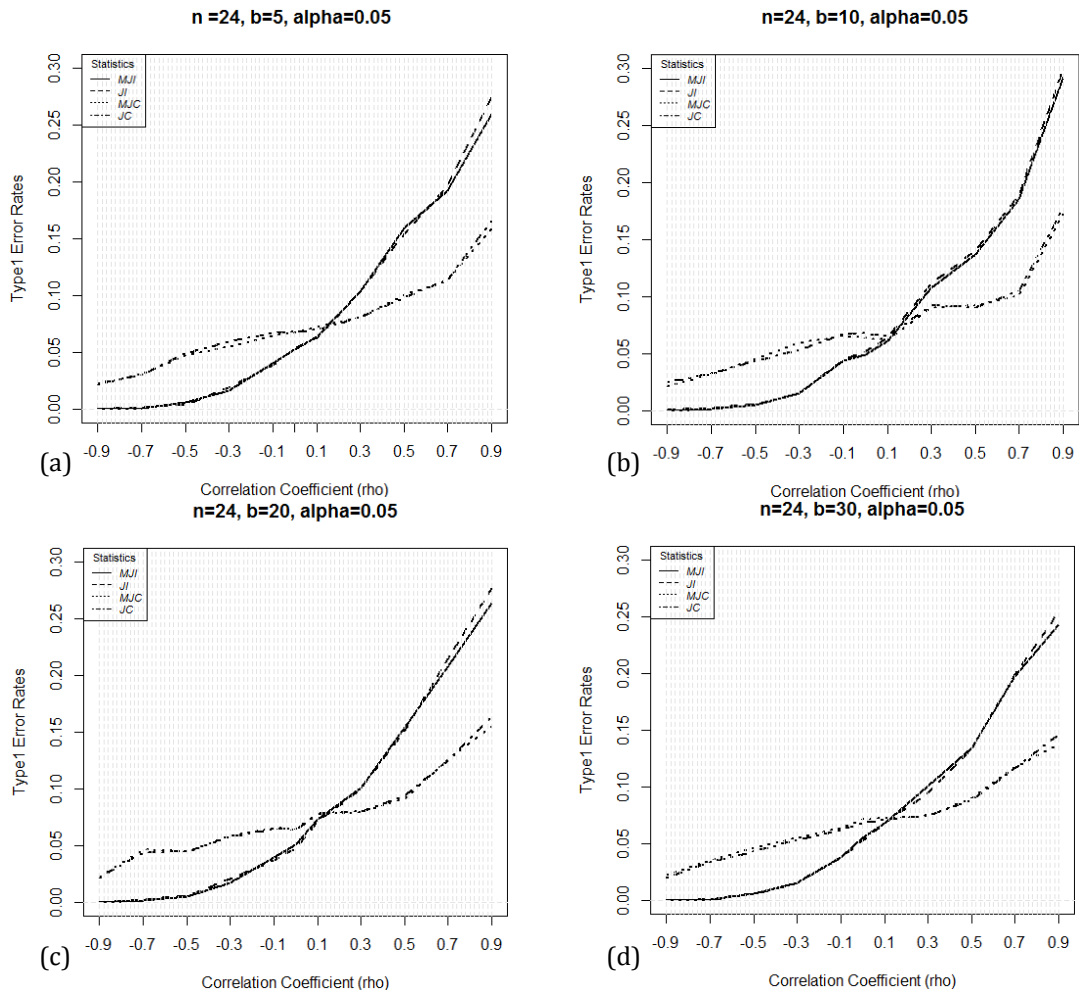
A numerical example is given in order to apply the test statistics to dataset that are generated in Table 1. Consider that six depression patients are given a drug that increases levels of ‘happy chemical’ in the brain. At baseline, all six patients have similar levels of this chemical. Researcher measures brain-chemical levels at three subsequent time points after given drug: at 2 months, 3 months and 6 months post-baseline.

Since there is a natural personal variability, we can take each subject as a block. Because experiment subject stays same for each time period, we need to take into consideration the dependence in repeated observations. The both Jonckheere test statistics are appropriate for such a situation. For this dataset, we can define our null and ordered alternative hypothesis as given below based on 3000 bootstrap samples:

$$H_0: \tau_1 = \tau_2 = \tau_3 = \tau_4; \quad H_1: \tau_1 \leq \tau_2 \leq \tau_3 \leq \tau_4$$

the bootstrap p-value results of these four statistics is obtained as shown in Table 2.





**Figure 4(a).**Type 1 Error Rates of  $n = 24$  and  $b = 5$  for different correlation coefficients of  $AR(1)$  model, **(b).**Type 1 Error Rates of  $n = 24$  and  $b = 10$  for different correlation coefficients of  $AR(1)$  model, **(c).**Type 1 Error Rates of  $n = 24$  and  $b = 20$  for different correlation coefficients of  $AR(1)$  model, **(d).**Type 1 Error Rates of  $n = 24$  and  $b = 30$  for different correlation coefficients of  $AR(1)$  model.

**Table 1.** Hypothetical dataset for repeated measures of blocks.

	Time Periods			
	Baseline	Time1	Time2	Time3
Subject1	1000	1100	1200	1300
Subject2	1000	1000	1005	950
Subject3	1000	1999	800	1700
Subject4	1000	1100	1150	1100
Subject5	1000	1000	1050	1010
Subject6	1000	1100	1109	1500

**Table 2.** p-values for assessing time effect for the numerical example.

MJI	0.0126
MJC	0.0053
JI	0.0050
JC	0.0023

All test statistics results indicate that we have enough evidence to reject the null hypothesis at the significance level. It can be concluded that the drug given the patients during the depression treatment period increases the happy chemical level in the patients brain. Hence, it can be used as a depression medicine for patients treatment based on this result of study.

## 6. Discussion and Conclusion

In clinical or medical research, dataset is usually gathered serially or longitudinally on subjects. The group of repeated measures can be analyzed by parametric methods to uncover the mean profile change with ordered alternatives. The modification of Jonckheere tests were developed in order to make use of information contained blocks that have dependent observations. In the literature, some authors investigate large time interval ( $n = 25, 50, 100$ ) [3],[14]. Due to this attitude, they miss out IID bootstrap importance for small time intervals ( $n = 6$  or  $9$ ). Therefore, we take into consideration small time intervals in this study. As seen from the tables, the strong correlation increases the type1 error rates. For negative values of  $\rho$ , type1 error rates are very smaller than the nominal level, which shows a very conservative test. While, opposite is true, the rates are highly larger than nominal level that indicates a very liberal test. These problems moderately can be solved for larger block sizes. The accuracy of the estimated type 1 error rates improves when block size is increased to 20 or 30. These tests are mostly liberal for small blocks sizes for small value of  $n$ . Type 1 error rates for IID bootstrap-

based tests vary in the range of [0.000; 0.0447] for negative correlation terms and [0.043; 0.3] for positive correlation terms.

As the time intervals are increased, MJC provides lower rates of type1 error for higher positive correlation terms. Generally, when block size gets larger for all the time intervals ( $n = 6, 9, 12, 24$ ), type1 error rates get lower.

After moderate values of  $\rho$ , the bootstrap methods become quickly distinguishable from their type1 error rates. Whereas circular bootstrap method is reasonable for larger correlation terms ( $\rho > 0.5$ ), IID bootstrap can be preferred for smaller values of ( $\rho < 0.5$ ). It is seen that the type1 error rates for these tests are affected by the value of correlation term. Generally, the type1 error rates for these tests are somewhat higher than the nominal. It appears that there is quite difference between the two bootstrap-based methods. As far as we have realized that a similar behavior has been observed for unweighted trend statistics as in [3].

Since any trend will be eliminated due to shuffling, circular bootstrap method is preferred for the dataset which has strong dependent structure as in explained Section 2 broadly. As seen in Figures, circular bootstrap-based tests provide lower type1 error rates for more number of positive correlation terms as  $n$  is increased. For instance, MJC and JC generate lower type1 error rates for combination of  $n = 9$  and  $\rho = 0.7, 0.9$ . For  $n = 24$ , MJC and JC provide lower type1 error rates for  $\rho = 0.3, 0.5, 0.7, 0.9$ . We also should note that the tests will mostly be liberal for small block sizes. These tests should be applied with caution when both  $n$  and  $b$  are small. Especially, for high positive correlation terms, estimated type error rates are increased for large time intervals.

## References

- [1] Jonckheere, A. R. 1954. A test of significance for the relation between  $m$  rankings and  $k$  ranked categories, *British Journal of Statistical Psychology*, 7(1954), 93-100.
- [2] Page, E. B. 1963. Ordered hypotheses for multiple treatments: a significance test for linear ranks. *Journal of the American Statistical Association*, 58(301), 216-230.
- [3] Zhang, Y., Cabilio, P. 2013. A generalized Jonckheere test against ordered alternatives for repeated measures in randomized blocks. *Statistics in medicine*, 32(10), 1635-1645.
- [4] Agresti, A., Pendergast, J. 1986. Comparing mean ranks for repeated measures data. *Communications in Statistics-Theory and Methods*, 15(5), 1417-1433.
- [5] Conover, W. J., Iman, R. L. 1981. Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician*, 35(3), 124-129.
- [6] Pellegrini, A. D., Long, J. D. 2002. A longitudinal study of bullying, dominance, and victimization during the transition from primary school through secondary school. *British journal of developmental psychology*, 20(2), 259-280.
- [7] Kendall, M. G. 1955. Rank correlation methods. Second edition, revised and enlarged. Hafner Publishing Co, Newyork, 196pp.
- [8] Von Storch, H. 1999. Misuses of statistical analysis in climate research. pp 11-26. Von Storch, H., Navarra, A. 1999. Analysis of climate variability: applications of statistical techniques. Springer Science & Business Media, 337p.
- [9] Hamed, K. H., Rao, A. R. 1998. A modified Mann-Kendall trend test for autocorrelated data. *Journal of Hydrology*, 204(1), 182-196.
- [10] Skillings, J. H., Wolfe, D. A. 1978. Distribution-free tests for ordered alternatives in a randomized block design. *Journal of the American Statistical Association*, 73(362), 427-431.
- [11] Kunsch, H. R. 1989. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 1217-1241.
- [12] Politis, D. N., Romano, J. P. 1992. A circular block-resampling procedure for stationary data. pp. 263-270. Lepage, R., Billard, L. 1992. Exploring the limits of bootstrap (Vol. 270). John Wiley & Sons, New York, 437p.
- [13] Dehling, H., Wendler, M. 2010. Central limit theorem and the bootstrap for U-statistics of strongly mixing data. *Journal of Multivariate Analysis*, 101(1), 126-137
- [14] Önöz, B., & Bayazit, M. 2012. Block bootstrap for Mann-Kendall trend test of serially dependent data. *Hydrological Processes*, 26(23), 3552-3560.