Pak.j.stat.oper.res. Vol.16 No.2 2020 pp 387-408 DOI: http://dx.doi.org/10.18187/pjsor.v16i2.2746

Pakistan Journal of Statistics and Operation Research

Permutation Tests for Two-sample Location Problem Under Extreme Ranked Set Sampling

Monjed H. Samuh[1*], Ridwan A. Sanusi[2]

*Corresponding author

1. Department of Applied Mathematics & Physics, Palestine Polytechnic University, Palestine.
2. Department of Community Health Sciences, University of Manitoba, Winnipeg, MB, Canada.

Abstract

In this paper, permutation test of comparing two-independent samples is investigated in the context of extreme ranked set sampling (ERSS). Three test statistics are proposed. The statistical power of these new test statistics are evaluated numerically. The results are compared with the statistical power of the classical independent two-sample t-test, Mann-Whitney U test, and the usual two-sample permutation test under simple random sampling (SRS). In addition, the method of computing a confidence interval for the two-sample permutation problem under ERSS is explained. The performance of this method is compared with the intervals obtained by SRS and Mann-Whitney procedures in terms of empirical coverage probability and expected length. The comparison shows that the proposed statistics outperform their counterparts. Finally, the application of the proposed statistics is illustrated using a real life example.

Key Words: Permutation Test; Extreme Ranked Set Sampling; Power Level; Type I Error Probability.

Mathematical Subject Classification: 62D05, 62G05, 62G09.

1. Introduction

McIntyre (1952) introduced the concept of ranked set sampling (RSS) as a new sampling scheme for data collection. Due to its importance for a variety of applications in statistics, it is republished in McIntyre (2005) to estimate the mean of Australian pasture yields. As claimed by McIntyre, McIntyre (1952, 2005), the mean of the RSS is an unbiased estimator of the population mean. Also, the variance of the RSS mean is smaller than in simple random sampling (SRS) with equal measurement elements. This sampling scheme is useful when it is difficult to measure large number of elements but visually (without inspection) ranking some of them is easier. For example, in McIntyre's experiment the yields of pasture plots can be assessed without the actual laborious process of weighing and mowing the hay for a lot of plots. Moreover, the RSS scheme is also highly applicable in instances where measuring a variable of interest is difficult and risky to measure. For example, in studying some diseases such as the yellowing of the body of an infant, one of the main steps is to measure the bilirubin level of the infant by taking their blood samples. However, it is risky and excruciating to take the blood samples. It is rather easy to rank the babies and take the measurement of the bilirubin level on their urine samples (Paul and Thomas, 2017).

The RSS scheme involves randomly selecting m sets, each of size m elements, from a study population (typically m is in the range 2 to 5). The elements of each set are ordered with regards to the variable of interest by any negligible cost method or visually without measurements. Finally, the i-th minimum from the i-th set, $i = 1, 2, \ldots, m$, is identified for measurement. The obtained sample is referred to as a ranked set sample of set size m. Takahasi and Wakimoto (1968) explained the mathematical theory behind the claims of McIntyre, McIntyre (1952, 2005) by showing that the

efficiency of the RSS mean with respect to SRS, defined by the ratio of the variances of the two sample means, is bounded by 1 and $(m+1)/2$.

Some authors estimate the parameters of a specific distribution using RSS. Bhoj (1997) obtained the estimates of the location and the scale parameters of the extreme value distribution using RSS. Similarly, Abu-Dayyeh et al. (2004) proposed some estimators for estimating the location and the scale parameters of the Logistic distribution using SRS, RSS and some of its other modifications. For more examples, see Lam et al. (1994), Bhoj and Ahsanullah (1996), Chacko and Thomas (2007), Chacko and Thomas (2008), Al-Saleh and Diab (2009), and Sarikavanij et al. (2014), among others.

To better improve the efficiency of the estimators, some variations of RSS were implemented. For example (but not limited to): Samawi et al. (1996) introduced a more practical and efficient variation of RSS which is referred to as the extreme RSS (ERSS); Muttlak (1997) proposed a median RSS as a modification of RSS to decrease ranking error and to improve the efficiency of the estimators being estimated; Al-Saleh and Al-Kadiri (2000) suggested double RSS as a method that improves efficiency of the RSS estimators while keeping m fixed; and Al-Saleh and Al-Omari (2002) suggested a multi-stage RSS as a generalization of double RSS. These variations of RSS were later used to estimate the parameters of some distributions. Shaibu and Muttlak (2004) used ERSS and median RSS to propose linear unbiased estimators and maximum likelihood estimators of the parameters of location-scale family of distributions like; exponential, normal and gamma distributions. They showed that their estimators dominate other existing estimators under ERSS and the estimators are most efficient under median RSS. For more examples, see Adatia (2000).

In the context of testing hypothesis, Koti and Jogesh Babu (1996) derived the exact distribution of the sign test statistic based on RSS. It was reported that the test is more powerful than the counterpart sign test statistic of SRS. Liangyong and Xiaofang (2010) used the sign test statistic of RSS for testing hypotheses about the quantiles of a population distribution. Bohn and Wolfe (1992, 1994) and Bohn and Wolfe (1994) suggested the RSS analogue of the classical two-sample Wilcoxon test and studied its relative properties under perfect and imperfect judgment. Öztürk (1999) studied the effect of the RSS on two-sample sign test statistic. Öztürk and Wolfe (2000) presented an optimal RSS allocation scheme for a two-sample RSS median test. They derived the exact distribution of the two-sample median test statistic in the context of RSS and tabled it for some sample sizes. Samuh (2012), Samuh (2017), and Amro and Samuh (2017) investigated the two-sample permutation test within the context of RSS and multistage RSS. In this paper, a new testing procedure for the two-sample design within ERSS is investigated.

The rest of the paper is structured as follows. The procedure of the ERSS is described in Section 2. The independent two-sample problem is introduced in Section 3. Permutation test for two-sample ERSS with three proposed test statistics is discussed in Section 4. The method of computing a confidence interval for the two-sample permutation problem under ERSS is explained in Section 5. Simulation study that shows the benefits of permutation test of the extreme ranked set two-sample design is reported in Section 6. Illustrative example is used to show the application of this research in Section 7. Finally, summary and concluding remarks are provided in Section 8.

2. Extreme ranked set sampling scheme

Following Samawi et al. (1996), the procedure of the ERSS is described as follows:

1. Randomly select m sets of size m elements each from the study population. These may be denoted as set $1 = \{Y^*_{11}, Y^*_{12}, \ldots, Y^*_{1m}\}$, set $2 = \{Y^*_{21}, Y^*_{22}, \ldots, Y^*_{2m}\}$, and so on till the last set, set $m = \{Y^*_{m1}, Y^*_{m2}, \ldots, Y^*_{mm}\}$. It is assumed that the largest and the lowest elements in each set can be determined virtually or by any negligible cost method. This is, of course, a simple and practical approach.

2. If m is even, measure the lowest ranked element in set 1. Repeat this procedure for set 2 till set $(m/2)$. Represent the measured elements as $Y_1, Y_2, \ldots, Y_{(m/2)}$. Furthermore, measure the largest ranked element in set $(m/2+1)$. Repeat this procedure for set $(m/2+2)$ till the last set, set m. Represent the measured elements as $Y_{(m/2+1)}, Y_{(m/2+2)}, \ldots, Y_m$.

3. If m is odd, measure the lowest ranked element in set 1. Repeat this procedure for set 2 till set $((m-1)/2)$. Represent the measured elements as $Y_1, Y_2, \ldots, Y_{((m-1)/2)}$. Furthermore, measure the largest ranked element in set $((m+1)/2)$. Repeat this procedure for set $((m+3)/2)$ till set $(m-1)$. Represent the measured elements as $Y_{((m+1)/2)}, Y_{((m+3)/2)}, \ldots, Y_{(m-1)}$. Element in the last set can be measured in two different ways:

   (a) Select the average of the measures of the lowest and the largest ranked elements, or

(b) Measure the median ranked element, say Y_m. In this paper, we consider this way.

The acquired sample, $\{Y_1, Y_2, \ldots, Y_m\}$, is called an ERSS of size m.

4. Independently repeat the steps h cycles, if needed, to acquire an ERSS of size $n = h \times m$.

It is worth to note that although m^2 elements are sampled in the first step, only m of them are considered for measurement. In case of perfect ranking (no error was made in the ranking mechanism) the measured elements are called the *order statistics* and they are not ordered (See Navarro et al. (2007) for some examples where order statistics are not ordered); we denote the i-th order statistic acquired in the j-th cycle by Y_{ji}, $i = 1, 2, \ldots, m$, and $j = 1, 2, \ldots, h$. To this end, the ERSS scheme produces a data set as follows

$$\mathbf{Y} = \{Y_{ji}\} = \begin{pmatrix} Y_{11} & Y_{12} & \cdots & Y_{1m} \\ Y_{21} & Y_{22} & \cdots & Y_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{h1} & Y_{h2} & \cdots & Y_{hm} \end{pmatrix}$$

If m is even, the first $m/2$ columns have the distribution of the 1-st order statistic and the last $m/2$ columns have the distribution of the m-th order statistic. If m is odd, the first $(m-1)/2$ columns have the distribution of the 1-st order statistic, the second $(m-1)/2$ columns have the distribution of the m-th order statistic, and the last column has the distribution of the $(m/2)$-th order statistic. Therefore, the data in the same column are identically distributed. Also, all the data are mutually independent.
In this paper, we assume perfect judgment ranking in the selection of the data points for the ERSS. A violation of this assumption is quite interesting and a subject of future work.

3. The independent two-sample design

In this section, the independent two-sample problem is introduced, and the classical independent t-test, Mann-Whitney U test, and the two-sample permutation test are reviewed.
Let us consider the testing problems for one-sided alternative hypotheses as produced by treatments with non-negative effect size δ. Particularly, let $\mathbf{X}_1 = (X_{11}, X_{12}, \ldots, X_{1n_1})$ and $\mathbf{X}_2 = (X_{21}, X_{22}, \ldots, X_{2n_2})$ be independent random samples from $F(x_1)$ and $G(x_2) = F(x_2 - \delta)$, and we wish to test

$$H_0 : \delta = 0 \text{ against } H_1 : \delta > 0. \tag{1}$$

Under the normality assumption of the underlying distribution, the likelihood ratio test statistic for testing the null hypothesis in Equation 1 is given by

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where $\bar{X}_1$ and $\bar{X}_2$ are the sample means of the two samples $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively, and the pooled standard deviation is

$$S_p = \sqrt{(\sum_{i=1}^{n_1}(X_{1i} - \bar{X}_1)^2 + \sum_{i=1}^{n_2}(X_{2i} - \bar{X}_2)^2)/(n_1 + n_2 - 2)}.$$

When H_0 is true, T is distributed as Student's distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom. H_0 is rejected when $|T| > t_\nu^\alpha$, where t_ν^α is the upper α critical value, and α is the level of significance. The statistical power level is given by

$$W(\delta; n_1, n_2, \alpha) = 1 - F_t(t_\nu^\alpha, \nu, ncp),$$

where F_t is the cumulative distribution of the Student t, $\nu = n_1 + n_2 - 2$ is the degrees of freedom, and $ncp = \delta\left(S_p^2(1/n_1 + 1/n_2)\right)^{-1/2}$ is the non-centrality parameter. Note that W is a function of δ for a given sample sizes n_1 and n_2, and preassigned level of significance α. The power level measures how likely to get a significant result given that the alternative hypothesis is true; It measures the probability that the true value of δ will be detected by the test.

In a permutation framework, we can relax the normality assumption. Permutation tests form a subclass of nonparametric tests which do not rely on any particular distribution. To obtain an exact permutation solution, the exchangeability assumption (often referred to as equality in distributions) in the null hypothesis is required to the data points. This assumption is generally assured by randomly assigning experimental units to treatments in experimental studies. In case of observational studies, exchangeability in the null hypothesis shall be assumed. For testing the null hypothesis in Equation 1, a suitable test statistic should be chosen such that, without loss of generality, large values of it are considered to be against H_0. For more details about the choice of the test statistic in the permutation framework, see Page 84 of Pesarin and Salmaso (2010). One may choose $T = \bar{X}_1 - \bar{X}_2$ as a test statistic. For determining the exact p-value, an appropriate reference distribution is needed which is called the permutation distribution. Indeed, the following steps are used to carry out the permutation test for two-sample design.

1. For the given two-independent samples, $\mathbf{X}_1$ and $\mathbf{X}_2$, calculate the observed test statistic, $T_0 = T(\mathbf{X}_1, \mathbf{X}_2)$.
2. Write down the set of all possible permutations of the $n = n_1 + n_2$ observations between the two samples; i.e. the permutation sample space $\mathcal{X}$. The cardinality of this space is $n!$.
3. For each permutation in $\mathcal{X}$, compute the test statistic, $T^* = T(\mathbf{X}_1^*, \mathbf{X}_2^*)$. The cardinality of related space is $\binom{n}{n_1}$.
4. The true p-value is calculated as
$$\lambda_T = \frac{\text{number of } T^*\text{'s} \geq T_0}{\binom{n}{n_1}}.$$
5. For a given preassigned significance level α, the test is declared to be significant if α is greater than the p-value.

Since it is tedious to write down and enumerate the whole members of permutation sample space $\mathcal{X}$, conditional Monte Carlo simulation (Algorithm 1) can be used to approximate the p-value at any desired accuracy.

Algorithm 1 Conditional Monte Carlo (CMC)

1. For the given samples, $\mathbf{X}_1$ and $\mathbf{X}_2$, compute the observed test statistic, $T_0 = T(\mathbf{X}_1, \mathbf{X}_2)$.
2. From $\mathcal{X}$, take a random permutation $(\mathbf{X}_1^*, \mathbf{X}_2^*)$ of $(\mathbf{X}_1, \mathbf{X}_2)$, and compute the corresponding permutation test statistic $T^* = T(\mathbf{X}_1^*, \mathbf{X}_2^*)$.
3. Independently repeat Step 2 a large number of times, say B, giving B values for T^*, say $\{T_b^*, b = 1, \ldots, B\}$.
4. The estimated permutation p-value is
$$\hat{\lambda}_T = \frac{\sum_{b=1}^{B} \mathbb{I}(T_b^* \geq T_0)}{B},$$
where $\mathbb{I}(\cdot)$ is the indicator function.

Note that $\hat{\lambda}_T$ is an unbiased estimate of the true λ_T and, due to the Glivenko-Cantelli theorem (Shorack and Wellner, 1986), as B diverges, it is strongly consistent. A $100(1-\alpha)\%$ approximate confidence interval for λ_T is

$$\hat{\lambda}_T \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{\lambda}_T \left(1 - \hat{\lambda}_T\right)}{B}},$$

where $\sqrt{\hat{\lambda}_T \left(1 - \hat{\lambda}_T\right) / B}$ is the estimated standard error of $\hat{\lambda}_T$, and $z_{\frac{\alpha}{2}}$ is the upper $\alpha/2$ critical level of the standard normal distribution.

In order to properly define the power function, the underlying population distribution must be fully specified; defined in its analytical form and all its parameters. But this is not the case in permutation framework. In practice, the power of permutation test is based on repeated random sampling from some population. The p-value of the permutation test is conditional upon the observations for each sample, and the power is the proportion of p-values that are less than or equal α. Algorithm 2 is used for evaluating the power based on a standard Monte Carlo simulation.

Algorithm 2 Power Function of Permutation Test

1. For the given samples, $\mathbf{X}_1$ and $\mathbf{X}_2$, find an estimate of δ, say $\hat{\delta}$. Then, consider the consequent empirical deviates $\hat{Z} = (\mathbf{X}_1 - \hat{\delta}, \mathbf{X}_2)$.
2. Take a random permutation $\hat{Z}^*$ of $\hat{Z}$. Then for any chosen δ the corresponding dataset $(\mathbf{X}_1^* + \delta, \mathbf{X}_2^*)$ is considered.
3. Use the CMC algorithm (Algorithm 1) to estimate the permutation p-value $\hat{\lambda}_T$.
4. Independently repeat Steps 2 and 3 a large number of times, say R, giving R estimated p-values, say $\{\hat{\lambda}_{T(r)}, r = 1, \ldots, R\}$.
5. Finally, the estimated power level is given by

$$\hat{W}(\delta; n_1, n_2, \alpha, T) = \frac{\sum_{r=1}^{R} \mathbb{I}\left(\hat{\lambda}_{T(r)} \leq \alpha\right)}{R}.$$

6. To obtain the power as a function of δ, Steps 1-5 are repeated for different values of δ.

For more details see Pesarin and Salmaso (2010) (See also Samuh and Pesarin, 2018).

Another nonparametric test, which does not assume a normal distribution, is Mann-Whitney U test. The test statistic is $U = \min(U_1, U_2)$, where

$$U_i = n_1 n_2 + \frac{n_i(n_i + 1)}{2} - R_i, \quad i = 1, 2,$$

and R_i is the sum of the ranks for sample i. A large value of U is considered to be against H_0. To evaluate the power of the Mann-Whitney U test, a standard Monte Carlo method can be used.

4. Two-sample permutation test based on ERSS

Let $\mathbf{Y}^t = \{Y_{ji}^t\}$ and $\mathbf{Y}^c = \{Y_{ji}^c\}$ be two independent ERSS groups. The first denotes the treatment group and the second denotes the control group. In each group, the data are all mutually independent and, aside that, the data in the same column are identically distributed. Therefore, under the null hypothesis of no effect, the exchangeability assumption holds within columns and hence the permutation test can be applied. The data has to be permuted carefully taking into account whether m is odd or even to maintain the diversity of distributions. To this end, if m is even then the first $m/2$ columns of $\mathbf{Y}^t = \{Y_{ji}^t\}$ are permuted with the first $m/2$ columns of $\mathbf{Y}^c = \{Y_{ji}^c\}$ and the other $m/2$ columns of $\mathbf{Y}^t = \{Y_{ji}^t\}$ are permuted with the other $m/2$ columns of $\mathbf{Y}^c = \{Y_{ji}^c\}$. while if m is odd, the first $(m-1)/2$ columns of $\mathbf{Y}^t = \{Y_{ji}^t\}$ are permuted with the first $(m-1)/2$ columns of $\mathbf{Y}^c = \{Y_{ji}^c\}$, the other $(m-1)/2$ columns of $\mathbf{Y}^t = \{Y_{ji}^t\}$ are permuted with the other $(m-1)/2$ columns of $\mathbf{Y}^c = \{Y_{ji}^c\}$, and the last column of $\mathbf{Y}^t = \{Y_{ji}^t\}$ is permuted with the last column of $\mathbf{Y}^c = \{Y_{ji}^c\}$. To carry out the permutation test for this extreme ranked set two-sample design, Algorithm 3 is used.

To this end, three test statistics are proposed:

1. The first proposal is based on the difference between overall means of the two groups;

$$T_1 = \bar{Y}^t - \bar{Y}^c,$$

where $\bar{Y}^k = \sum_{i=1}^{n_k} \sum_{j=1}^{h} Y_{ji}^k / hm$, $k = t, c$ (Assuming balanced design).

2. The second proposal is based on the studentized statistic;

$$T_2 = \begin{cases} \sum_{r=1}^{2} \left(\frac{\bar{Y}_{Er}^t - \bar{Y}_{Er}^c}{\hat{\sigma}_{Er}}\right)^2, & \text{if } m \text{ is even} \\ \\ \sum_{s=1}^{3} \left(\frac{\bar{Y}_{Os}^t - \bar{Y}_{Os}^c}{\hat{\sigma}_{Os}}\right)^2, & \text{if } m \text{ is odd} \end{cases}$$

Algorithm 3 Extreme ranked set two-sample permutation test

1. For the acquired extreme ranked set two-sample data sets $\mathbf{Y}^t$ and $\mathbf{Y}^c$, compute the test statistic, T_0.
2. Form a new matrix $\mathbf{Y} = \mathbf{Y}^t \biguplus \mathbf{Y}^c$ by concatenating $\mathbf{Y}^t$ and $\mathbf{Y}^c$ vertically (Note that the two matrices have m columns).
3. Randomly permute $\mathbf{Y}$ as explained above to get $\mathbf{Y}^*$.
4. Split $\mathbf{Y}^*$ into $\mathbf{Y}^{t*}$ and $\mathbf{Y}^{c*}$ such that $\mathbf{Y}^{t*}$ and $\mathbf{Y}^{c*}$ contain the same number of rows as in $\mathbf{Y}^t$ and $\mathbf{Y}^c$, respectively.
5. Compute the test statistic $T^* = T(\mathbf{Y}^*)$ based on $\mathbf{Y}^* = \mathbf{Y}^{t*} \biguplus \mathbf{Y}^{c*}$.
6. Independently repeat Steps 3-5 a large number of times, say B, giving B test statistics, say $\{T_b^*, b = 1, \ldots, B\}$.
7. The estimated p-value is
$$\hat{\lambda}(\mathbf{Y}) = \frac{\sum_{b=1}^{B} \mathbb{I}\left(T_b^* \geq T_0\right)}{B}.$$

where

$$\hat{\sigma}_{Er}^2 = \sum_{k\in\{t,c\}} \sum_{i\in V_r} \sum_{j=1}^{h} (Y_{ji}^k - \bar{Y}_{Er}^k)^2/(mh-2),$$

$$\hat{\sigma}_{Os}^2 = \sum_{k\in\{t,c\}} \sum_{i\in W_s} \sum_{j=1}^{h} (Y_{ji}^k - \bar{Y}_{Os}^k)^2/(n_s-2),$$

$$\bar{Y}_{Er}^k = 2\sum_{i\in V_r} \sum_{j=1}^{h} Y_{ji}^k/m, \quad \bar{Y}_{Os}^k = 2\sum_{i\in W_s} \sum_{j=1}^{h} Y_{ji}^k/m,$$

$$W_1 = \{1, 2, \ldots, (m-1)/2\}, \quad W_2 = \{(m+1)/2, \ldots, m-1\}, \quad W_3 = \{m\},$$

$$V_1 = \{1, 2, \ldots, m/2\}, \quad V_2 = \{(m+2)/2, \ldots, m\},$$

and

$$n_1 = n_2 = (m-1)h, \quad n_3 = 2h.$$

It is worth to note that the rationale behind this proposal is due to the structure of the data points obtained by ERSS. For example, when m is even, the first $m/2$ columns of $\mathbf{Y}^t$ and $\mathbf{Y}^c$ have the distribution of the 1-st order statistic and the last $m/2$ columns have the distribution of the m-th order statistic. Therefore, the studentized statistic is proposed for more accuracy and to maintain the diversity of distributions.

3. The third proposal is based on partial tests. If m is even, the null hypothesis in Equation 1 is partitioned into 2 independent sub-hypotheses as follows.

$$H_{0r} : \delta_r = 0 \text{ against } H_{1r} : \delta_r > 0, \quad r = 1, 2,$$

where δ_1 (δ_2) is the true difference between the means of the 1-st (m-th) order statistic in the treatment and control groups. Thus, the suggested test statistic will be

$$T_{3r} = \bar{Y}_r^t - \bar{Y}_r^c, \quad r = 1, 2.$$

If m is odd, the null hypothesis in Equation 1 is partitioned into 3 independent sub-hypotheses as follows.

$$H_{0s} : \delta_s = 0 \text{ against } H_{1s} : \delta_s > 0, \quad s = 1, 2, 3,$$

where δ_1 (δ_2, δ_3) is the true difference between the means of the 1-st (m-th, $(m/2)$-th) order statistic in the

treatment and control groups. Thus, the suggested test statistic will be

$$T_{3s} = \bar{Y}_s^t - \bar{Y}_s^c, \quad s = 1, 2, 3.$$

Now, Algorithm 1 is used and this leads to 2 or 3 independent p-values. Finally, these p-values have to be combined for testing the overall hypothesis in Equation 1. The following approaches are considered for combining p-values.

(a) The Fisher approach (Fisher, 1934). It is based on the statistic $X^2 = -2\sum_{i=1}^{m} \log \lambda_i$. Under the null hypothesis, $X^2 \sim \chi^2_{(2m)}$. So, the combined p-value is given by

$$\lambda_F = P(\chi^2_{(2m)} > X^2).$$

(b) The Liptak approach (Liptak, 1958). It is based on the statistic $L = \sum_{i=1}^{m} \Phi^{-1}(1-\lambda_i)$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function. Under the null hypothesis, $L/\sqrt{m} \sim N(0,1)$. So, the combined p-value is given by

$$\lambda_L = P\left(Z > \frac{L}{\sqrt{m}}\right),$$

where Z is the standard normal random variable.

(c) The logistic approach (Mudholkar and George, 1977). It is based on the logit statistic $t = C^{-1}\sum_{i=1}^{m} \log\left[(1-\lambda_i)/\lambda_i\right]$, where

$$C = \sqrt{\frac{m\pi^2(5m+2)}{3(5m+4)}}.$$

Under the null hypothesis, t follows approximately a Student's distribution with $(5m+4)$ degrees of freedom. Hence, the combined p-value is given by

$$\lambda_M = P\left(T_{(5m+4)} > t\right).$$

5. Permutation confidence interval for δ

Suppose that $\hat{\delta}_L$ and $\hat{\delta}_U$ are the lower and upper limits, respectively, for the parameter δ. The interval $(\hat{\delta}_L, \hat{\delta}_U)$ is called a $(1-\alpha)100\%$ confidence interval for δ if

$$P(\hat{\delta}_L \leq \delta \leq \hat{\delta}_U) = 1 - \alpha.$$

The confidence interval for δ contains all values of δ_0 for which the null hypothesis $H_0 : \delta = \delta_0$ versus $H_1 : \delta \neq \delta_0$ is not rejected at level α. The one-sided confidence interval for δ contains all values of δ_0 for which the null hypothesis $H_0 : \delta = \delta_0$ versus $H_1 : \delta > (\text{or } <)\delta_0$ is not rejected at level α.
Pesarin and Salmaso (2010) provided a method to compute a confidence interval for the usual two-sample permutation problem. The method can be adapted to construct a permutation confidence interval for δ under ERSS. Algorithm 4 summarizes this method. To carry out this algorithm, two types of different tolerance specifications must be defined;

1. An error $\epsilon > 0$ to control the reliability of the permutation p-value. It must be related to the number of permutations used in CMC algorithm (B); the smaller ϵ is, the larger B is.
2. A real number η.

Two criteria are used to evaluate the performance of this algorithm; (1) by estimating the empirical coverage probability of the resulting confidence interval, and (2) by calculating the length of the interval. This is done by simulation.

6. Simulation study

A simulation study is carried out to assess the significance level and the power of the proposed test statistics for the two-sample ERSS design and to compare them with the usual two-sample permutation test, Mann-Whitney U test, and the classical two-sample t-test.

Algorithm 4 Permutation Confidence Interval for δ

1. For the given samples, $\mathbf{Y}^t$ and $\mathbf{Y}^c$, find an estimate of δ, say $\hat{\delta}$.
2. Choose a negative value of η and subtract $(\hat{\delta}+\eta)$ from every value of the treatment group $\mathbf{Y}^t$.
3. Use the CMC algorithm to estimate the permutation p-value, $\hat{\lambda}(\eta)$, based on the data sets $\mathbf{Y}^t-(\hat{\delta}+\eta)$ and $\mathbf{Y}^c$.
4. If $|\hat{\lambda}(\eta)-\alpha/2|<\epsilon$, then assign $\hat{\delta}_L=\hat{\delta}+\eta$. Otherwise, repeat Steps 2 and 3 with different value of η.
5. To obtain the upper confidence limit of δ, repeat Steps 2 and 3 with positive values for η until the condition $|1-\hat{\lambda}(\eta)-\alpha/2|<\epsilon$ is satisfied and then assign $\hat{\delta}_U=\hat{\delta}+\eta$.

6.1. Simulation conditions

Different configurations are considered in the simulation study. For each combination of $m=\{3,\,4\}$ and $h=\{3,\,5,\,10\}$, four different distributions are considered; uniform distribution $U(-1/3,1/3)$, normal distribution $N(0,1)$, exponential distribution $Exp(1)$, and gamma distribution $G(4,1)$. Several other combinations were also performed but not reported here, and the results follow the same behavior. The simulation study is performed based on $R=5000$ data sets. The permutation is based on $B=1000$ replications. To examine the significance level of the tests, we set $\delta=0$, while to investigate the power behavior, we select values of δ in the set $\{0.2,\,0.4,\,0.6,\,0.8\}$. The nominal significance level was set to $\alpha=\{0.05,\,0.1,\,0.4\}$. Tables 1-8 contain the results of the study.

It is worth to point out that the power levels of the proposed test statistics are obtained for the same generated two-sample ERSS. Moreover, balanced designs are considered in computing the power levels; that is, each sample of the two-sample ERSS is with set size m and number of cycles h. Also, the size of each sample in the two-sample SRS is $h\times m$, so that power comparisons between the considered test statistics under ERSS and SRS are done by maintaining the same number of observations in both schemes (to insure that the two schemes have the same cost).

6.2. Simulation results

As reported in Tables 1 and 2, the proposed test statistics, for all considered distributions, control the type I error probability at the nominal α level except the third proposed test statistic when the sample size is small (especially when $h<4$). Thus, the test based on the third proposed statistic is conservative for small sample size. For instance, when $h=3$ for all distributions considered, the empirical level of significance of the statistics (SRS, T_1 and T_2) are the closest to their corresponding nominal significance level, while the empirical level of significance of the third proposed statistic (T_3(Fisher), T_3(Liptak), T_3(Logit)) deviates the most from α. Nevertheless, the empirical level of significance of T_3(Fisher) gets closer to the nominal level for high values of α. Furthermore, when $h=5$ under all distributions considered, the empirical level of significance of the third proposed statistic is now closer to α, just as the other test statistics. In fact, they are even much closer than other estimators for high values of α under normal and exponential distributions. Generally, when $h>3$, the empirical level of significance of all the statistics are almost the same as their corresponding nominal level for high values of α.

The empirical power levels are shown in Tables 3-8. For fixed m, h, and α, the power levels of the permutation test statistics based on ERSS (T_1, T_2, T_3), are strictly higher than the power level of the permutation test statistic within SRS and MW for all given δ except at very small sample size ($h<4$) and small nominal significance level. For all considered test statistics, the power levels increase as the effect size δ increases. Moreover, the power levels based on ERSS increase as m, h, and α increase. It can be also seen that the power levels of T_1 and T_2 behave the same for the uniform and normal distributions, and T_2 is more appropriate for the exponential and gamma distributions (which are asymmetric distributions) than for the uniform and normal distributions (which are symmetric distributions). In addition, T_2 behaves better for the exponential distribution than for the gamma distribution in the sense that the skewness is greater for the exponential distribution than for the gamma distribution. Among the considered combining functions, the Liptak combining function is the best for the uniform and normal distributions. All considered combining functions behave almost the same for large sample size. Finally, under normality assumption, Tables 9-11 report the exact power levels for the parametric one-sided two-sample t-test for different values of the sample size. Apparently, the power levels of the permutation test statistics based on ERSS are higher than in the parametric t-test. Contrarily, the power levels of the permutation test under SRS are equivalent to the parametric t-test for a large sample size, but

Table 1: Empirical level of significance from the simulation study, $m = 3$, $h = 3$.

		Nominal level α						
Distribution	Test stat	0.05	0.10	0.20	0.40	0.60	0.80	0.90
Uniform	SRS	0.058	0.109	0.197	0.394	0.615	0.813	0.906
	MW	0.041	0.094	0.194	0.389	0.545	0.797	0.863
	T_1	0.052	0.101	0.197	0.400	0.592	0.802	0.893
	T_2	0.048	0.103	0.205	0.403	0.600	0.803	0.896
	T_3(Fisher)	0.016	0.042	0.131	0.347	0.556	0.752	0.866
	T_3(Liptak)	0.024	0.062	0.150	0.344	0.526	0.709	0.788
	T_3(Logit)	0.023	0.056	0.147	0.340	0.529	0.712	0.795
Normal	SRS	0.049	0.109	0.200	0.420	0.629	0.814	0.911
	MW	0.038	0.089	0.190	0.392	0.549	0.791	0.862
	T_1	0.046	0.097	0.193	0.391	0.590	0.805	0.901
	T_2	0.049	0.097	0.199	0.395	0.595	0.806	0.908
	T_3(Fisher)	0.011	0.044	0.127	0.336	0.541	0.774	0.888
	T_3(Liptak)	0.023	0.058	0.147	0.325	0.523	0.716	0.798
	T_3(Logit)	0.017	0.055	0.145	0.323	0.525	0.721	0.805
Exponential	SRS	0.045	0.105	0.203	0.404	0.603	0.801	0.900
	MW	0.039	0.094	0.194	0.392	0.546	0.799	0.867
	T_1	0.045	0.094	0.187	0.379	0.575	0.793	0.899
	T_2	0.044	0.098	0.194	0.383	0.589	0.795	0.898
	T_3(Fisher)	0.009	0.034	0.122	0.339	0.536	0.763	0.875
	T_3(Liptak)	0.023	0.056	0.143	0.317	0.511	0.700	0.791
	T_3(Logit)	0.015	0.049	0.134	0.312	0.510	0.709	0.797
Gamma	SRS	0.055	0.108	0.217	0.403	0.600	0.778	0.892
	MW	0.041	0.098	0.191	0.383	0.540	0.793	0.865
	T_1	0.047	0.095	0.198	0.399	0.599	0.808	0.901
	T_2	0.045	0.094	0.196	0.397	0.601	0.815	0.910
	T_3(Fisher)	0.011	0.044	0.124	0.339	0.560	0.776	0.891
	T_3(Liptak)	0.025	0.058	0.138	0.334	0.518	0.719	0.800
	T_3(Logit)	0.021	0.052	0.133	0.330	0.522	0.725	0.805

slightly higher for lower sample sizes.

To evaluate the performance of the confidence intervals of δ obtained by Algorithm 4, the empirical coverage probability and the expected length (based on 2000 simulations) of 90% confidence intervals are calculated. Data are simulated from uniform distribution, normal distribution, exponential distribution, and gamma distribution. The results are reported in Tables 12 and 13 for different values of m, h, and δ. It can be seen that, comparing ERSS to SRS, the confidence intervals obtained by ERSS are, on average, shorter than the one obtained by SRS of equivalent sample size. Intervals obtained by Mann-Whitney procedure are the shortest. Everything else being fixed, the length of the intervals decrease as the sample size (the set size m and/or the number of cycle h) increases. The coverage probability (percent of intervals containing δ) is at least as the nominal 90% level for all considered test statistics. When $m = 4$, the coverage probability is a bit lower than the nominal 90% level for those intervals obtained by T_2 and T_3 statistics for the exponential distribution. Moreover, the coverage probability obtained by T_1 does not differ substantially from the one obtained by SRS, but the length of the interval obtained by T_1 is shorter than the one obtained by SRS.

Table 2: Empirical level of significance from the simulation study, $m = 3,\ h = 5$**.**

		Nominal level α						
Distribution	Test stat	0.05	0.10	0.20	0.40	0.60	0.80	0.90
Uniform	SRS	0.049	0.100	0.202	0.407	0.615	0.814	0.910
	MW	0.045	0.096	0.189	0.394	0.592	0.773	0.870
	T_1	0.049	0.100	0.203	0.394	0.587	0.789	0.891
	T_2	0.051	0.101	0.198	0.392	0.589	0.788	0.893
	T_3(Fisher)	0.043	0.098	0.196	0.393	0.586	0.787	0.886
	T_3(Liptak)	0.048	0.095	0.192	0.388	0.580	0.781	0.886
	T_3(Logit)	0.047	0.094	0.194	0.386	0.580	0.782	0.887
Normal	SRS	0.057	0.106	0.203	0.406	0.601	0.804	0.897
	MW	0.046	0.098	0.186	0.387	0.592	0.775	0.867
	T_1	0.056	0.109	0.201	0.399	0.605	0.807	0.900
	T_2	0.059	0.107	0.203	0.406	0.601	0.805	0.900
	T_3(Fisher)	0.053	0.100	0.198	0.397	0.602	0.803	0.906
	T_3(Liptak)	0.054	0.101	0.199	0.398	0.596	0.795	0.896
	T_3(Logit)	0.053	0.102	0.199	0.398	0.598	0.797	0.896
Exponential	SRS	0.053	0.106	0.202	0.395	0.601	0.796	0.898
	MW	0.048	0.099	0.187	0.388	0.593	0.775	0.872
	T_1	0.054	0.104	0.206	0.400	0.608	0.799	0.905
	T_2	0.053	0.109	0.203	0.405	0.600	0.802	0.900
	T_3(Fisher)	0.048	0.094	0.198	0.394	0.599	0.798	0.900
	T_3(Liptak)	0.049	0.101	0.200	0.398	0.594	0.795	0.890
	T_3(Logit)	0.047	0.099	0.198	0.397	0.594	0.795	0.891
Gamma	SRS	0.057	0.103	0.195	0.398	0.603	0.798	0.900
	MW	0.044	0.097	0.184	0.392	0.596	0.778	0.870
	T_1	0.044	0.095	0.195	0.398	0.602	0.804	0.902
	T_2	0.047	0.092	0.197	0.395	0.600	0.800	0.901
	T_3(Fisher)	0.041	0.088	0.188	0.392	0.590	0.802	0.899
	T_3(Liptak)	0.043	0.089	0.191	0.385	0.593	0.799	0.897
	T_3(Logit)	0.044	0.087	0.187	0.385	0.593	0.798	0.898

Table 3: Empirical power levels from the simulation study, the nominal significance level is $\alpha = 0.05$ – Uniform and normal distributions

				Uniform				Normal			
m	h	Test Stat	$\delta \rightarrow$	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
3	3	SRS		0.100	0.196	0.314	0.472	0.107	0.210	0.350	0.504
		MW		0.056	0.100	0.171	0.279	0.057	0.104	0.184	0.304
		T_1		0.143	0.292	0.517	0.746	0.139	0.282	0.530	0.724
		T_2		0.127	0.267	0.466	0.679	0.132	0.250	0.477	0.656
		T_3(Fisher)		0.042	0.114	0.242	0.405	0.042	0.096	0.242	0.397
		T_3(Liptak)		0.084	0.194	0.367	0.558	0.075	0.169	0.354	0.538
		T_3(Logit)		0.070	0.166	0.331	0.514	0.063	0.145	0.318	0.500
	5	SRS		0.135	0.274	0.480	0.690	0.133	0.269	0.482	0.679
		MW		0.070	0.163	0.306	0.479	0.077	0.171	0.324	0.520
		T_1		0.180	0.449	0.733	0.916	0.184	0.428	0.727	0.900
		T_2		0.177	0.432	0.714	0.901	0.177	0.411	0.705	0.887
		T_3(Fisher)		0.160	0.388	0.663	0.865	0.155	0.366	0.650	0.850
		T_3(Liptak)		0.178	0.433	0.713	0.901	0.170	0.401	0.701	0.886
		T_3(Logit)		0.173	0.424	0.700	0.896	0.168	0.396	0.688	0.880
	10	SRS		0.188	0.450	0.734	0.927	0.195	0.452	0.751	0.918
		MW		0.119	0.306	0.570	0.809	0.114	0.324	0.603	0.849
		T_1		0.297	0.702	0.949	0.997	0.284	0.682	0.939	0.996
		T_2		0.294	0.702	0.945	0.996	0.283	0.679	0.937	0.994
		T_3(Fisher)		0.268	0.659	0.926	0.993	0.257	0.639	0.915	0.992
		T_3(Liptak)		0.299	0.704	0.945	0.996	0.284	0.683	0.936	0.994
		T_3(Logit)		0.292	0.694	0.942	0.996	0.281	0.676	0.933	0.994
4	3	SRS		0.116	0.246	0.386	0.582	0.117	0.235	0.426	0.599
		MW		0.067	0.132	0.247	0.393	0.070	0.137	0.262	0.419
		T_1		0.217	0.528	0.812	0.950	0.160	0.392	0.657	0.847
		T_2		0.227	0.524	0.805	0.946	0.156	0.384	0.648	0.836
		T_3(Fisher)		0.213	0.500	0.783	0.936	0.144	0.364	0.616	0.810
		T_3(Liptak)		0.225	0.531	0.810	0.947	0.156	0.386	0.648	0.842
		T_3(Logit)		0.229	0.525	0.805	0.943	0.154	0.382	0.641	0.835
	5	SRS		0.144	0.338	0.580	0.805	0.153	0.348	0.594	0.791
		MW		0.093	0.218	0.407	0.626	0.088	0.229	0.436	0.673
		T_1		0.296	0.705	0.944	0.996	0.235	0.557	0.856	0.971
		T_2		0.297	0.705	0.944	0.995	0.231	0.553	0.848	0.971
		T_3(Fisher)		0.277	0.678	0.936	0.994	0.220	0.524	0.826	0.965
		T_3(Liptak)		0.298	0.702	0.944	0.996	0.233	0.553	0.853	0.971
		T_3(Logit)		0.295	0.698	0.942	0.996	0.231	0.549	0.846	0.970
	10	SRS		0.223	0.551	0.843	0.975	0.220	0.542	0.839	0.970
		MW		0.139	0.391	0.702	0.910	0.141	0.402	0.740	0.929
		T_1		0.473	0.930	0.999	1.000	0.366	0.811	0.986	1.000
		T_2		0.477	0.931	0.999	1.000	0.365	0.809	0.985	0.999
		T_3(Fisher)		0.447	0.913	0.998	1.000	0.350	0.787	0.980	1.000
		T_3(Liptak)		0.477	0.929	0.999	1.000	0.365	0.810	0.986	0.999
		T_3(Logit)		0.473	0.927	0.999	1.000	0.367	0.803	0.984	0.999

Table 4: Empirical power levels from the simulation study, the nominal significance level is $\alpha = 0.05$ – exponential and gamma distributions

				Exponential				Gamma			
m	h	Test Stat	$\delta \rightarrow$	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
3	3	SRS		0.122	0.252	0.400	0.563	0.109	0.225	0.358	0.527
		MW		0.079	0.194	0.346	0.514	0.059	0.118	0.219	0.361
		T_1		0.149	0.318	0.539	0.713	0.145	0.302	0.501	0.714
		T_2		0.196	0.462	0.739	0.892	0.143	0.311	0.515	0.711
		T_3(Fisher)		0.068	0.218	0.426	0.613	0.039	0.130	0.264	0.453
		T_3(Liptak)		0.120	0.305	0.541	0.707	0.077	0.215	0.388	0.602
		T_3(Logit)		0.103	0.276	0.507	0.685	0.063	0.190	0.346	0.561
	5	SRS		0.163	0.311	0.528	0.706	0.145	0.289	0.496	0.706
		MW		0.121	0.336	0.570	0.767	0.091	0.193	0.374	0.592
		T_1		0.191	0.442	0.692	0.861	0.187	0.442	0.705	0.894
		T_2		0.296	0.683	0.905	0.981	0.205	0.482	0.753	0.931
		T_3(Fisher)		0.281	0.662	0.890	0.976	0.174	0.435	0.708	0.906
		T_3(Liptak)		0.288	0.647	0.879	0.967	0.196	0.468	0.746	0.925
		T_3(Logit)		0.288	0.659	0.894	0.974	0.191	0.462	0.738	0.925
	10	SRS		0.207	0.488	0.751	0.916	0.191	0.455	0.758	0.916
		MW		0.222	0.605	0.872	0.972	0.133	0.370	0.679	0.897
		T_1		0.279	0.654	0.895	0.979	0.278	0.672	0.928	0.990
		T_2		0.446	0.889	0.992	1.000	0.313	0.740	0.966	0.997
		T_3(Fisher)		0.450	0.895	0.993	1.000	0.292	0.707	0.955	0.996
		T_3(Liptak)		0.449	0.885	0.992	1.000	0.314	0.738	0.963	0.997
		T_3(Logit)		0.455	0.897	0.993	1.000	0.312	0.736	0.964	0.997
4	3	SRS		0.140	0.281	0.465	0.647	0.121	0.249	0.434	0.610
		MW		0.108	0.262	0.481	0.660	0.073	0.156	0.302	0.497
		T_1		0.153	0.351	0.553	0.736	0.152	0.379	0.619	0.820
		T_2		0.357	0.767	0.946	0.986	0.188	0.457	0.739	0.915
		T_3(Fisher)		0.390	0.791	0.954	0.989	0.182	0.448	0.739	0.910
		T_3(Liptak)		0.354	0.725	0.898	0.960	0.191	0.451	0.731	0.904
		T_3(Logit)		0.372	0.774	0.937	0.983	0.190	0.457	0.742	0.914
	5	SRS		0.176	0.376	0.616	0.800	0.155	0.349	0.598	0.801
		MW		0.161	0.442	0.711	0.879	0.103	0.261	0.504	0.748
		T_1		0.207	0.459	0.728	0.888	0.213	0.517	0.797	0.945
		T_2		0.482	0.905	0.990	1.000	0.260	0.637	0.906	0.989
		T_3(Fisher)		0.522	0.933	0.995	1.000	0.253	0.634	0.911	0.991
		T_3(Liptak)		0.488	0.900	0.989	1.000	0.260	0.633	0.900	0.987
		T_3(Logit)		0.508	0.920	0.992	1.000	0.260	0.640	0.913	0.991
	10	SRS		0.246	0.569	0.844	0.968	0.233	0.552	0.842	0.970
		MW		0.291	0.732	0.951	0.994	0.161	0.474	0.807	0.960
		T_1		0.297	0.672	0.926	0.987	0.325	0.762	0.966	0.998
		T_2		0.718	0.988	1.000	1.000	0.423	0.894	0.996	1.000
		T_3(Fisher)		0.758	0.995	1.000	1.000	0.419	0.897	0.996	1.000
		T_3(Liptak)		0.721	0.989	1.000	1.000	0.421	0.893	0.996	1.000
		T_3(Logit)		0.749	0.994	1.000	1.000	0.428	0.903	0.997	1.000

Table 5: Empirical power levels from the simulation study, the nominal significance level is $\alpha = 0.10$ – Uniform and normal distributions

				Uniform				Normal			
m	h	Test Stat	$\delta \rightarrow$	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
3	3	SRS		0.191	0.317	0.472	0.650	0.195	0.331	0.477	0.650
		MW		0.117	0.190	0.291	0.440	0.126	0.198	0.322	0.472
		T_1		0.244	0.447	0.678	0.858	0.246	0.456	0.660	0.840
		T_2		0.224	0.420	0.634	0.828	0.228	0.422	0.619	0.795
		T_3(Fisher)		0.117	0.256	0.459	0.656	0.117	0.244	0.424	0.621
		T_3(Liptak)		0.162	0.341	0.551	0.745	0.162	0.323	0.524	0.715
		T_3(Logit)		0.148	0.318	0.530	0.728	0.145	0.305	0.495	0.697
	5	SRS		0.220	0.411	0.626	0.810	0.218	0.399	0.634	0.806
		MW		0.151	0.262	0.443	0.632	0.133	0.269	0.472	0.659
		T_1		0.302	0.604	0.848	0.962	0.298	0.571	0.812	0.956
		T_2		0.291	0.590	0.837	0.954	0.297	0.562	0.802	0.946
		T_3(Fisher)		0.272	0.549	0.804	0.936	0.268	0.507	0.761	0.922
		T_3(Liptak)		0.285	0.593	0.838	0.951	0.286	0.555	0.797	0.940
		T_3(Logit)		0.278	0.587	0.830	0.951	0.285	0.548	0.789	0.938
	10	SRS		0.303	0.595	0.851	0.971	0.306	0.591	0.861	0.966
		MW		0.196	0.439	0.702	0.890	0.186	0.442	0.739	0.909
		T_1		0.426	0.804	0.976	0.999	0.416	0.807	0.973	0.996
		T_2		0.421	0.799	0.974	0.999	0.416	0.806	0.971	0.997
		T_3(Fisher)		0.401	0.773	0.966	0.998	0.392	0.775	0.963	0.996
		T_3(Liptak)		0.426	0.801	0.975	0.999	0.419	0.808	0.973	0.997
		T_3(Logit)		0.420	0.799	0.973	0.998	0.412	0.805	0.973	0.997
4	3	SRS		0.215	0.371	0.567	0.721	0.206	0.382	0.570	0.730
		MW		0.128	0.218	0.351	0.537	0.122	0.236	0.380	0.537
		T_1		0.336	0.677	0.894	0.982	0.278	0.535	0.766	0.932
		T_2		0.341	0.678	0.889	0.981	0.280	0.527	0.758	0.928
		T_3(Fisher)		0.329	0.657	0.868	0.974	0.262	0.506	0.731	0.917
		T_3(Liptak)		0.343	0.681	0.891	0.981	0.280	0.531	0.757	0.928
		T_3(Logit)		0.340	0.677	0.883	0.979	0.280	0.526	0.756	0.925
	5	SRS		0.246	0.486	0.702	0.893	0.263	0.481	0.729	0.884
		MW		0.161	0.320	0.537	0.734	0.156	0.339	0.560	0.783
		T_1		0.436	0.833	0.981	0.999	0.349	0.704	0.915	0.988
		T_2		0.433	0.832	0.980	0.999	0.341	0.695	0.913	0.987
		T_3(Fisher)		0.418	0.808	0.976	0.998	0.328	0.671	0.904	0.983
		T_3(Liptak)		0.435	0.834	0.981	0.999	0.343	0.695	0.912	0.987
		T_3(Logit)		0.428	0.830	0.980	0.999	0.342	0.687	0.913	0.986
	10	SRS		0.346	0.695	0.914	0.993	0.351	0.696	0.915	0.990
		MW		0.219	0.515	0.804	0.946	0.220	0.535	0.825	0.960
		T_1		0.608	0.976	1.000	1.000	0.502	0.894	0.994	1.000
		T_2		0.605	0.974	1.000	1.000	0.503	0.897	0.993	1.000
		T_3(Fisher)		0.586	0.965	1.000	1.000	0.484	0.883	0.991	1.000
		T_3(Liptak)		0.605	0.973	1.000	1.000	0.503	0.894	0.994	1.000
		T_3(Logit)		0.600	0.972	1.000	1.000	0.499	0.894	0.993	1.000

Table 6: Empirical power levels from the simulation study, the nominal significance level is $\alpha = 0.10$ – exponential and gamma distributions

				Exponential				Gamma			
m	h	Test Stat	$\delta \rightarrow$	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
3	3	SRS		0.205	0.368	0.531	0.677	0.194	0.328	0.514	0.641
		MW		0.164	0.329	0.501	0.656	0.129	0.216	0.355	0.517
		T_1		0.259	0.460	0.651	0.802	0.236	0.453	0.673	0.834
		T_2		0.352	0.642	0.844	0.951	0.242	0.455	0.695	0.858
		T_3(Fisher)		0.200	0.431	0.675	0.825	0.127	0.276	0.496	0.711
		T_3(Liptak)		0.251	0.496	0.711	0.850	0.173	0.357	0.592	0.779
		T_3(Logit)		0.233	0.480	0.702	0.843	0.161	0.342	0.572	0.763
	5	SRS		0.240	0.438	0.670	0.830	0.229	0.447	0.655	0.810
		MW		0.210	0.463	0.709	0.854	0.157	0.295	0.524	0.735
		T_1		0.302	0.574	0.801	0.922	0.296	0.573	0.813	0.950
		T_2		0.422	0.792	0.963	0.994	0.310	0.614	0.857	0.972
		T_3(Fisher)		0.425	0.791	0.962	0.992	0.286	0.577	0.825	0.965
		T_3(Liptak)		0.417	0.773	0.951	0.987	0.303	0.606	0.852	0.970
		T_3(Logit)		0.422	0.783	0.955	0.990	0.298	0.599	0.848	0.970
	10	SRS		0.327	0.604	0.838	0.963	0.328	0.598	0.847	0.963
		MW		0.348	0.715	0.929	0.987	0.209	0.489	0.791	0.939
		T_1		0.414	0.764	0.945	0.990	0.415	0.788	0.960	0.997
		T_2		0.573	0.942	0.997	1.000	0.453	0.845	0.981	0.999
		T_3(Fisher)		0.581	0.943	0.997	1.000	0.432	0.829	0.976	0.998
		T_3(Liptak)		0.576	0.941	0.997	1.000	0.455	0.847	0.981	0.999
		T_3(Logit)		0.587	0.944	0.998	1.000	0.452	0.846	0.981	0.999
4	3	SRS		0.224	0.404	0.576	0.770	0.220	0.379	0.582	0.761
		MW		0.197	0.382	0.603	0.767	0.127	0.246	0.424	0.622
		T_1		0.267	0.475	0.683	0.823	0.266	0.525	0.757	0.900
		T_2		0.517	0.866	0.971	0.995	0.317	0.617	0.864	0.962
		T_3(Fisher)		0.537	0.886	0.979	0.996	0.300	0.607	0.863	0.962
		T_3(Liptak)		0.512	0.842	0.956	0.987	0.316	0.616	0.854	0.952
		T_3(Logit)		0.525	0.869	0.969	0.994	0.316	0.616	0.860	0.959
	5	SRS		0.266	0.503	0.748	0.885	0.264	0.482	0.735	0.884
		MW		0.244	0.569	0.803	0.934	0.172	0.361	0.640	0.829
		T_1		0.304	0.590	0.812	0.935	0.341	0.656	0.881	0.975
		T_2		0.632	0.951	0.997	1.000	0.410	0.770	0.953	0.997
		T_3(Fisher)		0.665	0.965	0.999	1.000	0.398	0.770	0.956	0.997
		T_3(Liptak)		0.633	0.948	0.996	1.000	0.408	0.765	0.951	0.996
		T_3(Logit)		0.656	0.955	0.998	1.000	0.411	0.771	0.955	0.996
	10	SRS		0.349	0.711	0.919	0.987	0.342	0.681	0.914	0.986
		MW		0.407	0.830	0.969	0.998	0.251	0.601	0.877	0.981
		T_1		0.428	0.783	0.963	0.995	0.478	0.860	0.987	1.000
		T_2		0.823	0.996	1.000	1.000	0.580	0.947	0.999	1.000
		T_3(Fisher)		0.863	0.999	1.000	1.000	0.566	0.950	0.999	1.000
		T_3(Liptak)		0.824	0.996	1.000	1.000	0.578	0.946	0.999	1.000
		T_3(Logit)		0.847	0.998	1.000	1.000	0.577	0.951	0.999	1.000

Table 7: Empirical power levels from the simulation study, the nominal significance level is $\alpha = 0.40$ – Uniform and normal distributions

				Uniform				Normal			
m	h	Test Stat	$\delta \rightarrow$	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
3	3	SRS		0.557	0.726	0.837	0.925	0.550	0.727	0.849	0.922
		MW		0.438	0.527	0.635	0.766	0.420	0.530	0.654	0.774
		T_1		0.646	0.833	0.935	0.984	0.627	0.819	0.931	0.982
		T_2		0.637	0.825	0.924	0.983	0.628	0.809	0.919	0.978
		T_3(Fisher)		0.568	0.772	0.894	0.969	0.554	0.735	0.882	0.961
		T_3(Liptak)		0.572	0.775	0.895	0.970	0.561	0.752	0.885	0.962
		T_3(Logit)		0.570	0.773	0.894	0.968	0.555	0.748	0.882	0.962
	5	SRS		0.621	0.796	0.919	0.974	0.612	0.788	0.916	0.971
		MW		0.451	0.596	0.759	0.883	0.441	0.592	0.766	0.887
		T_1		0.703	0.907	0.981	1.000	0.704	0.899	0.977	0.998
		T_2		0.699	0.899	0.978	1.000	0.705	0.896	0.975	0.997
		T_3(Fisher)		0.674	0.891	0.976	0.999	0.685	0.880	0.967	0.995
		T_3(Liptak)		0.693	0.900	0.979	0.999	0.699	0.893	0.974	0.997
		T_3(Logit)		0.691	0.899	0.979	0.999	0.698	0.893	0.974	0.997
	10	SRS		0.697	0.904	0.979	0.998	0.698	0.898	0.982	0.998
		MW		0.518	0.741	0.910	0.978	0.525	0.755	0.919	0.984
		T_1		0.793	0.973	0.999	1.000	0.806	0.967	0.998	1.000
		T_2		0.792	0.971	0.999	1.000	0.808	0.967	0.998	1.000
		T_3(Fisher)		0.780	0.965	0.998	1.000	0.783	0.963	0.997	1.000
		T_3(Liptak)		0.792	0.970	0.999	1.000	0.806	0.967	0.998	1.000
		T_3(Logit)		0.790	0.970	0.999	1.000	0.807	0.967	0.998	1.000
4	3	SRS		0.605	0.765	0.878	0.946	0.586	0.586	0.901	0.956
		MW		0.428	0.557	0.702	0.825	0.429	0.560	0.705	0.841
		T_1		0.725	0.929	0.986	1.000	0.670	0.670	0.969	0.994
		T_2		0.732	0.932	0.986	1.000	0.672	0.672	0.968	0.992
		T_3(Fisher)		0.721	0.923	0.983	0.999	0.661	0.661	0.960	0.992
		T_3(Liptak)		0.733	0.933	0.987	1.000	0.674	0.674	0.968	0.992
		T_3(Logit)		0.733	0.933	0.987	1.000	0.672	0.672	0.968	0.993
	5	SRS		0.634	0.839	0.950	0.990	0.641	0.848	0.954	0.989
		MW		0.491	0.667	0.818	0.930	0.484	0.651	0.857	0.943
		T_1		0.802	0.979	0.999	1.000	0.733	0.948	0.992	1.000
		T_2		0.807	0.979	0.999	1.000	0.729	0.944	0.992	1.000
		T_3(Fisher)		0.786	0.974	0.998	1.000	0.726	0.941	0.990	0.999
		T_3(Liptak)		0.806	0.979	0.999	1.000	0.729	0.944	0.992	1.000
		T_3(Logit)		0.806	0.978	0.999	1.000	0.728	0.943	0.992	1.000
	10	SRS		0.733	0.942	0.993	1.000	0.747	0.945	0.991	1.000
		MW		0.555	0.808	0.949	0.992	0.555	0.826	0.957	0.996
		T_1		0.903	0.998	1.000	1.000	0.848	0.991	1.000	1.000
		T_2		0.906	0.998	1.000	1.000	0.848	0.991	1.000	1.000
		T_3(Fisher)		0.896	0.997	1.000	1.000	0.837	0.987	1.000	1.000
		T_3(Liptak)		0.904	0.998	1.000	1.000	0.850	0.990	1.000	1.000
		T_3(Logit)		0.904	0.998	1.000	1.000	0.849	0.990	1.000	1.000

Table 8: Empirical power levels from the simulation study, the nominal significance level is $\alpha = 0.40$ – exponential and gamma distributions

			Exponential				Gamma			
m	h	Test Stat $\delta \rightarrow$	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
3	3	SRS	0.583	0.740	0.847	0.915	0.562	0.726	0.849	0.921
		MW	0.486	0.655	0.799	0.899	0.430	0.540	0.688	0.820
		T_1	0.638	0.791	0.899	0.963	0.624	0.822	0.933	0.976
		T_2	0.746	0.917	0.981	0.996	0.645	0.848	0.945	0.987
		T_3(Fisher)	0.695	0.893	0.971	0.994	0.567	0.793	0.920	0.979
		T_3(Liptak)	0.689	0.872	0.958	0.984	0.567	0.787	0.922	0.975
		T_3(Logit)	0.687	0.871	0.958	0.984	0.564	0.786	0.922	0.975
	5	SRS	0.613	0.801	0.919	0.973	0.612	0.807	0.921	0.974
		MW	0.537	0.770	0.913	0.971	0.468	0.634	0.822	0.925
		T_1	0.673	0.877	0.969	0.989	0.688	0.891	0.976	0.995
		T_2	0.813	0.967	0.997	1.000	0.707	0.921	0.983	0.998
		T_3(Fisher)	0.807	0.964	0.996	1.000	0.692	0.911	0.982	0.997
		T_3(Liptak)	0.811	0.965	0.997	1.000	0.697	0.919	0.983	0.997
		T_3(Logit)	0.814	0.966	0.997	1.000	0.698	0.918	0.984	0.997
	10	SRS	0.706	0.905	0.976	0.996	0.708	0.904	0.982	0.998
		MW	0.669	0.911	0.987	0.998	0.531	0.808	0.950	0.990
		T_1	0.771	0.949	0.996	1.000	0.789	0.966	0.997	1.000
		T_2	0.892	0.994	1.000	1.000	0.817	0.982	0.999	1.000
		T_3(Fisher)	0.884	0.996	1.000	1.000	0.803	0.976	0.998	1.000
		T_3(Liptak)	0.891	0.994	1.000	1.000	0.817	0.982	0.999	1.000
		T_3(Logit)	0.892	0.995	1.000	1.000	0.816	0.981	0.999	1.000
4	3	SRS	0.595	0.760	0.869	0.953	0.604	0.767	0.898	0.961
		MW	0.493	0.719	0.865	0.940	0.440	0.578	0.765	0.878
		T_1	0.637	0.815	0.918	0.971	0.667	0.846	0.962	0.989
		T_2	0.848	0.979	0.997	0.999	0.724	0.901	0.988	0.998
		T_3(Fisher)	0.858	0.983	0.998	1.000	0.713	0.902	0.988	0.998
		T_3(Liptak)	0.851	0.980	0.997	0.999	0.721	0.900	0.989	0.997
		T_3(Logit)	0.852	0.980	0.997	0.999	0.721	0.900	0.989	0.998
	5	SRS	0.648	0.846	0.953	0.982	0.647	0.848	0.955	0.990
		MW	0.596	0.842	0.956	0.990	0.497	0.703	0.881	0.961
		T_1	0.685	0.889	0.973	0.994	0.727	0.922	0.988	0.998
		T_2	0.892	0.997	1.000	1.000	0.795	0.963	0.997	1.000
		T_3(Fisher)	0.906	0.998	1.000	1.000	0.795	0.968	0.996	1.000
		T_3(Liptak)	0.893	0.996	1.000	1.000	0.794	0.962	0.997	1.000
		T_3(Logit)	0.894	0.996	1.000	1.000	0.795	0.963	0.997	1.000
	10	SRS	0.725	0.934	0.990	0.999	0.734	0.941	0.992	0.999
		MW	0.726	0.956	0.997	1.000	0.593	0.858	0.974	0.998
		T_1	0.786	0.963	0.999	0.999	0.835	0.983	0.999	1.000
		T_2	0.973	1.000	1.000	1.000	0.894	0.996	1.000	1.000
		T_3(Fisher)	0.979	1.000	1.000	1.000	0.892	0.996	1.000	1.000
		T_3(Liptak)	0.971	1.000	1.000	1.000	0.893	0.996	1.000	1.000
		T_3(Logit)	0.972	1.000	1.000	1.000	0.895	0.996	1.000	1.000

Table 9: The power levels of the parametric one-sided two-sample t-test. The nominal significance level is $\alpha = 0.05$.

	δ			
$n = m \times h$	0.2	0.4	0.6	0.8
9	0.108	0.203	0.335	0.492
12	0.121	0.243	0.413	0.600
15	0.133	0.282	0.483	0.689
20	0.153	0.344	0.587	0.799
30	0.190	0.455	0.743	0.922
40	0.224	0.551	0.845	0.971

Table 10: The power levels of the parametric one-sided two-sample t-test. The nominal significance level is $\alpha = 0.10$.

	δ			
$n = m \times h$	0.2	0.4	0.6	0.8
9	0.193	0.325	0.483	0.645
12	0.212	0.374	0.564	0.739
15	0.229	0.420	0.632	0.810
20	0.256	0.488	0.724	0.889
30	0.304	0.601	0.847	0.964
40	0.348	0.691	0.917	0.989

Table 11: The power levels of the parametric one-sided two-sample t-test. The nominal significance level is $\alpha = 0.40$.

	δ			
$n = m \times h$	0.2	0.4	0.6	0.8
9	0.568	0.724	0.846	0.925
12	0.593	0.766	0.888	0.956
15	0.616	0.800	0.918	0.974
20	0.648	0.844	0.950	0.989
30	0.699	0.902	0.981	0.998
40	0.739	0.938	0.992	1.000

Table 12: Empirical coverage probability (expected length between parentheses) of 90% confidence intervals – Uniform and normal distributions

			Uniform		Normal	
m	h	Test Stat $\delta \rightarrow$	0.0	0.6	0.0	0.6
3	3	SRS	0.986 (2.605)	0.984 (2.613)	0.992 (2.562)	0.985 (2.594)
		MW	0.905 (1.789)	0.886 (1.779)	0.903 (1.694)	0.899 (1.688)
		T_1	0.993 (2.025)	0.992 (2.009)	0.988 (2.040)	0.997 (2.045)
		T_2	0.979 (1.877)	0.978 (1.875)	0.973 (1.891)	0.983 (1.895)
		T_3(Fisher)	0.996 (2.361)	0.999 (2.379)	0.994 (2.426)	0.995 (2.397)
		T_3(Liptak)	0.997 (2.096)	0.996 (2.080)	0.997 (2.125)	0.993 (2.140)
		T_3(Logit)	0.998 (2.179)	0.998 (2.180)	0.994 (2.236)	0.995 (2.201)
	5	SRS	0.993 (2.094)	0.995 (2.097)	0.995 (2.092)	0.993 (2.103)
		MW	0.898 (1.343)	0.902 (1.333)	0.895 (1.278)	0.903 (1.281)
		T_1	0.999 (1.649)	0.998 (1.658)	0.998 (1.688)	0.998 (1.694)
		T_2	0.992 (1.508)	0.992 (1.515)	0.993 (1.542)	0.988 (1.533)
		T_3(Fisher)	0.993 (1.572)	0.993 (1.561)	0.992 (1.591)	0.996 (1.591)
		T_3(Liptak)	0.988 (1.516)	0.995 (1.514)	0.993 (1.540)	0.990 (1.534)
		T_3(Logit)	0.994 (1.522)	0.989 (1.517)	0.989 (1.542)	0.993 (1.533)
	10	SRS	0.999 (1.636)	0.999 (1.635)	0.996 (1.641)	0.998 (1.642)
		MW	0.892 (0.915)	0.904 (0.912)	0.912 (0.882)	0.904 (0.889)
		T_1	0.999 (1.267)	1.000 (1.267)	0.999 (1.286)	1.000 (1.280)
		T_2	0.998 (1.227)	0.996 (1.221)	0.999 (1.227)	0.999 (1.232)
		T_3(Fisher)	0.998 (1.233)	0.996 (1.224)	0.996 (1.238)	0.998 (1.238)
		T_3(Liptak)	0.998 (1.221)	0.998 (1.221)	0.998 (1.229)	0.999 (1.225)
		T_3(Logit)	0.999 (1.218)	0.998 (1.218)	0.999 (1.227)	0.998 (1.225)
4	3	SRS	0.992 (2.312)	0.988 (2.293)	0.989 (2.289)	0.990 (2.281)
		MW	0.916 (1.545)	0.921 (1.553)	0.903 (1.481)	0.909 (1.462)
		T_1	0.998 (1.551)	0.998 (1.556)	0.998 (1.785)	0.998 (1.777)
		T_2	0.995 (1.386)	0.994 (1.397)	0.993 (1.589)	0.996 (1.596)
		T_3(Fisher)	0.992 (1.418)	0.992 (1.401)	0.986 (1.658)	0.990 (1.631)
		T_3(Liptak)	0.994 (1.379)	0.995 (1.396)	0.994 (1.580)	0.993 (1.591)
		T_3(Logit)	0.986 (1.382)	0.995 (1.386)	0.990 (1.592)	0.992 (1.614)
	5	SRS	0.995 (1.912)	0.993 (1.914)	0.997 (1.887)	0.993 (1.883)
		MW	0.903 (1.156)	0.923 (1.140)	0.900 (1.103)	0.913 (1.109)
		T_1	0.999 (1.284)	0.999 (1.287)	0.998 (1.523)	0.998 (1.519)
		T_2	1.000 (1.223)	0.997 (1.219)	0.994 (1.318)	0.997 (1.332)
		T_3(Fisher)	0.999 (1.211)	1.000 (1.213)	0.998 (1.383)	0.992 (1.375)
		T_3(Liptak)	1.000 (1.216)	0.997 (1.219)	0.998 (1.310)	0.994 (1.313)
		T_3(Logit)	0.999 (1.219)	0.998 (1.215)	0.996 (1.329)	0.995 (1.339)
	10	SRS	1.000 (1.586)	0.997 (1.586)	0.998 (1.560)	0.998 (1.552)
		MW	0.892 (0.791)	0.900 (0.788)	0.896 (0.766)	0.898 (0.765)
		T_1	1.000 (1.212)	1.000 (1.213)	0.999 (1.232)	1.000 (1.232)
		T_2	1.000 (1.017)	1.000 (1.022)	0.999 (1.217)	1.000 (1.218)
		T_3(Fisher)	1.000 (1.091)	0.999 (1.073)	1.000 (1.190)	1.000 (1.183)
		T_3(Liptak)	1.000 (1.007)	1.000 (1.008)	0.999 (1.222)	0.999 (1.216)
		T_3(Logit)	0.998 (1.027)	1.000 (1.026)	1.000 (1.205)	0.998 (1.202)

Table 13: Empirical coverage probability (expected length between parentheses) of 90% confidence intervals – Exponential and gamma distributions

			Exponential		Gamma	
m	h	Test Stat $\delta \rightarrow$	0.0	0.6	0.0	0.6
3	3	SRS	0.997 (2.485)	0.992 (2.470)	0.987 (2.559)	0.988 (2.536)
		MW	0.902 (1.348)	0.894 (1.337)	0.904 (1.605)	0.915 (1.597)
		T_1	0.997 (2.057)	0.994 (2.073)	0.997 (2.036)	0.993 (2.054)
		T_2	0.912 (1.475)	0.899 (1.454)	0.969 (1.787)	0.979 (1.807)
		T_3(Fisher)	0.983 (2.036)	0.975 (2.054)	0.997 (2.347)	0.994 (2.299)
		T_3(Liptak)	0.983 (1.841)	0.997 (1.872)	0.994 (2.076)	0.990 (2.054)
		T_3(Logit)	0.984 (1.909)	0.989 (1.910)	0.993 (2.135)	0.992 (2.127)
	5	SRS	0.992 (2.036)	0.996 (2.024)	0.997 (2.080)	0.992 (2.060)
		MW	0.919 (0.907)	0.898 (0.915)	0.904 (1.175)	0.910 (1.187)
		T_1	0.997 (1.709)	0.998 (1.723)	0.999 (1.700)	0.996 (1.695)
		T_2	0.916 (1.238)	0.931 (1.236)	0.977 (1.448)	0.983 (1.449)
		T_3(Fisher)	0.916 (1.264)	0.923 (1.270)	0.983 (1.516)	0.974 (1.510)
		T_3(Liptak)	0.947 (1.274)	0.959 (1.283)	0.988 (1.462)	0.992 (1.462)
		T_3(Logit)	0.937 (1.255)	0.951 (1.283)	0.990 (1.475)	0.989 (1.475)
	10	SRS	0.998 (1.620)	1.000 (1.628)	1.000 (1.643)	0.995 (1.643)
		MW	0.891 (0.585)	0.895 (0.592)	0.886 (0.808)	0.889 (0.801)
		T_1	1.000 (1.375)	0.999 (1.372)	1.000 (1.328)	0.999 (1.324)
		T_2	0.959 (1.055)	0.961 (1.045)	0.994 (1.197)	0.993 (1.196)
		T_3(Fisher)	0.926 (1.030)	0.944 (1.046)	0.990 (1.185)	0.984 (1.194)
		T_3(Liptak)	0.957 (1.059)	0.961 (1.047)	0.995 (1.200)	0.997 (1.202)
		T_3(Logit)	0.947 (1.053)	0.959 (1.045)	0.994 (1.193)	0.992 (1.191)
4	3	SRS	0.992 (2.221)	0.996 (2.219)	0.991 (2.254)	0.996 (2.266)
		MW	0.906 (1.095)	0.891 (1.104)	0.913 (1.366)	0.915 (1.382)
		T_1	0.999 (1.978)	0.998 (1.984)	0.995 (1.846)	0.993 (1.836)
		T_2	0.833 (1.171)	0.866 (1.199)	0.955 (1.459)	0.959 (1.462)
		T_3(Fisher)	0.824 (1.144)	0.807 (1.145)	0.945 (1.479)	0.954 (1.477)
		T_3(Liptak)	0.891 (1.227)	0.880 (1.204)	0.979 (1.460)	0.975 (1.479)
		T_3(Logit)	0.837 (1.166)	0.868 (1.205)	0.959 (1.489)	0.950 (1.456)
	5	SRS	1.000 (1.840)	0.996 (1.836)	0.994 (1.859)	0.992 (1.874)
		MW	0.893 (0.779)	0.906 (0.758)	0.904 (1.010)	0.916 (1.020)
		T_1	0.999 (1.663)	0.999 (1.656)	0.997 (1.565)	0.998 (1.560)
		T_2	0.869 (1.039)	0.886 (1.053)	0.975 (1.234)	0.978 (1.262)
		T_3(Fisher)	0.829 (1.030)	0.851 (1.033)	0.954 (1.238)	0.954 (1.255)
		T_3(Liptak)	0.882 (1.035)	0.891 (1.051)	0.971 (1.250)	0.980 (1.243)
		T_3(Logit)	0.863 (1.053)	0.851 (1.035)	0.960 (1.231)	0.968 (1.236)
	10	SRS	0.998 (1.486)	0.999 (1.488)	0.999 (1.519)	0.996 (1.514)
		MW	0.901 (0.495)	0.912 (0.499)	0.897 (0.696)	0.912 (0.693)
		T_1	0.999 (1.333)	0.998 (1.332)	1.000 (1.244)	1.000 (1.244)
		T_2	0.895 (0.900)	0.895 (0.919)	0.995 (1.073)	0.991 (1.072)
		T_3(Fisher)	0.882 (0.909)	0.890 (0.887)	0.979 (1.076)	0.979 (1.068)
		T_3(Liptak)	0.908 (0.915)	0.905 (0.908)	0.991 (1.065)	0.989 (1.068)
		T_3(Logit)	0.909 (0.911)	0.890 (0.895)	0.986 (1.064)	0.987 (1.078)

7. Data application

Height is an important factor in basketball and football. The heights of basketball players are in the range of 184-210 cm, and the heights of football players are in the range of 170-188 cm. Talent players could have height outside these ranges. So, in general, the average height of basketballers is slightly greater than that for footballers. To examine this among Palestinian athletes, a SRS of 15 footballers and 15 basketballers are selected from an athletics club in Palestine. Moreover, an ERSS of 15 footballers and 15 basketballers ($m = 3$, and $h = 5$) are selected in the following way. A group of three players is randomly selected, ranked according to their heights, and the height of the shortest player is measured. A second group of three players is randomly selected, ranked according to their heights, and the height of the tallest player is measured. Finally, a third group is randomly selected, ranked according to their heights, and the height of the middle player is measured. This is a single cycle of ERSS of size $m = 3$. This process is repeated 5 times ($h = 5$), for both footballers and basketballers, to acquire an ERSS of size $m \times h = 15$. The data are shown in Table 14.

Table 14: The height of 15 footballers and 15 basketballers chosen from an athletics club in Palestine by SRS and ERSS

	Basketballers					Footballers				
	185	183	195	201	211	175	191	183	185	173
SRS	185	191	198	201	191	203	193	180	191	206
	203	201	213	208	196	193	185	196	201	183
	191	201	191	196	203	175	196	193	160	198
ERSS	198	185	201	183	175	196	180	201	193	188
	201	206	195	213	206	203	191	173	198	178

To construct a 95% confidence interval for $\delta = \mu_B - \mu_F$, the true difference between the two population means, we set $\epsilon = 0.001$ and $B = 1000$. The results are reported in Table 15. It can be seen that, for testing $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$, all p-values are significant at $\alpha = 0.05$. In addition, all confidence intervals are positive abd do not include zero; that is, the expected height of basketballers is greater than that for footballers.

Table 15: 95% confidence interval for $\delta = \mu_B - \mu_F$.

	observed test stat	p-value	CI
SRS	8.267	0.024	$(0.022, 16.757)$
MW	162.5	0.039	$(0.000, 10.600)$
T_1	8.133	0.012	$(0.854, 15.412)$
T_2	3.257	0.006	$(1.460, 14.807)$
T_3(Fisher)	13.2	0.015	$(1.258, 15.008)$
T_3(Liptak)	6.60	0.006	$(1.460, 14.807)$
T_3(Logit)	4.60	0.009	$(1.258, 15.008)$

8. Summary and Conclusion

Two-sample permutation test was previously investigated within the context of RSS and multistage RSS. It was shown to be highly efficient and applicable in the context of RSS. In this paper, we extend the applicability of the permutation test to ERSS. The concept of ERSS is first introduced. Then, we review the independent two-sample design, such as, the classical independent t-test, Mann-Whitney U test, and the two-sample permutation test. Further, three different permutation test statistics for two-sample ERSS are suggested. The first statistic is based on the difference between overall means of two groups. The second is based on the studentized statistic, while the third is based on partial tests. The third statistic yields 2 or 3 independent p-values, which are combined to a single p-value for testing hypothesis. The following methods are explored for combining the p-values; the Fisher approach, the Liptak approach, and the logistic approach. The suggested statistics are compared with the usual permutation test statistic, the Mann-Whitney U test and the classical t-test. These are done under different distributions, such as, uniform, normal, exponential, and gamma distributions. In summary, our simulation results assert that the power levels of the permutation test statistics using ERSS are higher than the power levels of the classical two sample independent t-test statistic and the permutation test statistic using SRS. In addition, we compute the confidence interval for the two-sample permutation problem under ERSS and observe that the length of the confidence interval are, on average, lesser than the one obtained under SRS of

equivalent sample size. The effect of sample size is considered and it is observed that the performance of the proposed statistics improves with increase in sample size. Real life application of this research is shown using an illustrative example. To this end, it is recommended to apply permutation test within the framework of ERSS in lieu of SRS. It is worthy of note that we assume perfect judgment ranking in the selection of the data points for the ERSS. A violation of this assumption is quite interesting and a subject of future work.

Acknowledgement

The authors are very thankful to the Associate Editor and the anonymous referees for the comments, which led to a considerable improvement of an earlier version of this paper. Moreover, the authors would like to acknowledge their universities (Palestine Polytechnic University, University of Manitoba) for giving them moral and technical support to carry out research work.

References

1. Abu-Dayyeh, W. A., Al-Subh, S. A., and Muttlak, H. A. (2004). Logistic parameters estimation using simple random sampling and ranked set sampling data. *Applied Mathematics and Computation*, 150(2):543–554.
2. Adatia, A. (2000). Estimation of parameters of the half-logistic distribution using generalized ranked set sampling. *Computational Statistics & Data Analysis*, 33(1):1–13.
3. Al-Saleh, M. F. and Al-Kadiri, M. A. (2000). Double-ranked set sampling. *Statistics & Probability Letters*, 48(2):205–212.
4. Al-Saleh, M. F. and Al-Omari, A. I. (2002). Multistage ranked set sampling. *Journal of Statistical Planning and Inference*, 102(2):273–286.
5. Al-Saleh, M. F. and Diab, Y. A. (2009). Estimation of the parameters of downton's bivariate exponential distribution using ranked set sampling scheme. *Journal of Statistical Planning and Inference*, 139(2):277–286.
6. Amro, L. and Samuh, M. H. (2017). More powerful permutation test based on multistage ranked set sampling. *Communications in Statistics - Simulation and Computation*, 46(7):5271–5284.
7. Bhoj, D. S. (1997). Estimation of parameters of the extreme value distribution using ranked set sampling. *Communications in Statistics-Theory and Methods*, 26(3):653–667.
8. Bhoj, D. S. and Ahsanullah, M. (1996). Estimation of parameters of the generalized geometric distribution using ranked set sampling. *Biometrics*, 52(2):685–694.
9. Bohn, L. L. and Wolfe, D. A. (1992). Nonparametric two-sample procedures for ranked-set samples data. *Journal of the American Statistical Association*, 87(418):552–561.
10. Bohn, L. L. and Wolfe, D. A. (1994). The effect of imperfect judgment rankings on properties of procedures based on the ranked-set samples analog of the Mann-Whitney-Wilcoxon statistic. *Journal of the American Statistical Association*, 89(425):168–176.
11. Chacko, M. and Thomas, P. Y. (2007). Estimation of a parameter of bivariate pareto distribution by ranked set sampling. *Journal of Applied Statistics*, 34(6):703–714.
12. Chacko, M. and Thomas, P. Y. (2008). Estimation of a parameter of morgenstern type bivariate exponential distribution by ranked set sampling. *Annals of the Institute of Statistical Mathematics*, 60(2):301–318.
13. Fisher, R. A. (1934). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.
14. Koti, K. M. and Jogesh Babu, G. (1996). Sign test for ranked-set sampling. *Communications in Statistics-Theory and Methods*, 25(7):1617–1630.
15. Lam, K., Sinha, B. K., and Wu, Z. (1994). Estimation of parameters in a two-parameter exponential distribution using ranked set sample. *Annals of the Institute of Statistical Mathematics*, 46(4):723–736.
16. Liangyong, Z. and Xiaofang, D. (2010). Optimal ranked set sampling design for the sign test. , 26(3):225–233.
17. Liptak, T. (1958). On the combination of independent tests. *Magyar Tud Akad Mat Kutato Int Kozl*, 3:171–197.
18. McIntyre, G. (1952). A method for unbiased selective sampling, using ranked sets. *Australian Journal of Agricultural Research*, 3(4):385–390.
19. McIntyre, G. (2005). A method for unbiased selective sampling, using ranked sets. *The American Statistician*, 59:230–232.

20. Mudholkar, G. S. and George, E. O. (1977). The logit statistic for combining probabilities-an overview. Technical report, DTIC Document.
21. Muttlak, H. (1997). Median ranked set sampling. *Journal of Applied Statistical Science*, 6:245–255.
22. Navarro, J., Rychlik, T., and Shaked, M. (2007). Are the order statistics ordered? a survey of recent results. *Communications in Statistics - Theory and Methods*, 36(7):1273–1290.
23. Öztürk, Ö. (1999). Two-sample inference based on one-sample ranked set sample sign statistics. *Journal of Nonparametric Statistics*, 10(2):197–212.
24. Öztürk, Ö. and Wolfe, D. A. (2000). Optimal allocation procedure in ranked set two-sample median test. *Journal of Nonparametric Statistics*, 13(1):57–76.
25. Paul, J. and Thomas, P. Y. (2017). Concomitant record ranked set sampling. *Communications in Statistics - Theory and Methods*, 46(19):9518–9540.
26. Pesarin, F. and Salmaso, L. (2010). *Permutation Tests for Complex Data: Theory, Application and Software*. John Wiley & Sons, Ltd., Chichester.
27. Samawi, H. M., Ahmed, M. S., and Abu-Dayyeh, W. (1996). Estimating the population mean using extreme ranked set sampling. *Biometrical Journal*, 38(5):577–586.
28. Samuh, M. H. (2012). *Some Advances in Permutation Testing*. PhD thesis, Department of Statistical Science, Padua University.
29. Samuh, M. H. (2017). Ranked set two-sample permutation test. *Statistica*, 77(3):237–249.
30. Samuh, M. H. and Pesarin, F. (2018). Applications of conditional power function of two-sample permutation test. *Computational Statistics*, 33(4):1847–1862.
31. Sarikavanij, S., Kasala, S., Sinha, B. K., and Tiensuwan, M. (2014). Estimation of location and scale parameters in two-parameter exponential distribution based on ranked set sample. *Communications in Statistics-Simulation and Computation*, 43(1):132–141.
32. Shaibu, A. B. and Muttlak, H. A. (2004). Estimating the parameters of the normal, exponential and gamma distributions using median and extreme ranked set samples. *Statistica*, 64(1):75–98.
33. Shorack, G. R. and Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics*. Wiley Series in Probability & Mathematical Statistics, New York.
34. Takahasi, K. and Wakimoto, K. (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering. *Annals of the Institute of Statistical Mathematics*, 20(1):1–31.