Comparing the Performance of Randomization Tests and Traditional Tests: A Simulation Study

W.B.M.R.D. Wijesuriya[1], C.H.Magalla[2], D. Kasturiratna[3]

[1,2]*Department of Statistics, University of Colombo, Sri Lanka*

[3]*Department of Mathematics and Statistics, Northern Kentucky University, USA*

[1]rush,wijesuriya@gmail.com, [2]champa@stat.cmb.ac.lk,[3]Kasturirad1@nku.edu

Abstract

Being non parametric in nature, the randomization tests (RTs) differ from the parametric statistical tests in many aspects and are often assumed to be more robust than parametric tests when their assumptions are violated. However, this ideology lacks sufficient evidence and the virtues of the RTs continue to be debated in the literature often with different conclusions. As a result researchers are often reluctant to employ RTs which are different from status quo and opt to use the traditional tests, regardless of the characteristics of their data. Hence this study compares the robustness, in terms of type I error rate and the power, of the most widely used classical parametric tests; pooled t test, unpooled t test, paired t test and one way ANOVA F test with their respective randomization counterpart using simulations under several trial conditions. While highlighting the seldom unrecognised potential of the RTs, the results concluded that, although the RTs are more robust in the presence of certain parametric assumption violations, this should not be a general rule and hence should only be used under the appropriate conditions for each test as demonstrated.

Keywords: **Randomization tests, Permutation tests, t test, ANOVA F test, Type I error, Power**

I. INTRODUCTION

Parametric statistical tests such as the t test and F test, assumes that the variable in question has a known underlying distribution that can be defined. In addition to that, the parametric tests also have other assumptions about homogeneity of variances and independence of observations (Ludbrook & Dudley, 1998; Berry, Mielke, & Mielke jr, 2001). These assumptions of the parametric tests are indispensable for two reasons. First, they place constraints on the interpretation of the results of the test (Snijders, 2001). Second, the characteristics of the population sampled, are used to draw inferences. Hence the parametric assumptions are important in deriving the optimal parametric test.

Randomization tests (RTs) became known through R.A Fisher's (1935) demonstration that the assumption of normality is not a must for analyzing data (David 2008). RTs make no reference to a population and hence do not require random sampling (Potvin & Roff, 1993; Ludbrook & Dudley, 1998).

RTs only require random assignment of treatments to experimental units. Few experiments in behavioural sciences such as biology, education, psychology, medicine or any other field use randomly selected subjects (Edgington & Onghena, 2007; Huo, Noortgate, Heyvaert, & Onghena, 2010;Huo & Onghena, 2012). According Hunter and May (1993), in most research, the population model of inference enters statistical analysis not because the experimenter wishes to generalize the results to a population, but because the model is so common that many assume it's the only method available for hypothesis testing. RTs also permit to assess statistical significance of nearly any parameter. (Peres-neto & Olden 2001). In addition to that, the inferences of the RTs refer only to the actual experimental units involved in the experiment/problem.

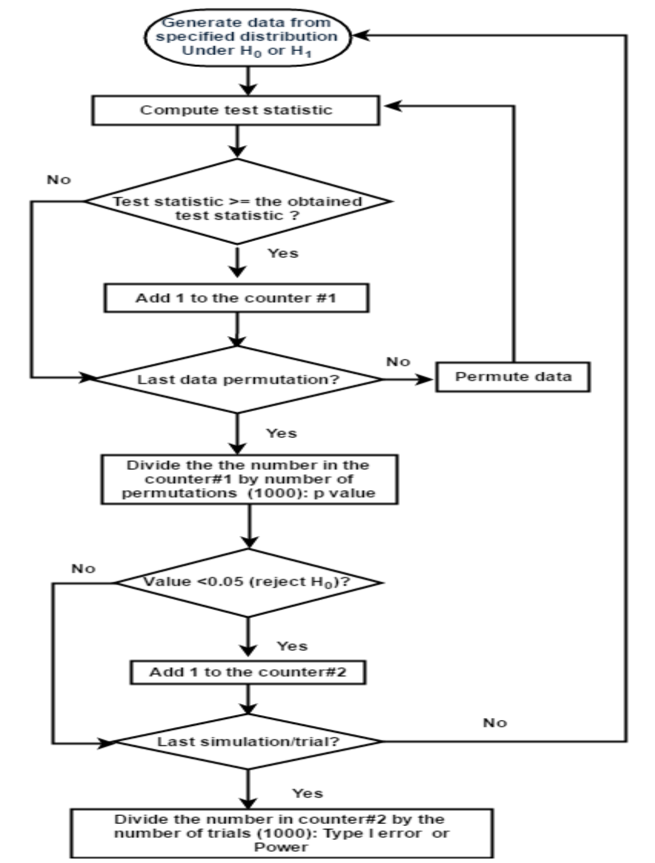
As many of the restrictions that were placed upon randomization tests are being resolved, more literature on randomization tests is timely so that students, researchers and statisticians in general are well versed with it. Therefore the main objective of this research is to compare the performance(type I error and power) of some of the most commonly used parametric tests; pooled t test, unpooled t test, paired t test and one way ANOVA F test with the randomization tests to discriminate the statistical conditions that support the two different tests.

II. THEORY AND SIMULATION PROCEDURE

Probability of type I error (α) of a test (size/ the nominal power) is defined to be the probability of the test rejecting the null hypothesis (H_0) when H_0 is true. This probability is often fixed by the statistician. Hence this probability should be at least in the proximity of the claimed value in order for the test to be relevant. (Higgins, 2004). The value of α was fixed at 5% in this study. Power of a statistical test ($1-\beta$) is defined as the probability of the test rejecting H_0 when H_0 is false. (Mood, Graybill, & Boes 1950).

RTs belong to a larger class of statistical tests called the permutation tests. The procedure for RT involves reshuffling/permuting the data and calculating the test statistic for each permutation, to compile the sampling distribution of the test statistic. Hence there is no requirement that the test statistic used, should conform to a mathematically definable probability distribution (Berry, Mielke, & Mielke jr, 2001). This means that the RTs can determine the p value directly from the data, without the use of reference tables based on probability distributions unlike t test or F test.

FIG.1 shows the basic steps followed in performing a RT.



Generate data from specified distribution Under H_0 or H_1
Compute test statistic
No
Test statistic >= the obtained test statistic ?
Yes
Add 1 to the counter #1
No
Last data permutation?
Permute data
Yes
Divide the the number in the counter#1 by number of permutations (1000): p value
No
Value <0.05 (reject H_0)?
Yes
Add 1 to the counter#2
No
Last simulation/trial?
Yes
Divide the number in counter#2 by the number of trials (1000): Type I error or Power

FIG.1Flow chart for the basic steps followed in the simulation procedure

In this study, the number of permutations and simulations considered were both 1000. In order to obtain the Type I error rates for the RTs, samples were generated under H_0 (Δ=0) for each trial. The p value for each trial was calculated as the proportion of permutations that generated a test statistic that were equal or bigger than the test statistic obtained for the original sample. This p value was used to reject or not reject H_0 in each trial. Type I error rates for each of the tests were determined by dividing the number of mistakenly rejected H_0 s in 1000 trials. To calculate the power, samples were generated under H_1 (Δ>0), the alternate hypothesis (H_0 false). Then the power was determined as the number of correctly rejected H_0 s in 1000 trials. The procedure was carried out using R.

The simulation results for the size and power estimates for the unpooled t test, paired t test, F test and the RT under different trial conditions such as sample size ratios, distribution shapes, effect sizes (Δ) and variance ratios are summarised in Tables I,II,III and IV. Four different sample sizes (5, 10, 30 and 50) and four types of distributions; Normal, Uniform (symmetric), Exponential and Gamma (skewed) were considered. As the results for both skewed distributions were similar, only the results generated for one type of skewed distribution are included.

III. SIMULATION RESULTS

TABLE I
SIZE AND POWER VALUES (%) OF POOLED T TEST AND RT

n	**5:5**		**10:10**		**30:30**		**50:50**	
Δ	**Normal(σ1=σ2=10)**							
	t test	RT	t test	RT	t test	RT	t test	RT
0	4.2	4.9	4.3	4.0	4.8	4.4	5.1	4.7
5	16.1	15.3	29.1	28.3	60.1	58.9	79.3	78.6
10	41.8	40.4	68.2	67.2	98.7	98.7	100.0	100.0
	Uniform(σ1=σ2=10): similar type I error and power values as Normal case							
	Exponential(σ1=σ2=10)							
	t test	RT	t test	RT	t test	RT	t test	RT
0	4.7	4.6	4.4	4.6	4.7	4.1	4.3	4.0
5	23.8	24.8	31.3	31.7	63.1	62.4	82.3	82.2
10	49.6	48.6	73.9	73.7	97.8	97.7	99.8	99.8
Δ	**Normal(σ1=10,σ2=20)**							
	t test	RT	t test	RT	t test	RT	t test	RT
0	5.3	4.8	5.4	5.4	5.0	4.9	5.0	4.9
5	12.6	11.7	16.7	16.5	35.6	35.1	47.3	47.4
10	26.6	26.0	40.6	39.6	79.3	78.1	94.4	94.0
	Uniform(σ1=10,σ2=20) : similar to the Normal case							
	Exponential(σ1=10,σ2=20)							
	t test	RT	t test	RT	t test	RT	t test	RT
0	11.8	11.4	9.6	9.6	9.4	9.0	6.1	6.1
5	22.7	21.9	24.5	24.6	38.5	38.3	52.0	52.0
10	38.3	37.0	46.3	46.0	76.4	76.1	89.2	88.9
n	**5:10**		**10:20**		**30:60**		**50:100**	
Δ	**Normal(σ1=σ2=10)**							
	t test	RT	t test	RT	t test	RT	t test	RT
0	5.6	5.6	5.3	5.3	4.5	4.7	4.4	4.8
5	21.7	21.4	33.0	32.3	72.7	71.9	89.1	88.6
10	55.1	54.4	80.0	79.5	99.8	99.7	100.0	100.0
	Uniform(σ1=σ2=10) : similar to the normal case							
	Exponential(σ1=σ2=10)							
	t test	RT	t test	RT	t test	RT	t test	RT
0	2.9	4.4	3.6	4.7	4.9	5.6	5.2	5.1
5	28.3	32.7	42.6	45.6	73.0	73.7	89.4	89.9
10	60.0	63.0	81.2	81.6	99.4	99.4	100.0	100.0

When the variances were homogenous, both the tests maintained the relevant type I error regardless of other trial conditions. Although the two tests had very similar powers, pooled t test was slightly more powerful except in small samples of asymmetric distributions. With unequal sample sizes (unbalanced designs) both the tests were relevant for all other trial conditions, only if the data was normal or symmetric. Whereas when the data was asymmetric, RT was more reliable in maintaining the relevant size. Pooled t test was slightly higher in power for normal or symmetric data while the RT was more powerful in capturing false hypotheses in asymmetric data. When the variances were heterogeneous, both the tests were only relevant as long as the data are not asymmetrically distributed. Although the two tests had very similar powers, pooled t test was slightly more powerful except in small samples of asymmetric distributions.

TABLE II
SIZE AND POWER VALUES (%) OF UNPOOLED T TEST AND RT

n	5:5		10:10		30:30		50:50	
Δ	**Normal($\sigma_1=5,\sigma_2=10$)**							
	t test	RT	t test	RT	t test	RT	t test	RT
0	4.6	4.9	4.6	4.6	5.3	5.3	5.3	5.3
5	20.8	21.3	38.9	38.9	79.1	79.0	92.6	92.2
10	54.6	55.8	83.8	84.8	99.9	99.9	100.0	100.0
	Uniform($\sigma_1=5,\sigma_2=10$) : similar to the Normal case							
	Exponential($\sigma_1=5,\sigma_2=10$)							
	t test	RT	t test	RT	t test	RT	t test	RT
0	10.5	11.3	10.1	10.1	8.0	7.9	8.1	7.6
5	32.8	33.4	45.2	45.4	76.5	76.8	88.9	88.6
10	60.5	61.2	80.3	81.6	99.4	99.6	100.0	100.0
	Normal($\sigma_1=5,\sigma_2=10$)							
n	**5:10**		**10:20**		**30:60**		**50:100**	
Δ	t test	RT	t test	RT	t test	RT	t test	RT
0	5.3	4.8	5.4	5.4	5.0	4.9	5.0	4.9
5	12.6	11.7	16.7	16.5	35.6	35.1	47.3	47.4
10	26.6	26.0	40.6	39.6	79.3	78.1	94.4	94.0
n	**10:5**		**20:10**		**60:30**		**100:50**	
Δ	t test	RT	t test	RT	t test	RT	t test	RT
0	5.0	8.5	5.1	8.7	5.0	8.9	5.1	8.4
	Uniform($\sigma_1=5,\sigma_2=10$)							
n	**5:10**		**10:20**		**30:60**		**50:100**	
Δ	t test	RT	t test	RT	t test	RT	t test	RT
0	4.4	2.5	5.4	2.8	5.0	2.9	4.8	2.1
5	30.9	20.0	54.2	38.8	93.5	88.6	99.2	98.4
10	74.9	62.3	98.1	94.9	100.0	100.0	100.0	100.0
n	**10:5**		**20:10**		**60:30**		**100:50**	
Δ	t test	RT	t test	RT	t test	RT	t test	RT
0	5.3	9.1	6.0	10.0	5.4	9.3	4.7	8.3
	Exponential($\sigma_1=5,\sigma_2=10$)							
n	**5:10**		**10:20**		**30:60**		**50:100**	
Δ	t test	RT	t test	RT	t test	RT	t test	RT
0	4.4	2.5	5.4	2.8	5.0	2.9	4.8	2.1
5	30.9	20.0	54.2	38.8	93.5	88.6	99.2	98.4
10	74.9	62.3	98.1	94.9	100.0	100.0	100.0	100.0
n	**10:5**		**20:10**		**60:30**		**100:50**	
Δ	t test	RT	t test	RT	t test	RT	t test	RT
0	7.3	6.8	6.0	4.3	5.9	3.6	5.2	3.1

When the variances were heterogeneous, both the tests only maintained the relevant size if the data comes from a normal or symmetric distributions. Although the powers of both the tests were quite similar the unpooled t test is slightly more powerful, except in small samples. With unequal sample sizes which are indirectly related to the variances (leading to high variances), the unpooled t test was more reliable in protecting the type I error rate in normal or symmetric data. In asymmetric data the unpooled t test was reliable for larger samples. Here the RT was more powerful for all distributions. With unequal sample sizes which are directly related to the variances (leading to low variances) also the unpooled t test was more reliable in protecting the type I error rate in normal or symmetric data and asymmetric large samples. Here however the unpooled t test was more powerful.

TABLE IIII
SIZE AND POWER VALUES (%) OF PAIRED T TEST AND RT

n	5:5		10:10		30:30		50:50	
Δ	**Normal($\sigma_1=\sigma_2=10$)**							
	t test	RT	t test	RT	t test	RT	t test	RT
0	5.0	3.8	5.4	5.4	5.3	5.1	5.2	4.9
5	44.1	30.5	74.4	73.8	99.4	99.4	99.9	99.9
10	91.6	77.5	100.0	100.0	100.0	100.0	100.0	100.0
	Uniform($\sigma_1=\sigma_2=10$): similar performance as Normal case							
	Exponential($\sigma_1=\sigma_2=10$)							
	t test	RT	t test	RT	t test	RT	t test	RT
0	3.0	2.8	4.6	4.8	5.2	4.9	5.1	4.8
5	51.6	39.4	75.3	76.1	98.0	97.8	99.9	99.8
10	84.6	72.0	96.1	96.0	100.0	100.0	100.0	100.0
Δ	**Normal($\sigma_1=10,\sigma_2=20$)**							
	t test	RT	t test	RT	t test	RT	t test	RT
0	4.8	3.0	4.7	4.6	5.8	5.8	4.9	4.7
5	18.9	12.2	30.6	29.2	65.1	64.2	83.3	82.8
10	40.8	28.5	72.6	71.6	99.1	98.9	100.0	100.0
	Uniform($\sigma_1=10,\sigma_2=20$): similar to the Normal case							
	Exponential($\sigma_1=10,\sigma_2=20$)							
	t test	RT	t test	RT	t test	RT	t test	RT
0	16.9	11.6	12.2	12.2	10.4	10.3	7.9	7.9
5	35.2	24.7	39.5	38.9	63.1	63.5	76.4	76.0
10	51.0	38.9	69.5	69.5	95.6	95.7	99.2	99.3

With homogenous variances, paired t test was relevant for normal and symmetric data regardless of the sample size. The RT was only relevant for moderate or large samples. When the distributions were skewed with homogenous variances, both the tests were only relevant for moderate or large samples. When the variances were heterogeneous, the paired t test was relevant for normal and symmetric data regardless of the sample size while the RT was only relevant for moderate or large samples. With heterogeneous variances in skewed data both the tests did not maintain the relevant size. The paired t test was more powerful regardless of the trial conditions.

TABLE IV
SIZE AND POWER VALUES (%) OF ANOVA F TEST AND RT

n	5:5:5		10:10:10		30:30:30		50:50:50	
Δ	**Normal($\sigma_1=\sigma_2=\sigma_3=10$)**							
	F test	RT	F test	RT	F test	RT	F test	RT
0	5.8	5.9	4.9	4.8	5.6	5.5	5.1	5.1
5	11.3	10.5	17.5	17.3	49.3	48.6	72.6	72.0
10	30.1	29.4	54.4	54.4	98.0	97.7	100.0	100.0
	Uniform ($\sigma_1=\sigma_2=\sigma_3=10$) :similar to the Normal case							

	Gamma(σ_1=σ_2= σ_3=10)							
	F test	RT	F test	RT	F test	RT	F test	RT
0	3.0	4.2	3.7	4.4	4.1	4.1	5.1	5.2
5	10.3	12.0	19.8	20.5	51.2	51.5	71.7	71.3
10	35.8	36.0	60.2	60.3	97.6	97.7	99.9	99.9
Δ	**Normal(σ_1=5,σ_2=10, σ_3=15)**							
	F test	RT	F test	RT	F test	RT	F test	RT
0	6.8	7.8	5.4	5.7	5.6	6.0	5.0	5.0
5	11.1	11.9	18.3	18.1	40.7	40.7	60.1	60.0
10	27.9	29.0	49.0	49.1	92.3	92.1	99.4	99.4
	Uniform (σ_1=5,σ_2=10, σ_3=15): similar to the Normal case							
Δ	**Gamma(σ_1=5,σ_2=10, σ_3=15)**							
	F test	RT	F test	RT	F test	RT	F test	RT
0	9.4	9.9	8.2	9.4	8.4	8.5	6.0	6.3
5	6.1	7.9	14.1	16.0	39.4	40.9	62.4	62.8
10	23.9	29.0	50.1	54.0	94.6	94.8	99.7	99.7
n	**5:10:15**		**10:20:30**		**30:60:90**		**50:100:150**	
Δ	**Normal(σ_1=σ_2= σ_3=10)**							
	F test	RT	F test	RT	F test	RT	F test	RT
0	5.1	5.0	5.3	5.3	4.4	4.3	5.5	5.5
5	18.6	18.4	38.4	37.6	83.8	83.4	97.7	97.7
10	65.0	64.8	93.4	93.4	100.0	100.0	100.0	100.0
	Uniform (σ_1=σ_2= σ_3=10):similar to the Normal case							
	Gamma(σ_1=σ_2= σ_3=10)							
	F test	RT	F test	RT	F test	RT	F test	RT
0	4.6	4.6	4.6	4.8	4.5	4.7	5.4	5.4
5	21.1	21.7	41.9	41.1	84.8	84.5	97.4	97.3
10	68.4	68.9	92.4	92.6	100.0	100.0	100.0	100.0

With homogenous variances, both the tests were relevant for normal and symmetrically distributed data but F test was better in terms of power. For skewed data with homogenous variances, RT was relevant for all sample sizes, while the F test was only relevant for moderate or higher sample sizes. When the variances were heterogeneous, both tests were not relevant for any distribution and the RT was more powerful for normal and skewed data. For uniform data with heterogeneous variances, both tests showed similar power. When the sample sizes were unequal, both the tests were relevant regardless of the distribution. Here for normal and symmetric data, the F test was more powerful and for skewed data, the RT was powerful for small samples while for larger samples F test was more powerful.

IV. CONCLUSION

The research concluded that both the tests are not robust in terms of the type I error rate for skewed distributions with unequal variances. The pooled t test was more powerful except in skewed distributions with small sample sizes or unequal sample sizes. The simulations illustrated that the RT was more powerful than the unpooled t test when the variances are high. Also the paired t test proved to be superior in power even when its assumptions were violated. For skewed distributions, the more appropriate test among the F test and RT was identified as the RT.

The study also showed that in the presence of violations of the assumptions embedded with the classical tests, RTs do not always necessarily exhibit higher robustness in terms of power and the type I error rates. For instance, the paired t test is more robust than the RT in terms of power even when the assumptions of the paired t test do not hold, for all types of distributions. Hence RTs should not be recommended as an alternative to the parametric tests whenever the assumptions are violated and should only be used under the appropriate conditions such as which were shown in this study. In certain instances where none of the parametric assumptions are violated, the RTs showed slightly lower but nearly similar robustness as the parametric tests such as the pooled t test in the presence of homogenous variances and unpooled t test in the presence of heterogeneous variances. This stands as evidence for the fact that RTs, also considering its advantages, has much potential as a statistical test and are underutilized compared to the traditional tests.

REFERENCES

[1] Adams, D. C., & Anthony, C. D. (1996). Using Randomization Techniques to Analyse Behavioural Data. *Animal Behaviour* , *51* (4), pp.733-738.

[2] Berry, K., Mielke, P., & Mielke, H. (2001). The Fisher-Pitman permutation test: An attractive alteranative to the F test. Psychological Reports, 90(2), pp.495-502.

[3] Boik, R. J. (1987). The Fisher-Pitman permutation test: A non-robust alternative to the normal theory F test when variances are heterogeneous. *British Journal of Mathematical and Statistical Psychology* , *40*, pp.26-42.

[4] David, H. (2008). The Beginnings of Randomization Tests. The American Statistician, 62(1), pp.70-72.

[5] Edgington, E. and Onghena, P. (2007). *Randomization tests*. 4th ed. CRC press.

[6] Higgins, J. (2004). An introduction to modern nonparametric statistics. 1st ed. Pacific Grove, CA: Brooks/Cole.

[7] Huo, M., & Onghena, P. (2012). RT4WIN: A Windows-based program for randomization tests. Psychologica Belgica, 52 (4),pp. 387-406.

[8] Huo, M., Noortgate, W. V., Heyvaert, M., & Onghena, P. (2010). A Systematic Review on Randomization and Permutation Tests in the Educational and Behavioral Sciences. Measuring Behavior , 19, p.456.

[9] Ludbrook, J. (1994). Advantages of permutation (randomization) tests in clinical and experimental pharmacology and physiology. *Clinical and Experimental Pharmacology and Physiology* , *21*,pp. 673-686.

[10] Ludbrook, J., & Dudley, H. (1998). Why Permutation Tests are Superior to t and F Tests in Biomedical Research. The American Statistician , 52 (2), pp.127-132.

[11] Manly, B. F. (1995). Randomization tests to compare the means with unequal variance. *Sankhya : The Indian Journal of Statistics* , *57*,pp. 200-222.

[12] May, R. and Hunter, M. (1993). Some advantages of permutation tests. *Canadian Psychology/Psychologie canadienne*, 34(4), pp.401-407.

[13] Mood, A., Graybill, F. and Boes, D. (1950). Introduction to the theory of statistics. Second edition. 3rd ed. McGraw-Hill Book Co., New York, N.Y.

[14] Peres-neto, P. R., & Olden, J. D. (2001). Assessing the robustness of randomization tests:examples from behavioural studies. Animal Behaviour, 61,pp. 79-86.

[15] Potvin, C., & Roff, D. A. (1993). Distribution-Free and Robust Statistical Methods: Viable Alternatives to Parametric Statistics. Ecology, 74 (6), pp.1617-1628.