An outlier in an independent samples design

Ben Derrick

Royal Statistical Society Conference, 2018

Abstract

There is a flaw with some of the most commonly performed statistical tests. A paradox of the one sample t-test is the contrariwise decrease in the p-value as the value of an outlier increases in the direction of the overall effect. Demonstration of this paradox is extended to the equal variances assumed and Welch's unrestricted to equal variances independent samples t-test. The phenomenon is explored using Monte-Carlo simulation, and compared with alternative two sample tests; the Mann-Whitney U test, and the Yuen-Welch t-test with 10% trimming per tail. Scenarios where the overall effect is concordant or discordant with the direction of the aberrant observation are considered.

Sample data is generated under normality, with the subsequent inclusion of an aberrant observation in one sample. The aberrant observation is systematically varied. The total sample sizes for each of the two samples within a factorial design are {10, 15, 20}. The variances within the factorial design are {1, 4}. For each parameter combination, the proportion of 10,000 iterations where the null hypothesis is rejected is calculated at the 5% significance level, two sided.

It is evidenced that the paradox for both forms of the independent samples t-test is exacerbated when the smaller sample size with the higher variance includes the aberrant observation, and as the imbalance between the sample sizes increases. Results also indicate that when the sample with the lower variance includes the aberrant observation, Welch's t-test and the Yuen-Welch t-test most closely retain Type I error robustness.

Recommendations on choice of test for independent samples designs are given, ending with discussion on how these results impact analyses for partially overlapping samples designs.

Introduction

An outlier is an observation that apparently deviates from other observations (Grubbs, 1969). An outlier can cause serious problems in statistical analyses. An outlier increases the variability within a sample, increasing the probability of making a Type II error, an issue that is exacerbated for small sample sizes (Cousineau and Chartier, 2010).

Zumbo and Jennings (2002) identify two types of non-normality: (i) samples from non-normal distributions, and (ii) samples from inherently normal distributions, but with outliers. The latter is considered here.

A paradox of the one sample t-test is the contrariwise decrease in the p-value as the value of an outlier increases in the direction of the overall effect (Derrick et al., 2017a). The paired samples t-test is also effected by the paradox, this is because the paired samples t-test is equivalent to the one sample t-test performed on sample differences. This phenomenon is referred to as as the extreme observation paradox (Derrick et al., 2017a).

An observation that may appear to be an outlier, may represent a location shift (Walfish, 2006). This location shift may be masked when performing the one sample t-test due to the increase in variability (Derrick et al., 2017a).

It follows that the independent samples t-test may also exhibit the extreme observation paradox. The form of the independent samples t-test with the assumption of equal variances relaxed is referred to as Welch's test, and is robust under normality (Derrick, Toher, and White, 2016; Ruxton, 2006), but may not be robust when outliers are present.

Other tests based on the t-test may also be subject to the same paradox. This includes the partially overlapping samples t-tests by Derrick (2017), Derrick et al. (2017b), and Derrick, White, and Toher (n.d.). These tests are used when there is a combination of independent observations and paired observations. For example, in a paired samples design where experimental error creates a scenario where both paired observations and independent observations are present. This example scenario can be considered as data missing by design, and therefore the assumption of missing completely at random (MCAR) holds (Kang, 2013). The partially overlapping samples t-test is an interpolation between the independent samples t-test and the paired samples t-test, so may be an appropriate alternative test under the conditions of normality and MCAR (Derrick, Toher, and White, 2017). Thus the properties of the independent samples t-test will impact the partially overlapping samples t-test.

Focus is on relatively small sample sizes, these are situations in which potentially extreme observations may have the greatest practical impact.

In this paper, simulation methodology for exploring the performance of

tests for a two independent samples designs is given. This is followed by results of the simulation. The extension to partially overlapping samples scenarios is finally considered.

Simulation methodology

The approach is to generate two independent samples under the normality assumption, then include one aberrant observation in one sample. This additional observation will systematically change in its observed value. Thus one observation is directly manipulated to create an extreme observation with otherwise normally distributed data. This aberrant observation may be compounded with outliers due to inherent variability within the other observations (Anscombe, 1960).

The tests performed are the independent samples t-test (equal variances assumed), Welch's test, the Mann-Whitney test, and the Yuen-Welch test. Under a nil-null hypothesis; the independent samples t-test and Welch's test are used to test a distribution mean difference of zero; the Yuen-Welch test is used to test the distribution of the trimmed means equal to zero. Under the same conditions, the Mann-Whitney test is used to test a null hypothesis of distribution differences symmetrically distributed around zero.

The Yuen-Welch test is performed using the R package 'PairedData' with 10% trimming per tail as outlined by (Wilcox, 2012), the other tests are performed using the 'stats' package.

Specifically, in Sample 1, n_a Standard Normal deviates are generated, and in Sample 2, n_{b-1} Standard Normal deviates are generated. The Mersenne-Twister algorithm (Matsumoto and Nishimura, 1998) generates uniform random deviates, then the Paley and Wiener (1934) transformation is applied to obtain Standard Normal deviates.

A fixed aberrant observation, x_b, is appended to the $x_1, x_2 \cdots, x_{b-1}$ observations to give a total sample size of n_b. For each simulated sample, the value of x_b is systematically varied from -8 to 8 in increments of 0.1. It is this value, x_b, which is referred to as the 'marching observation'. The values of x_b approximately range between ± 8 standard deviations from the mean and would therefore cover limits likely encountered in a practical environment.

If $\bar{x}_a - \bar{x}_{b-1} < 0$ then the observations in Sample 2 are multiplied by -1 to ensure a non-negative sample mean. This change of sign does not affect the validity of a two-sided test of a nil-null hypothesis for these data. This condition is to ensure that the concordance of effects ($\bar{x}_a - \bar{x}_{b-1} > 0, x_b > 0$) or discordance of effects ($\bar{x}_a - \bar{x}_{b-1} > 0, x_b < 0$) can be established.

The sample sizes of n_a and n_b that are varied within a factorial design

are {10, 15, 20}. The values of σ_1 and σ_2 that are varied within the factorial design are {1, 2}. The simulation is run 10,000 times for each parameter combination of n_a, n_b, x_b, σ_1, σ_2.

For each parameter combination and each test statistic, the interest is in the proportion of 10,000 iterations where the null hypothesis is rejected at the 5% significance level, two sided. This gives the Null Hypothesis Rejection Rate (NHRR). Note that the terminology NHRR is used rather than Type I error rate, because the inclusion of the marching observation strictly invalidate the underpinning assumptions.

The marching observation demonstrates the impact on the test statistics when the aberrant observations is close to a mean difference of zero, as well as what happens when extreme observations are included in a sample with a non-negative mean. The effect of gradually increasing the marching observation is to gradually violate the assumption of the nil-null hypothesis, large positive values of the marching observation would increase the NHRR. Negative values of x_b would cancel out the overall positive difference observed within the sample differences and decrease the NHRR.

Results

For illustrative purposes, the impact of the marching observation for a selection of parameter combinations from the independent samples simulation design are given.

As per the methodology outlined above, an aberrant observation is included in Sample 2.

Each scenario consists of a total sample size of 30. Scenarios 1-3 represent have equal sample sizes. Scenarios 4-6 have an imbalance in sample size. The six scenarios are as follows:

1. $n_a = 15,\ n_{b-1} = 14,\ \sigma_1 = 1,\ \sigma_2 = 1$
2. $n_a = 10,\ n_{b-1} = 19,\ \sigma_1 = 1,\ \sigma_2 = 1$
3. $n_a = 20,\ n_{b-1} = 9,\ \sigma_1 = 1,\ \sigma_2 = 1$
4. $n_a = 15,\ n_{b-1} = 14,\ \sigma_1 = 1,\ \sigma_2 = 2$
5. $n_a = 10,\ n_{b-1} = 19,\ \sigma_1 = 1,\ \sigma_2 = 2$
6. $n_a = 20,\ n_{b-1} = 9,\ \sigma_1 = 1,\ \sigma_2 = 2$

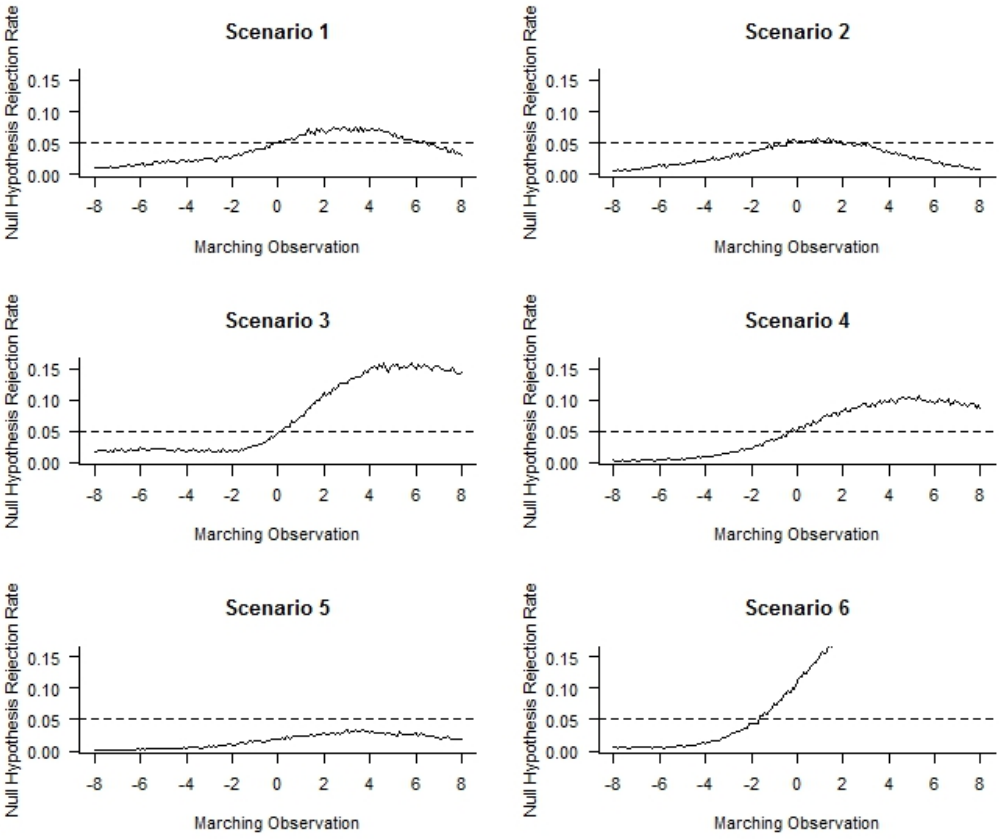


Figure 1: NHRR when performing the independent samples t-test

For each of the six scenarios; Figure 1 gives the NHRR when performing the independent samples t-test, Figure 2 gives the NHRR when performing Welch's test, Figure 3 gives the NHRR when performing the Yuen-welch test and Figure 4 gives the NHRR when performing the Mann-Whitney test.

Figure 1 shows that for Scenarios 1-4, when $x_b = 0$, the NHRR is approximately equal to the nominal Type I error rate of 5%. However, the extreme observation paradox can be observed through the contrariwise decrease in the NHRR as the value of an extreme observation increases in the direction of the overall effect. For positive sample means, as the value of x_b starts to increase above zero, the independent samples t-test has an increasingly higher NHRR, until a turning point is reached. This turning point has the effect that in some circumstances an extremely large aberrant observation becomes extreme enough to result in the null hypothesis being rejected (e.g. Scenario 1). For scenarios 5-6, when $x_b = 0$, the NHRR is not approximately equal to the nominal Type I error rate. This is due to the non-robustness of

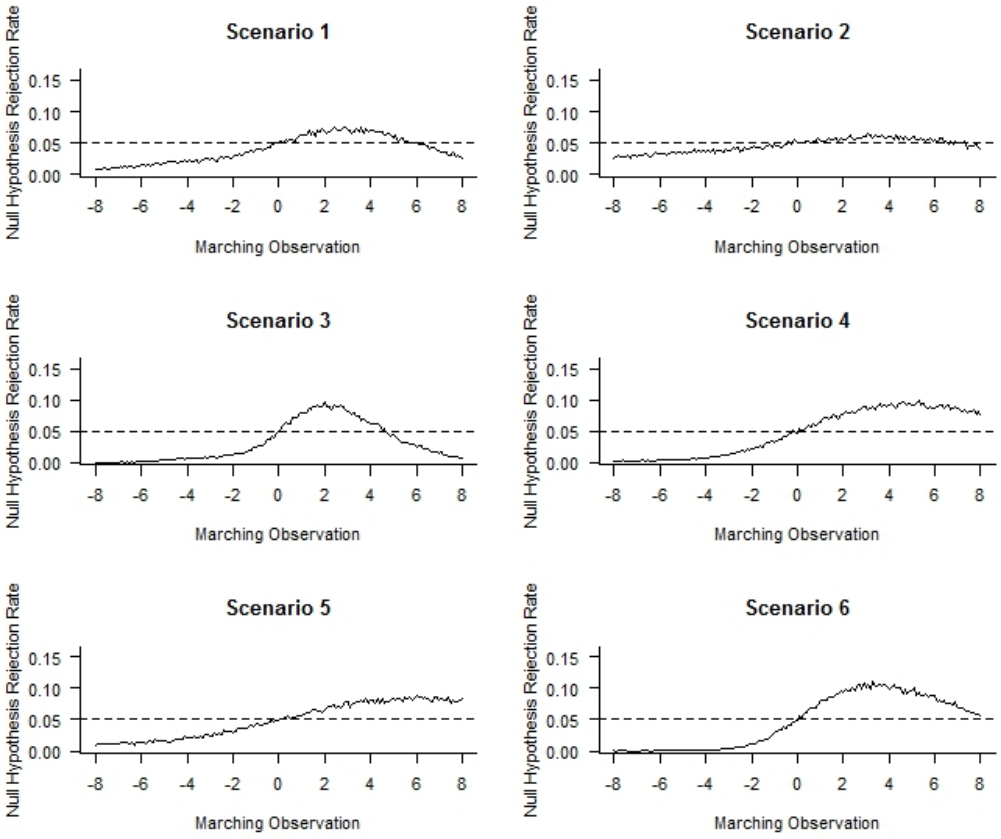


Figure 2: NHRR when performing Welch's test

the independent samples t-test under unequal variances.

Figure 2 shows that for each scenario, when $x_b = 0$ the NHRR is approximately equal to the nominal Type I error rate of 5%. This is anticipated given the known robustness of Welch's test. However, Figure 2 indicates that for increasing values of x_b, the paradox is also observed for Welch's test.

The only design difference between Scenario 2 and Scenario 3 is that the aberrant observation is in the larger sample for Scenario 2, and the smaller sample for Scenario 3. Figure 2 indicates that the test maintains closer robustness when the aberrant observation is in the larger sample. Likewise, The only design difference between Scenario 5 and Scenario 6 is the aberrant observation is in the larger sample for Scenario 5, and the smaller sample for Scenario 6. Figure 2 indicates that the paradox is exacerbated when the smaller sample size has the higher variance and includes the aberrant observation. Balanced samples sizes (Scenario 1, Scenario 4) are also preferable.

Figure 3 and Figure 4 show that the Mann-Whitney test and the Yuen-

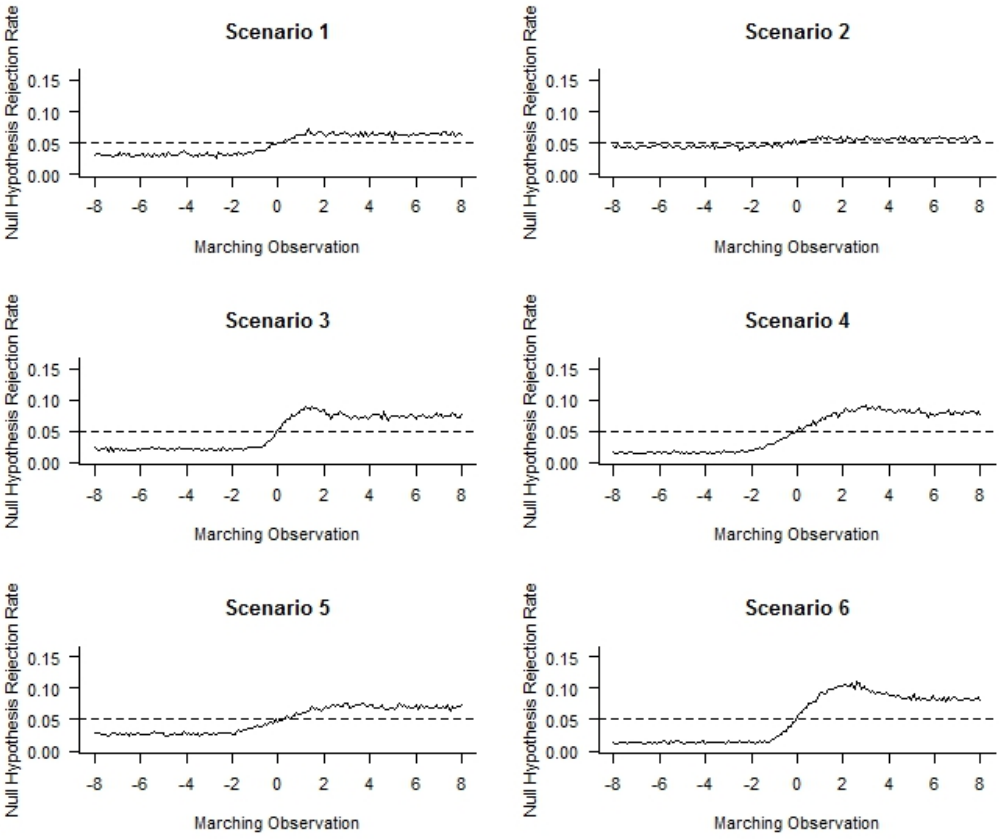


Figure 3: NHRR when performing the Yuen-Welch test

Welch test are liberal for positive values of the marching observation, and are conservative for negative values of the marching observation. The Mann-Whitney test and the Yuen-Welch test maintain NHRR close to the nominal significance level when sample sizes are equal or the larger sample size includes the marching observation. Both tests tend to a fixed value as $x_b \to \infty$, and both tests tend to a fixed value close to zero as $x_b \to -\infty$. Due to the use of rank values, the Mann-Whitney test is not greatly affected by the magnitude of the extreme observation. Similarly due to trimming, the Yuen-Welch test is not greatly affected by the magnitude of the extreme observation.

For independent samples, Bakker and Wicherts (2014) recommend proceeding with the Mann-Whitney test or the Yuen-Welch when outliers are present. However, Figure 3 and Figure 4 show that the fixed NHRR these tests tend to is dependent on sample size and variance. For Scenario 5 it can be seen that the NHRR when performing the Mann-Whitney test remains below the nominal significance level for all values of x_b. The results corrob-

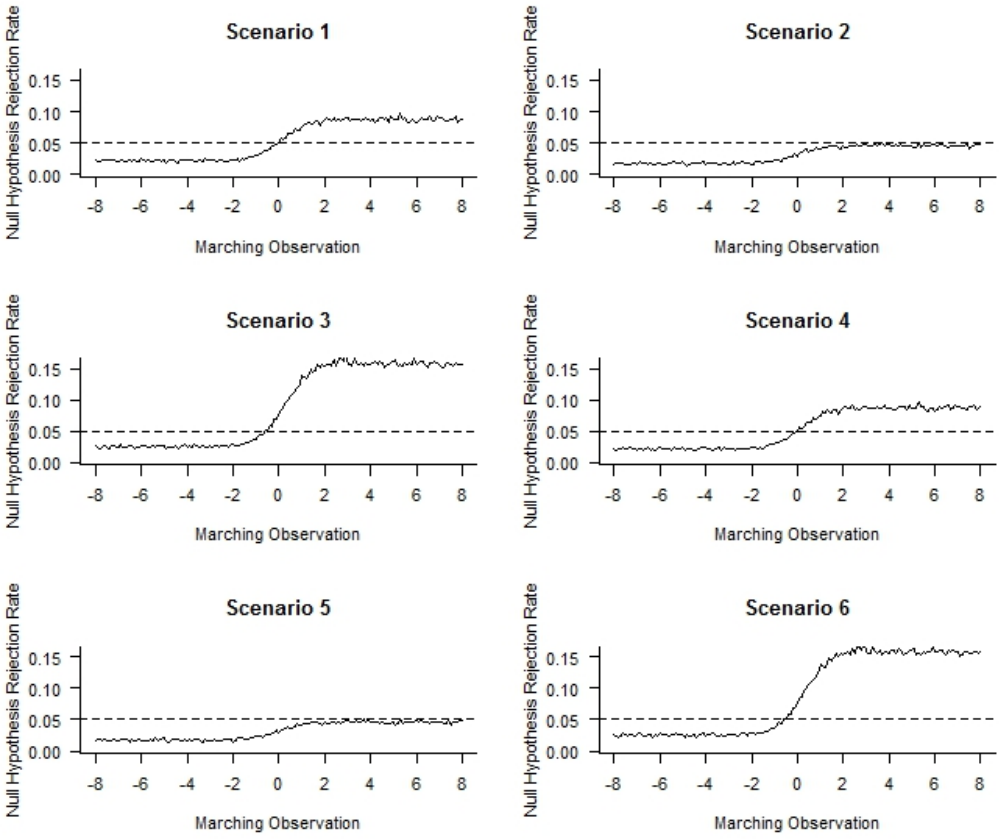


Figure 4: NHRR when performing the Mann-Whitney test

orate findings by Zimmerman (1998) that the Mann-Whitney test does not universally provide a robust alternative approach when an outlier is present.

Extension: partially overlapping samples

The simulation design is extended to the partially overlapping samples framework. The partially overlapping samples t-test assuming equal variances, T_{new1}, is assessed. This is compared against T_{RNK1}, substituting rank values into the test statistic as proposed by Derrick, White, and Toher (n.d.)

Under a two sided nil-null hypothesis; the parametric partially overlapping samples t-tests, T_{new1} is used to test for a distribution mean difference of zero. Under the same conditions, the non-parametric partially overlapping samples t-tests T_{RNK1} is used to test for a distribution of differences symmetrically distributed around zero.

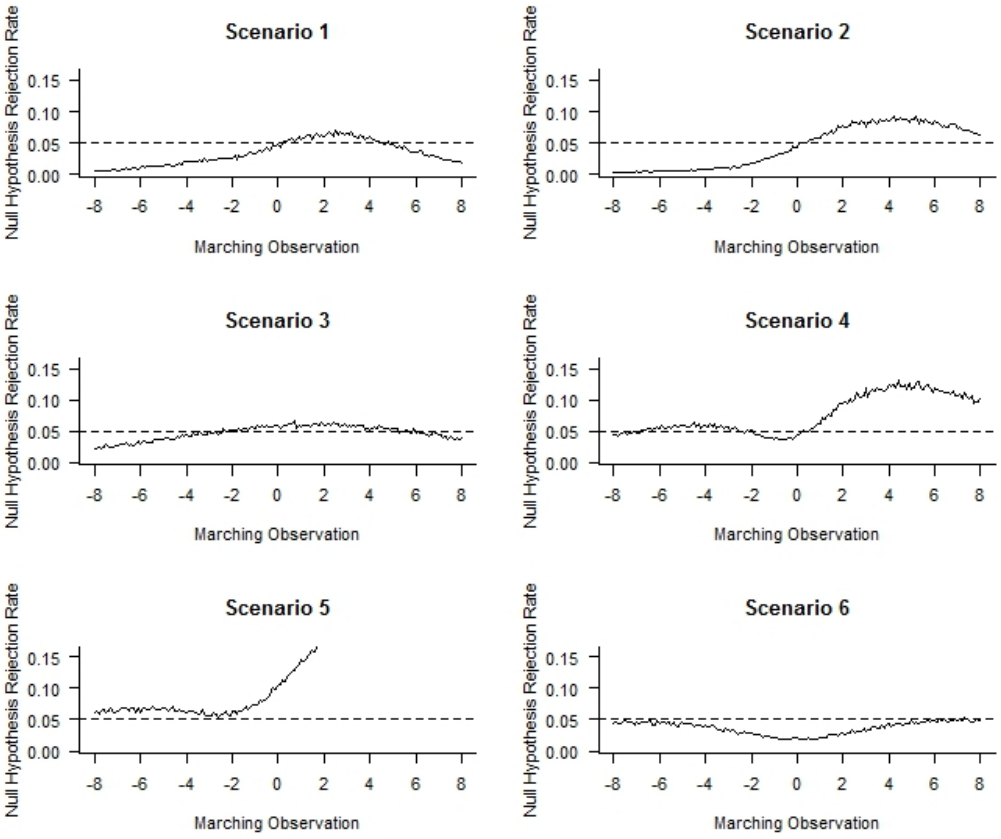


Figure 5: NHRR when performing T_{new1}

The approach is to simulate two groups of Normal deviates for a completely paired design with $n = 15$, $\rho = 0.5$. with $\sigma_1 = \{1, 2\}$ and $\sigma_2 = \{1, 2\}$.

Observations are deleted at random from the paired samples design with constraints so that remaining sample sizes are as per the scenario under consideration. The six scenarios are as follows:

1. $n_a = 5,\ n_{b-1} = 4,\ n_c = 5,\ \sigma_1 = 1,\ \sigma_2 = 1$
2. $n_a = 5,\ n_{b-1} = 4,\ n_c = 5,\ \sigma_1 = 1,\ \sigma_2 = 2$
3. $n_a = 5,\ n_{b-1} = 4,\ n_c = 5,\ \sigma_1 = 2,\ \sigma_2 = 1$
4. $n_a = 10,\ n_{b-1} = 3,\ n_c = 2,\ \sigma_1 = 1,\ \sigma_2 = 1$
5. $n_a = 10,\ n_{b-1} = 3,\ n_c = 2,\ \sigma_1 = 1,\ \sigma_2 = 2$
6. $n_a = 10,\ n_{b-1} = 3,\ n_c = 2,\ \sigma_1 = 2,\ \sigma_2 = 1$

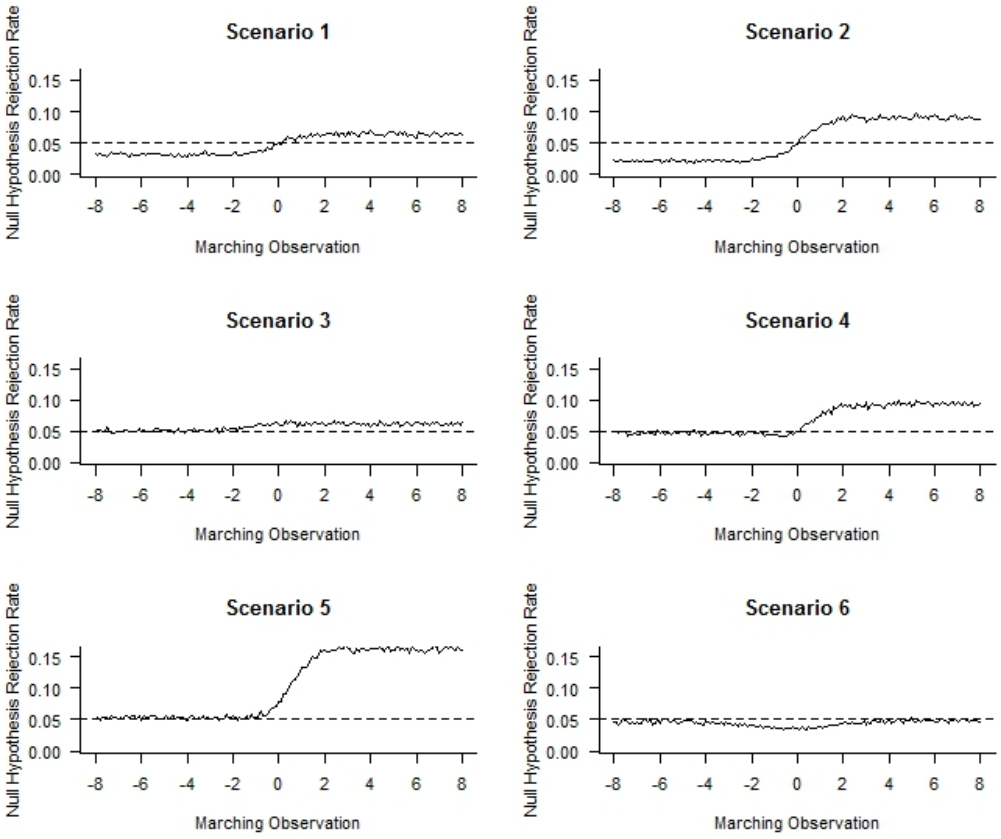


Figure 6: NHRR when performing T_{RNK1}

where n_a is the number of independent observations in Sample 1, n_b is the number of independent observations in Sample 2, and n_c is the number of pairs.

As previously, As previously, if $\bar{x}_a - \bar{x}_{b-1} < 0$ then the observations in Sample 2 are multiplied by -1 to ensure a non-negative sample mean. An additional observation, x_b, is appended to the n_{b-1}. For each simulated sample, the value of x_b is systematically varied from -8 to 8 in increments of 0.1. Again, it is this value, x_b, which is referred to as the 'marching observation'.

The six scenarios displayed are for indicative purposes of the behaviour of T_{new1} and T_{RNK1}. Scenarios 1-3 represent have equal sample sizes. Scenarios 4-6 have an imbalance in sample size. Figure 5 - Figure 6 display the NHRR for T_{new1} and T_{RNK1} respectively

The extreme observation paradox identified for the independent samples t-test is also observed in Figure 5. However, under unequal sample sizes

and unequal variances, further undesirable patterns are also observed. This can be explained by the non-robustness of the equal variances assumed test statistic in these conditions (Derrick et al., 2017b).

The poor outcomes for the parametric test are unsurprising following the conclusions based on paired samples tests and independent samples tests on which these tests are based.

Figure 6 show that the statistics T_{RNK1} tends to a fixed value for the NHRR. However, the fixed NHRR value is inflated when the smaller sample size is associated with the larger variance.

Summary

Given the debate in the literature regarding the removal of an outlier, it is important to be aware of the impact of an outlier on commonly used tests.

The results show that a single aberrant observation can potentially either mask true effects or show phantom significant effects.

There is a counter-intuitive decrease in the NHRR as the value of an extreme observation increases in the direction of the overall effect. This phenomenon is observed for the independent samples t-test and Welch's test. As a consequence, this paradox is also observed for the parametric partially overlapping samples t-test. Parametric tests display behaviour strongly dependent on the magnitude of the outlier.

Fagerland (2012) suggest that the problem is not the t-test itself, moreover it may be that in the presence of an outlier, the mean may be a poor measure of central location, and other measures of location may be more appropriate.

In contrast, test statistics making use of trimmed means or rank values do not suffer from the extreme observation paradox, and are not impacted by the magnitude of the outlier. However they are not necessarily robust for small sample sizes.

Typically the natural desire of a researcher is to prove significant effects, the researcher will often consider the removal of outliers in order to conclude a significant effect. It is demonstrated that the removal of an outlier may in fact produce the opposite outcome. The decision not to remove an outlier might be taken so that a significant effect is observed. Usually it is the removal of an outlier that requires justification, but in this respect a decision not to remove an observation should be considered with just as much vigour as the decision to remove an observation.

In textbooks listing the assumptions of the t-test, the assumption of no significant outliers is sometimes listed, but sometimes not. Given the results above, the assumption should be listed. The question of how to identify

a 'significant outlier' has no answer that is applied universally (Hodge and Austin, 2004), an extensive list is given by Bamnett and Lewis (1994), and is therefore an area of much debate that will continue.

References

Anscombe, F. J. (1960). "Rejection of outliers". In: *Technometrics* 2.2, pp. 123–146.

Bakker, M. and J. M. Wicherts (2014). "Outlier removal, sum scores, and the inflation of the type I error rate in independent samples t tests: The power of alternatives and recommendations." In: *Psychological methods* 19.3, p. 409.

Bamnett, V. and T. Lewis (1994). "Outliers in statistical data". In:

Cousineau, D. and S. Chartier (2010). "Outliers detection and treatment: a review." In: *International Journal of Psychological Research* 3.1, pp. 58–67.

Derrick, B. (2017). "Statistics: New t-tests for the comparison of two partially overlapping samples". In: *Faculty of Environment and Technology Degree Show, UWE, Frenchay Campus, UWE, 1 June 2017*.

Derrick, B., A. Broad, D. Toher, and P. White (2017a). "The impact of an extreme observation in a paired samples design". In: *metodološki zvezki-Advances in Methodology and Statistics* 14.

Derrick, B., B. Russ, D. Toher, and P. White (2017b). "Test statistics for the comparison of means for two samples which include both paired observations and independent observations". In: *Journal of Modern Applied Statistical Methods* 16.1, pp. 137–157.

Derrick, B., D. Toher, and P. White (2016). "Why Welch's test is Type I error robust". In: *The Quantitative Methods in Psychology* 12.1, pp. 30–38.

Derrick, B., D. Toher, and P. White (2017). "How to compare the means of two samples that include paired observations and independent observations: A companion to Derrick, Russ, Toher and White (2017)". In: *The Quantitative Methods in Psychology* 13.2, pp. 120–126.

Derrick, B., P. White, and D. Toher. "Parametric and non-parametric tests for the comparison of two samples which both include paired and unpaired observations". In: *Journal of Modern Applied Statistical Methods* [in press].

Fagerland, M. W. (2012). "t-tests, non-parametric tests, and large studies, a paradox of statistical practice?" In: *BMC medical research methodology* 12.1, p. 1.

Grubbs, F. E. (1969). "Procedures for detecting outlying observations in samples". In: *Technometrics* 11.1, pp. 1–21.

Hodge, V. J. and J. Austin (2004). "A survey of outlier detection methodologies". In: *Artificial intelligence review* 22.2, pp. 85–126.

Kang, H. (2013). "The prevention and handling of the missing data". In: *Korean journal of anesthesiology* 64.5, pp. 402–406.

Matsumoto, M. and T. Nishimura (1998). "Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator". In: *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 8.1, pp. 3–30.

Paley, R. E. A. C. and N. Wiener (1934). *Fourier transforms in the complex domain*. Vol. 19. American Mathematical Soc.

Ruxton, G. D. (2006). "The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test". In: *Behavioral Ecology* 17.4, pp. 688–690.

Walfish, S. (2006). "A review of statistical outlier methods". In: *Pharmaceutical technology* 30.11, p. 82.

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing*. Academic Press.

Zimmerman, D. W. (1998). "Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions". In: *The Journal of experimental education* 67.1, pp. 55–68.

Zumbo, B. D. and M. J. Jennings (2002). "The robustness of validity and efficiency of the related samples t-test in the presence of outliers". In: *Psicológica: Revista de metodología y psicología experimental* 23.2, pp. 415–450.

UWE Bristol | University of the West of England

Royal Statistical Society, International Conference, 2018

An outlier in an independent samples design

Ben.Derrick@uwe.ac.uk
Applied Statistics Group

Presented in Methods and Theory section at the Royal Statistical Society, Cardiff, 04/09/2018

Example

Sample A	Sample B
3	6
5	5
4	9
6	8
4	7

p-value

0.017 t-test equal variances assumed
0.019 t-test equal variances not assumed

Add one additional observation

Sample A	Sample B
-∞	∞

p-value

0.341 t-test equal variances assumed
0.363 t-test equal variances not assumed

Simulation study

Distribution: N(0,1) N(0,4)

‘marching observation’, additional observation within Sample 2, from -8 to 8 (increments of 1).

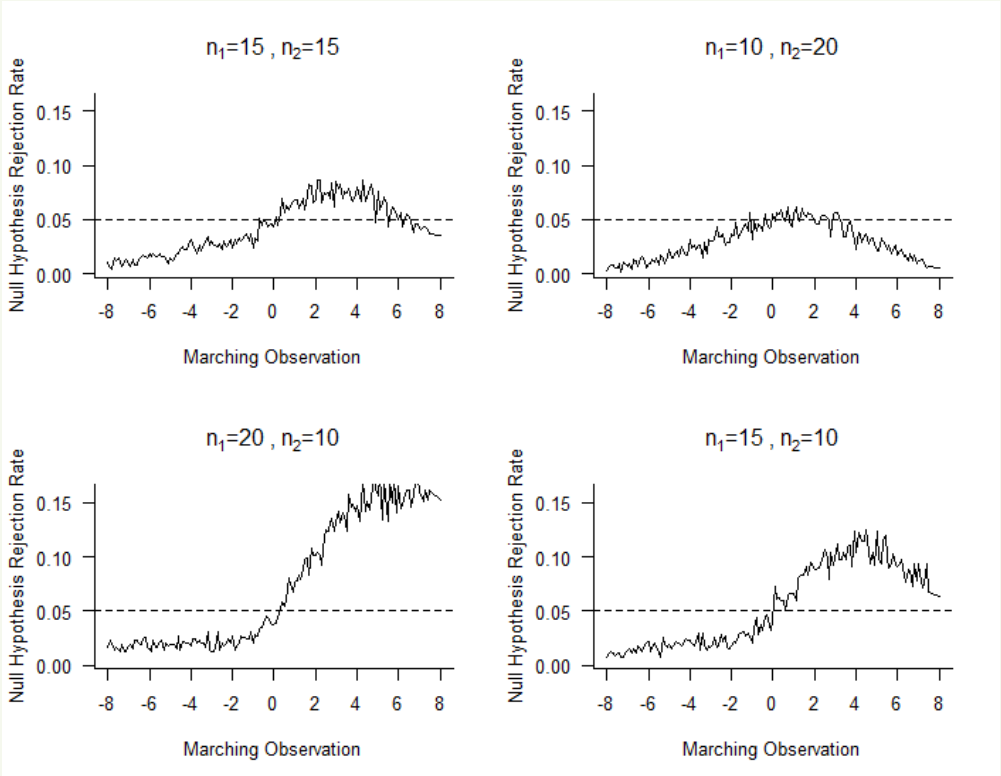
- Independent samples t-test (pooled variances)
- Welch’s test
- Mann-Whitney test

5% significance level, two sided

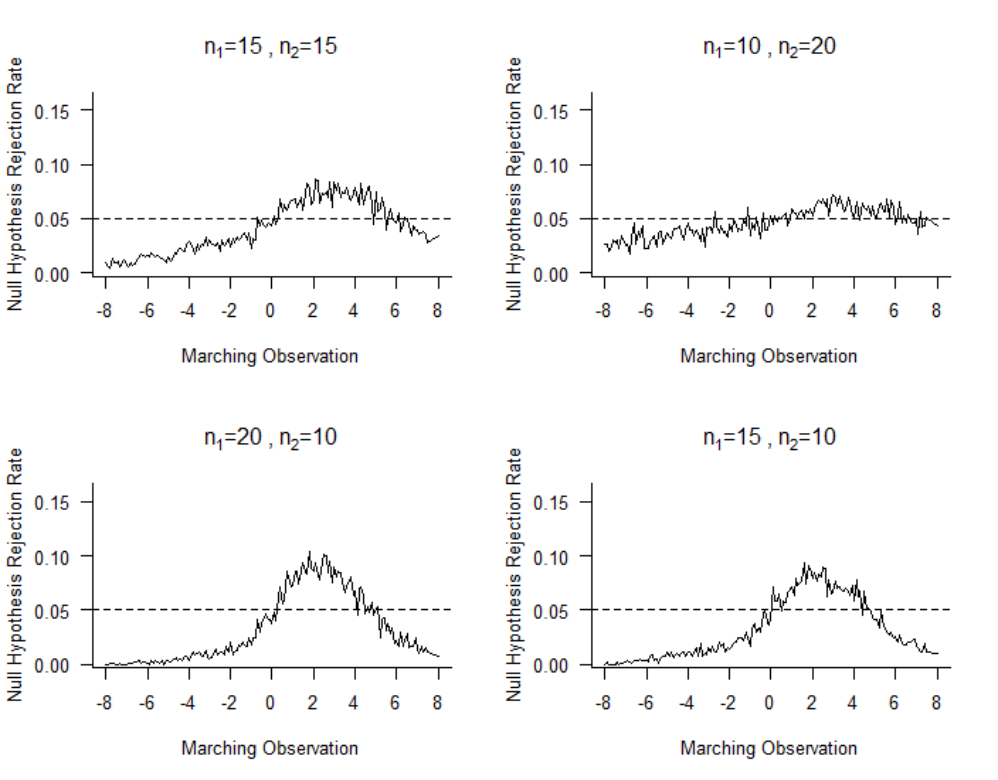
Sample sizes: 10, 15, 20

10,000 iterations.

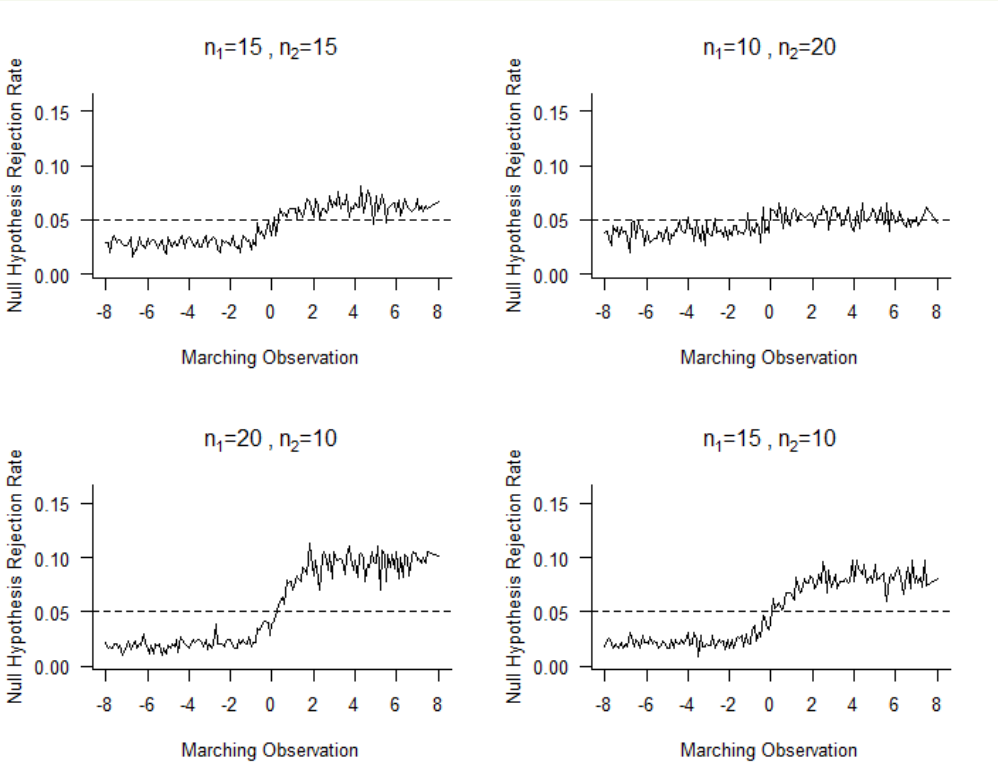
Results, Equal variances -Independent samples t-test



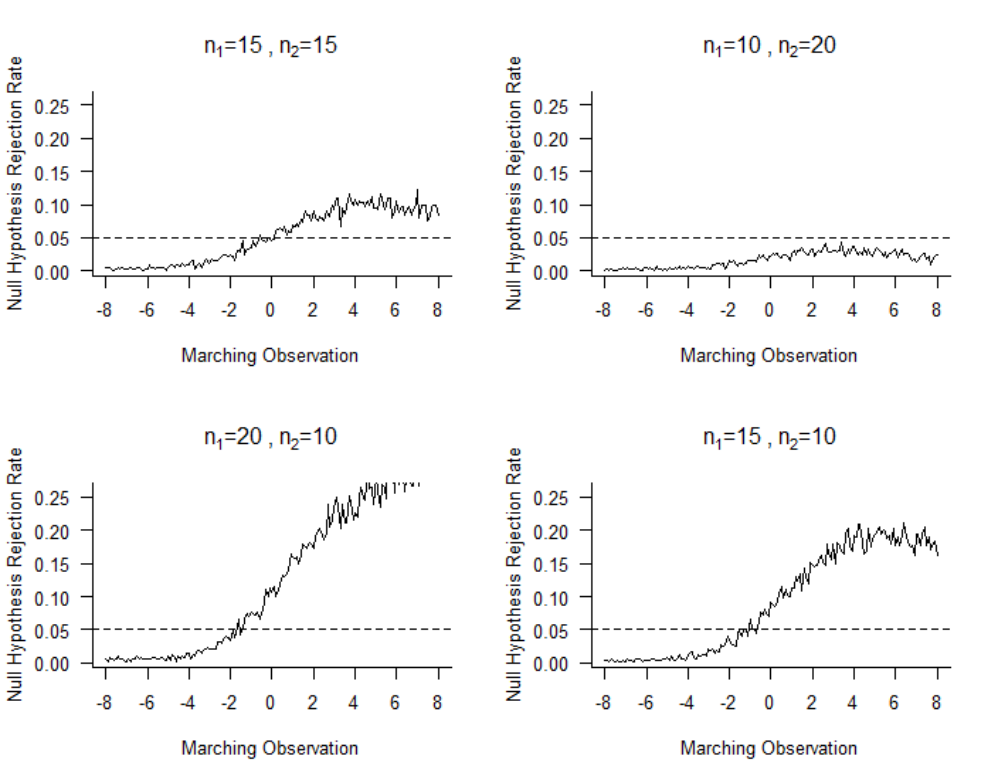
Results, Equal variances -Welch's test



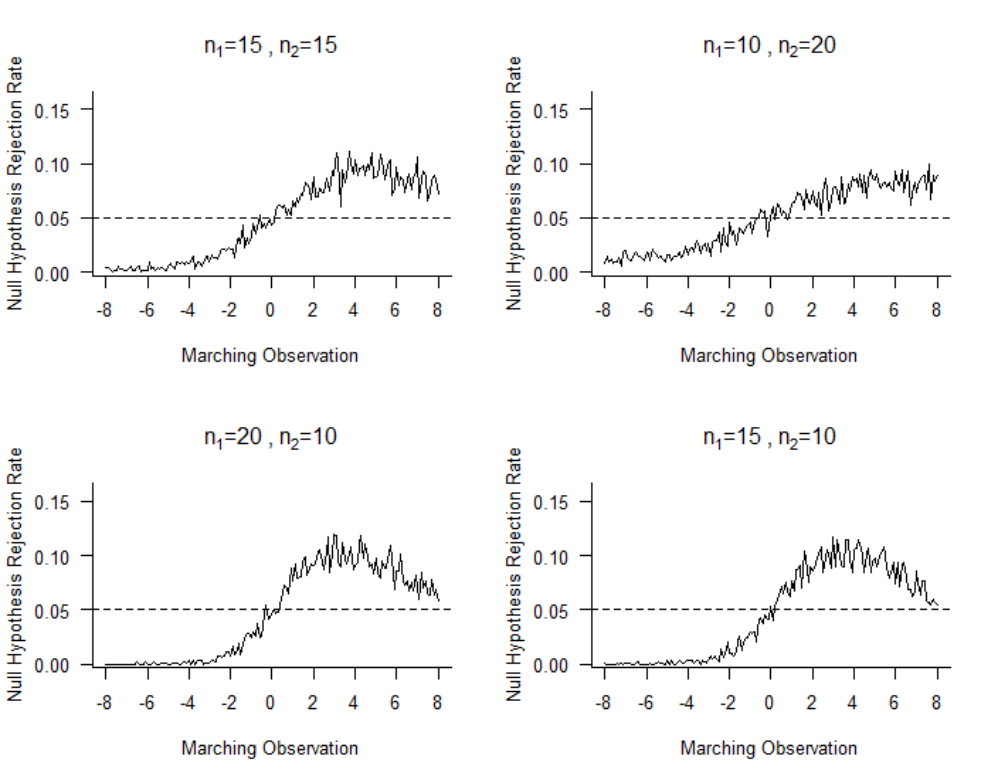
Results, Equal variances – Mann Whitney test



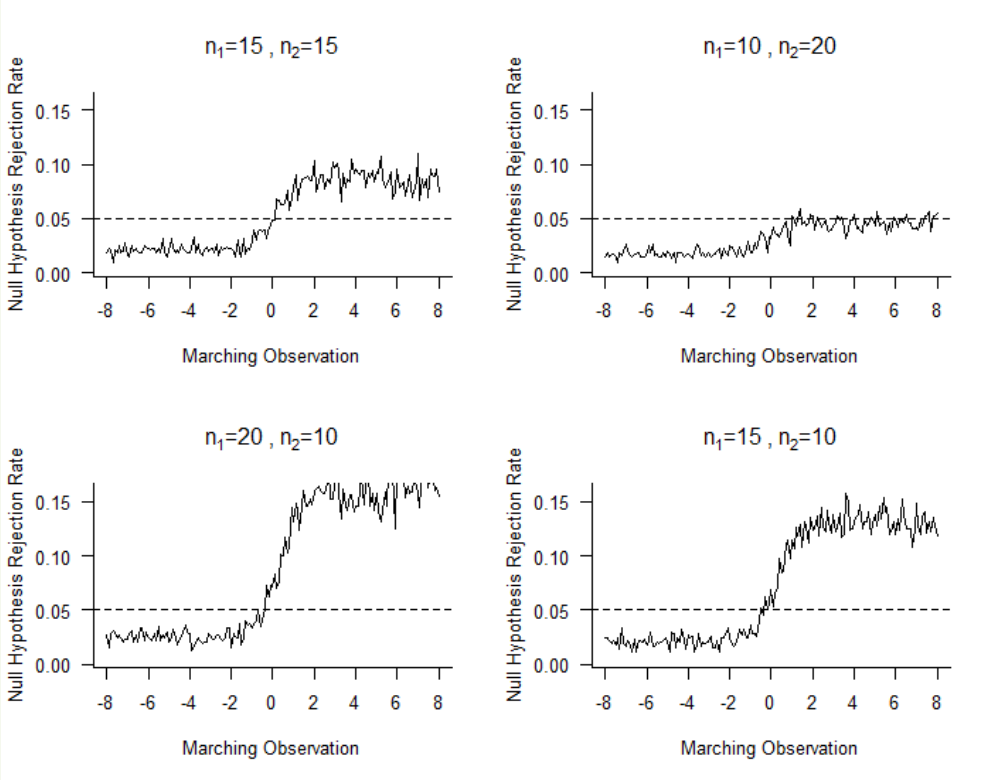
Sample 2 larger variance – Independent samples t-test



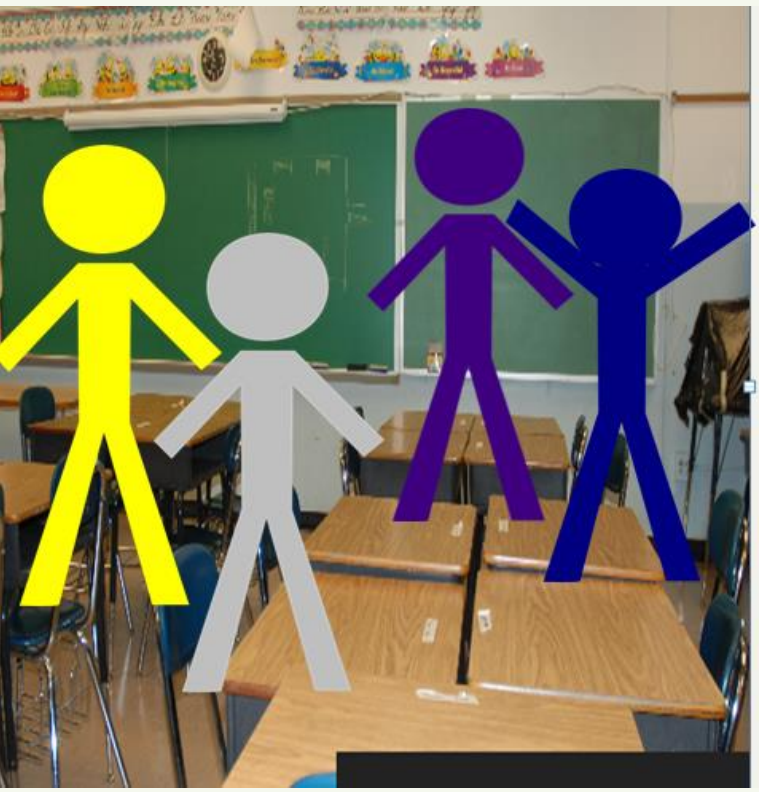
Sample 2 larger variance – Welch's test

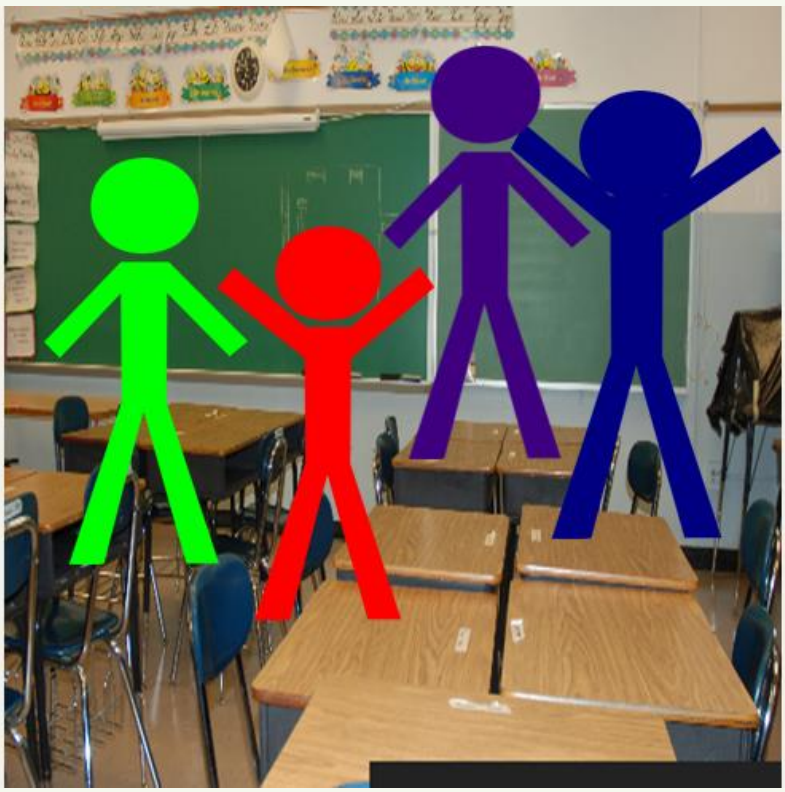


Sample 2 larger variance – Mann Whitney test



Extension:
Partially overlapping samples





Extension: Simulation methodology

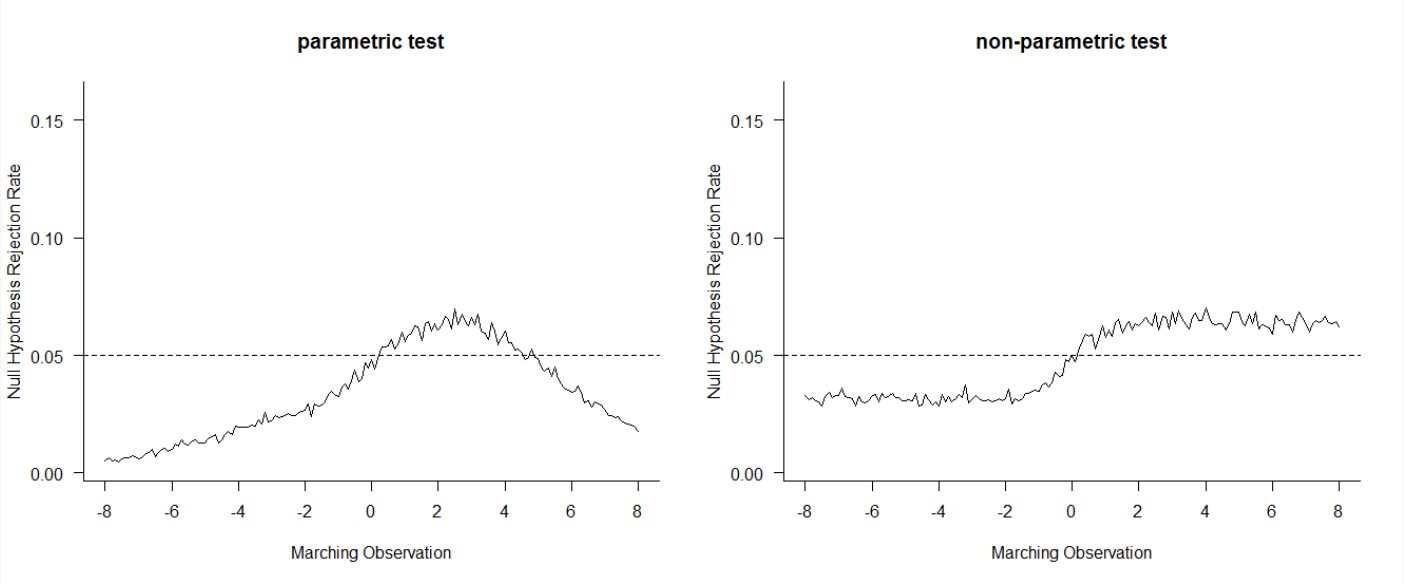
Paired samples design with observations deleted (MCAR)

Distribution: N(0,1) N(0,4)

‘marching observation’, additional observation within the sample, from -8 to 8 (increments of 1).

- Partially overlapping samples t-test
- Ranks applied to partially overlapping t-statistic

Partially overlapping samples, Equal sample size, equal variance



Conclusion

- extreme observation paradox, mask true effects or show phantom significant effects
- parametric tests, no outliers assumption of test
- Outlier detection methods subjective, decision not to remove ‘outlier’ should be considered with same vigour as decision to remove

Supporting material

Derrick, B. (2018). An outlier in an independent samples design. In: Royal Statistical Society, International Conference, 2018. Available from: http://eprints.uwe.ac.uk

References

Derrick, B., Broad, A., Toher, D. and White, P. (2017) The impact of an extreme observation in a paired samples design. Metodološki Zvezki - Advances in Methodology and Statistics, 14 (2). pp. 1-17.

Derrick, B., Russ, B., Toher, D. and White, P. (2017) Test statistics for the comparison of means for two samples which include both paired observations and independent observations. Journal of Modern Applied Statistical Methods, 16 (1). pp. 137-157.