# The impact of an extreme observation in a paired samples design

**Abstract**

The effect of systematically altering the value of a single observation within a paired differences design is considered. A paradox is observed for the paired samples t-test, where increasing the value of an observation in the direction of the true mean difference results in a higher p-value. Using simulation, deviations from robustness of the paired samples t-test is demonstrated, and is contrasted with Yuen's paired samples t-test and the Wilcoxon signed rank sum test.

## 1 Introduction

The paired samples t-test is logically and numerically equivalent to the one sample t-test performed on paired differences, and it is one of the most well-established and commonly performed statistical tests. Zimmerman (1997) demonstrated that the Type I error rate of the paired samples t-test remains close to the nominal significance level for varying correlation and sample sizes under normality. Under less idealised conditions, Posten (1979), Herrendörfer et al, (1983), Rasch and Guiard (2004), and Fradette et al. (2003) found that the paired samples t-test maintains Type I error robustness for a range of non-normal distributions. However, Blair and Higgins (1985) found the Wilcoxon signed rank sum test to also be Type I error robust and to have some power advantages over the paired samples t-test for a range of non-normal distributions. Chaffin and Rhiel (1993) demonstrated that the tails of the sampling distribution of the paired samples test statistic are skewness dependent, particularly with relatively small sample sizes.

Zumbo and Jennings (2002), using a novel contamination model, determined the effect of outliers on the validity and power of the paired samples t-test. They found the paired samples t-test to have robust validity for symmetric contamination, but with increasing inflation of the Type I error rate with increasing asymmetric contamination. This is coupled with degradation in power in the presence of outliers when the true effect is small and sample sizes are small. In their work the number of outliers in the sample is considered to be a random variable.

One of the assumptions of the paired samples t-test is that the differences between the two samples are normally distributed, or alternatively and in a practical sense, that the mean difference has a distribution which can reasonably be approximated by a Normal distribution. A closely related assumption is that there are no large outliers in the differences. When performing the paired samples t-test, there may be competition between the magnitude of the mean difference and the standard deviation of the differences. In particular, extreme observations within a dataset can distort the balance between these two elements of the test. To illustrate this, consider the example data in Table 1.

Table 1. Example data for six units within a paired design.

| Pair | Sample 1 | Sample 2 | Difference |
|---|---|---|---|
| 1 | 30 | 22 | 8 |
| 2 | 28 | 18 | 10 |
| 3 | 45 | 45 | 0 |
| 4 | 57 | 54 | 3 |
| 5 | 38 | 32 | 6 |
| 6 | 37 | 37 - X | X |

For the first five pairs, the mean of Sample 1 is greater than or equal to the mean of Sample 2. For the sixth pair, let the difference between the Sample 1 observation and the Sample 2 observation be denoted as X. Intuition might suggest that a positive value of X may contribute towards an overall significant difference in means being observed. If this were the case, a large positive value of X should seemingly contribute towards a significant effect. In the following, the value of X is systematically altered in order to demonstrate its impact on the paired samples t-test. The observation X will “march” through the data set and will be colloquially referred to as a marching observation. Table 2 shows the results of a two-sided paired samples t-test for negative values of X through to large positive values of X.

Table 2 Output for the paired samples t-test on 5 degrees of freedom for increasing values of X.

| X | t | p-value | X | t | p-value | X | t | p-value |
|---|---|---|---|---|---|---|---|---|
| -3 | 1.984 | 0.104 | ***11*** | ***3.670*** | ***0.014*** | 25 | 2.425 | 0.060 |
| -1 | 2.406 | 0.061 | ***13*** | ***3.461*** | ***0.018*** | 27 | 2.319 | 0.068 |
| ***1*** | ***2.870*** | ***0.035*** | ***15*** | ***3.240*** | ***0.023*** | 29 | 2.226 | 0.077 |
| ***3*** | ***3.321*** | ***0.021*** | ***17*** | ***3.033*** | ***0.029*** | 31 | 2.145 | 0.085 |
| ***5*** | ***3.671*** | ***0.014*** | ***19*** | ***2.848*** | ***0.036*** | 33 | 2.073 | 0.093 |
| ***7*** | ***3.840*** | ***0.012*** | ***21*** | ***2.687*** | ***0.043*** | 35 | 2.009 | 0.101 |
| **9** | ***3.820*** | ***0.012*** | 23 | 2.546 | 0.052 | 37 | 1.953 | 0.108 |

The values of X for which the null hypothesis of equal means is rejected at the 5% significance level are italicised and highlighted bold in Table 2. For low values of X it can be seen that as the value of X increases the p-value decreases. In this example, as the value of X increases beyond approximately 8, the p-value increases. As the value of the observed difference in the sixth pair increases (and hence as the mean difference increases), the p-value also increases. Observing an extreme value of X in the direction of the seemingly observed effect can increase the sample variance to such an extent that it impedes the test from giving a significant result. The extreme observation paradox is the contrariwise p-value increase as the value of an extreme observation increases in the direction of the overall effect.

As the absolute value of the marching observation increases, the assumptions of the paired samples t-test are increasingly violated. When the sample size is small or the assumptions of the paired samples t-test are violated, researchers often choose to perform the Wilcoxon signed rank sum test. Aguinis et al., (2013) summarise a comprehensive list of techniques for dealing with outliers and state that non-parametric tests give results that are robust in the presence of outliers. However, Zimmerman (2011) indicates that rank based methods do not necessarily eliminate the influence of outliers. Another alternative approach when outliers are present is to use Yuen's paired samples t-test. In this test, the principles of trimmed means outlined by Yuen (1974), are applied to the paired differences (Wilcox, 2005).

In this paper, simulation is used to explore the scenarios in which the extreme observation paradox is observed in a paired samples design. We are particularly interested in isolating those situations when two-sided hypothesis testing is undertaken (e.g. see Ringwalt et al., 2011), when sample sizes are relatively small (i.e. when outliers may have a greater effect on the paired samples t-test). The concept of a systematically marching observation similar to the demonstration in Table 2, is used to investigate the effects of an aberrant observation. In the simulation design, this aberrant observation is a forced additional observation not fitting with the simulated data, and is not due to inherent variability. Simulations are performed for an aberrant observation in the direction of the effect suggested by the rest of the sample, and secondly where an aberrant observation is in the opposing direction of the effect suggested by the rest of the sample. Thus situations where the sign of the marching observation is concordant or discordant with the mean of the other observations are considered. For comparative purposes, the paired samples t-test, the Wilcoxon signed rank sum test, and Yuen's paired samples t-test are included.

Null hypothesis significance testing is most frequently performed with a nil-null hypothesis specifying that no difference between groups is present, and a two directional alternative (Levine et al., 2008). Therefore the impact of an extreme observation for a two-sided test is the main emphasis of this paper. However, one-sided tests retain some practical utility, and the simulations are extended to a one-sided test.

We hypothesise that the seemingly paradoxical behaviour exhibited in Table 2 will be a feature of the paired samples t-test in general. In contrast we hypothesise that Yuen's paired samples t-test and the Wilcoxon signed rank sum test will be robust to a single aberrant observation.

In order to gain insight, we firstly investigate the mathematical limiting forms of each of the three test statistics under consideration as a single marching observation becomes increasingly large compared with the rest of the sample, and then proceed to a simulation investigation.

# 2 An unbounded marching observation

For development purposes consider a random sample $X_1$, $X_2$, $\cdots \cdots$, $X_{n-1}$, $X_n$, and let $X_{(1)} < X_{(2)} < \cdots \cdots X_{(n-1)} < X_{(n)}$ denote the order statistics. Further, let $X_k = Y_k$ for ($k$ = 1, 2, $\cdots$, $n-1$), let $Y_{(1)} < Y_{(2)} < \cdots \cdots Y_{(n-1)}$ be the corresponding order statistics, and let $X_n = \xi$ be the marching observation. In this notation, $Y_k$ ($k$ = 1, …, $n$-1) denotes the observations prior to the inclusion of the marching observation.

The following analytical exposition investigates the behaviour of the one sample t-test, Yuen's paired samples t- test, and the Wilcoxon signed rank sum test, as the marching observation $X_n = \xi$ becomes relatively large compared with the rest of the sample.

### 2.1 The t-test

Consider the single sample t-test test statistic on the paired differences, used to test $H_0 : \mu_X = 0$, defined by

$$T := \frac{\overline{X}}{\hat{\sigma}_{X+}} \sqrt{n}$$

where $\overline{X} := \dfrac{X_1 + X_2 + \ldots X_n}{n}$ and $\hat{\sigma}_{X+} := \sqrt{\dfrac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \ldots + (X_n - \overline{X})^2}{n-1}}$

Observe that: $\overline{X} = \frac{(n-1)\overline{Y}+\xi}{n}$, $\overline{X}-\overline{Y} = \frac{\xi-\overline{Y}}{n}$, $\xi-\overline{X} = \frac{(n-1)(\xi-\overline{Y})}{n}$.

Thus, $\hat{\sigma}_{X+} = \sqrt{\frac{(Y_1-\overline{X})^2+(Y_2-\overline{X})^2+\ldots\ldots+(Y_{n-1}-\overline{X})^2+(\xi-\overline{X})^2}{n-1}}$

Note that $\sum_{j=1}^{n-1}\left(Y_j-\overline{X}\right)^2 = \sum_{j=1}^{n-1}\left(Y_j-\overline{Y}+\overline{Y}-\overline{X}\right)^2$

$$= \sum_{j=1}^{n-1}\left(Y_j-\overline{Y}\right)^2+(n-1)(\overline{Y}-\overline{X})^2+2(\overline{Y}-\overline{X})\sum_{j=1}^{n-1}\left(Y_j-\overline{Y}\right)$$

$$=(Y_1-\overline{Y})^2+(Y_2-\overline{Y})^2+\ldots+(Y_{n-1}-\overline{Y})^2+(n-1)(\overline{Y}-\overline{X})^2+0$$

Hence, $\hat{\sigma}_{X+} = \sqrt{\frac{(Y_1-\overline{Y})^2+(Y_2-\overline{Y})^2+\ldots+(Y_{n-1}-\overline{Y})^2+(\xi-\overline{X})^2}{n-1}+(\overline{X}-\overline{Y})^2}$

For the $n-1$ values, define $\hat{\sigma}_Y := \sqrt{\frac{(Y_1-\overline{Y})^2+(Y_2-\overline{Y})^2+\ldots+(Y_{n-1}-\overline{Y})^2}{n-1}}$

Note that $\hat{\sigma}_Y$ does not have the "+" symbol, meaning that the marching observation is not included. An alternative definition for $\hat{\sigma}_Y$ could have $n-2$ in the denominator.

and so, $\hat{\sigma}_{X+} = \sqrt{\hat{\sigma}_Y^2+\frac{(\xi-\overline{X})^2}{n-1}+(\overline{X}-\overline{Y})^2}$

$$=\sqrt{\hat{\sigma}_Y^2+\frac{(n-1)^2}{n^2}\frac{(\xi-\overline{Y})^2}{(n-1)}+\frac{(\xi-\overline{Y})^2}{n^2}}$$

$$=\sqrt{\hat{\sigma}_Y^2+\frac{(\xi-\overline{Y})^2}{n}}$$

and hence $T = \frac{(n-1)\overline{Y}+\xi}{\sqrt{n\hat{\sigma}_Y^2+(\xi-\overline{Y})^2}}$.

It can be seen that as $\xi\to\infty$, $T\to 1$, and similarly as $\xi\to-\infty$, $T\to -1$. Accordingly, for any value of significance level likely to be encountered in practice the results $\xi\to\pm\infty$, $T\to\pm 1$ indicate that the null hypothesis would not be rejected under the stated conditions.

### 2.2 Yuen's paired samples t-test

Let $\gamma$ denote the per tail proportion of trimming, let $e := \lfloor \gamma n \rfloor$ and let $f := n - 2e$. Define the trimmed sample $X_{t1}$, $X_{t2}$, $\cdots\cdots$, $X_{tf-1}$, $X_{tf}$ as $X_{tk} := X_{(k+e)}$ ($k = 1, 2, \cdots, f$) and define the winsorised sample $X_{w1}$, $X_{w2}$, $\cdots\cdots$, $X_{wf}$ as

$$X_{wk} := \begin{cases} X_{(e+1)} & ; \ k = 1, 2, \cdots\cdots, e \\ X_{(k)} & ; \ k = e+1, e+2, \cdots\cdots, n-e \\ X_{(n-e)} & ; \ k = n-e+1, n-e+2, \cdots\cdots, n \end{cases}$$

Let $\overline{X}_t := \sum_{k=1}^{f} X_{tk} / f$, and $\overline{X}_w = \sum_{k=1}^{n} X_{wk} / n$ define the trimmed mean and winsorised mean respectively and let, $\hat{\sigma}^2_{Xw+} = \sum_{k=1}^{n} (X_{wk} - \overline{X}_w)^2 / (n-1)$ denote the winsorised variance. In this notation, Yuen's test statistic is given by $T_Y := \dfrac{\overline{X}_t}{\hat{\sigma}_{Xw+}} \sqrt{n}\ (1-2\gamma)$.

For $\xi < Y_{(e)}$

$$\overline{X}_t = \frac{Y_{(e)} + Y_{(e+1)} + Y_{(e+2)} + \cdots\cdots + Y_{(n-e-1)}}{f},$$

$$\overline{X}_w = \frac{eY_{(e)} + Y_{(e)} + Y_{(e+1)} + Y_{(e+2)} + \cdots\cdots + Y_{(n-e-1)} + eY_{(n-e-1)}}{n}, \text{ and}$$

$$\hat{\sigma}^2_{Xw+} = \frac{e(Y_{(e)} - \overline{X}_w)^2 + \sum_{k=e}^{n-e-1} (Y_{(k)} - \overline{X}_w)^2 + e\,(Y_{(n-e-1)} - \overline{X}_w)^2}{n-1}$$

For fixed values, $Y_1, Y_2, \cdots\cdots, Y_{n-1}$, as $\xi \to -\infty$, $T_Y := \dfrac{\overline{X}_t}{\hat{\sigma}_{Xw+}} \sqrt{n}\ (1-2\gamma)$ stabilises to some limiting value.

Similarly, for $\xi > Y_{(n-e)}$

$$\overline{X}_t = \frac{Y_{(e+1)} + Y_{(e+2)} + Y_{(e+2)} + \cdots\cdots + Y_{(n-e)}}{f},$$

$$\overline{X}_w = \frac{eY_{(e+1)} + Y_{(e+1)} + Y_{(e+2)} + Y_{(e+2)} + \cdots\cdots + Y_{(n-e)} + eY_{(n-e)}}{n}, \text{ and}$$

$$\hat{\sigma}^2_{Xw+} = \frac{e(Y_{(e+1)} - \overline{X}_w)^2 + \sum_{k=e+1}^{n-e}(Y_{(k)} - \overline{X}_w)^2 + e\,(Y_{(n-e)} - \overline{X}_w)^2}{n-1}$$

For fixed values, $Y_1,\ Y_2, \cdots\cdots, Y_{n-1}$ as $\xi \to -\infty$ , $T_Y := \dfrac{\overline{X}_t}{\hat{\sigma}_{Xw+}}\sqrt{n}\ (1-2\gamma)$ stabilises to some limiting value. Moreover, for a sufficiently large sample, the limit values for both directions of the marching observation should be close to each other. Hence the properties displayed as $\xi \to -\infty$ or $\xi \to -\infty$ are consistent with $T_Y$ being a robust test statistic.

### 2.3 The Wilcoxon signed rank sum test

Assuming no ties and no zero observations, then the test statistic for the Wilcoxon signed rank sum test, $W$, is defined as

$$W = R_1^X \operatorname{sgn}(X_1) + R_2^X \operatorname{sgn}(X_2) + \ldots + R_n^X \operatorname{sgn}(X_n)$$

where $R_k^X$ is the rank of $|X_k|$ among $|X_1|, |X_2|, \ldots, |X_n|$. If $X_1,\ X_2, \cdots\cdots,\ X_n$ are independent and follow the same symmetric continuous distribution, then $W$ follows a distribution with mean 0 and variance $n(n+1)(2n+1)/6$.

Denote by $R_k^Y$ the rank of $|Y_k|$ among $|Y_1|, |Y_2|, \ldots, |Y_{n-1}|$.
For $|\xi| > \max\{|Y_1|, |Y_2|, \cdots\cdots |Y_{n-1}|\}$,
$W = R_1^Y \operatorname{sgn}(Y_1) + R_2^Y \operatorname{sgn}(Y_2) + \ldots + R_n^Y \operatorname{sgn}(Y_n) + n \operatorname{sgn}(\xi)$.

Hence, under the stated conditions, for fixed values, $Y_1, Y_2, \cdots\cdots, Y_{n-1}$, the Wilcoxon signed rank sum statistic stabilises to some situation dependent limit value as $\xi \to +\infty$, and to some situation dependent limit value as $\xi \to -\infty$. The difference between these two values is $n - (-n) = 2n$, and the standardised values would differ by $\sqrt{24n/\{(n+1)(2n+1)\}}$. These are close to each other for sufficiently large $n$.

# 3 Simulation Methodology

The approach is to generate sample data meeting the assumptions of the paired samples t-test, and to then include an additional observation in the sample. This additional observation systematically changes in its observed value. The paired samples t-test, the Wilcoxon signed rank test, and Yuen's paired samples t-test, are performed for a two-sided nil-null hypothesis. Under a two-sided nil-null hypothesis; the paired samples t-test is used to test a distribution mean difference of zero; and Yuen's paired samples t-test is used to test the distribution of the trimmed mean equal to zero. Historically, the derivation of the Wilcoxon rank sum distribution has been made for continuous random variables under a null hypothesis of no distributional differences, and is sensitive to changes in central location (Gibbons and Chakraborti, 2011).

Within the simulation, the differences are generated rather than the paired observations themselves. Specifically, $n-1$ random Normal deviates $x_1, x_2, \ldots, x_{n-1}$ are generated using the Box–Muller (1958) transformation, where $n$ represents the sample size of the paired differences. Under $H_0$, the $n-1$ random Normal deviates have a population mean of zero ($\mu = 0$) and a standard deviation of one ($\sigma = 1$).

To isolate the phenomenon and behaviour of interest, if $\bar{x}_{n-1} = \sum_{i=1}^{n-1} x_i / (n-1) < 0$ then $x_1, x_2, \ldots, x_{n-1}$ 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.)

Under $H_1$, for each of the $n-1$ deviates, a constant $d$ is added to each of the values. The simulations are performed under normality so that the data fulfil the assumptions of the test with the exception of an aberrant observation.

An additional observation, $x_n$, is added to the $n-1$ observations to give a total sample size of $n$. For any simulated sample, the value of $x_n$ is systematically varied from -8 to 8 in increments of 0.1. It is this value, $x_n$, which is referred to as the 'marching observation'. The values of $x_n$ approximately range between +/- 8 standard deviations from the mean and would therefore cover limits likely encountered in a practical environment. Note that the condition of $\bar{x}_{n-1} > 0$ is to ensure that the concordance of effects ($\bar{x}_{n-1} > 0,\ x_n > 0$) or discordance of effects ($\bar{x}_{n-1} > 0,\ x_n < 0$) can be established.

A summary of the values of $n$, $x_n$ and $d$ used in the full factorial simulation design is given in Table 3. The simulation is run 10,000 times for each combination of sample size and mean difference.

In a second set of simulations, the impact of the marching observation is similarly assessed, removing the condition that the mean sample difference is positive, and performing a one-sided test. This is done as per the parameter combinations in Table 3 using upper tail critical values.

Table 3. Summary of simulation design.

| | |
|---|---|
| Sample size, $n$ | 10, 15, 20, 25 |
| Marching observation, $x_n$ | -8:8 (0.1) |
| Mean difference, $d$ | 0, 0.5 |
| Significance level | 5% |
| Number of Iterations | 10,000 |
| Programming Language | R version 3.1.3 |

For the paired samples t-test and the Wilcoxon signed rank sum test, the default 'stats' package in R is used. Yuen's paired samples t-test is performed using the R package 'PairedData' as outlined by Wilcox (2005). 10% trimming per tail is performed.

The proportion of the 10,000 iterations where the null hypothesis is rejected is calculated at the nominal significance level of 5%. This gives the Null Hypothesis Rejection Rate (NHRR). Note that the terminology NHRR is used and not Type I error rate, because the inclusion of the marching observation would strictly invalidate the underpinning assumptions of the resultant test. The effect of gradually increasing the marching observation is to gradually violate the assumption of the nil-null hypothesis.

The research question being asked is "How is the performance of the paired samples t-test, Yuen's paired samples t-test, and the Wilcoxon signed rank sum test affected by the presence of an aberrant observation?"

# 4 Results

The Null Hypothesis Rejection Rate (NHRR) is assessed for each of the three statistical tests under consideration for a two-sided test, firstly when $d = 0$ and secondly in the presence of a systematic effect size ($d = 0.5$).

Figure 1 gives the NHRR of the paired samples t-test when $d = 0$, using the nominal significance level of 5%.

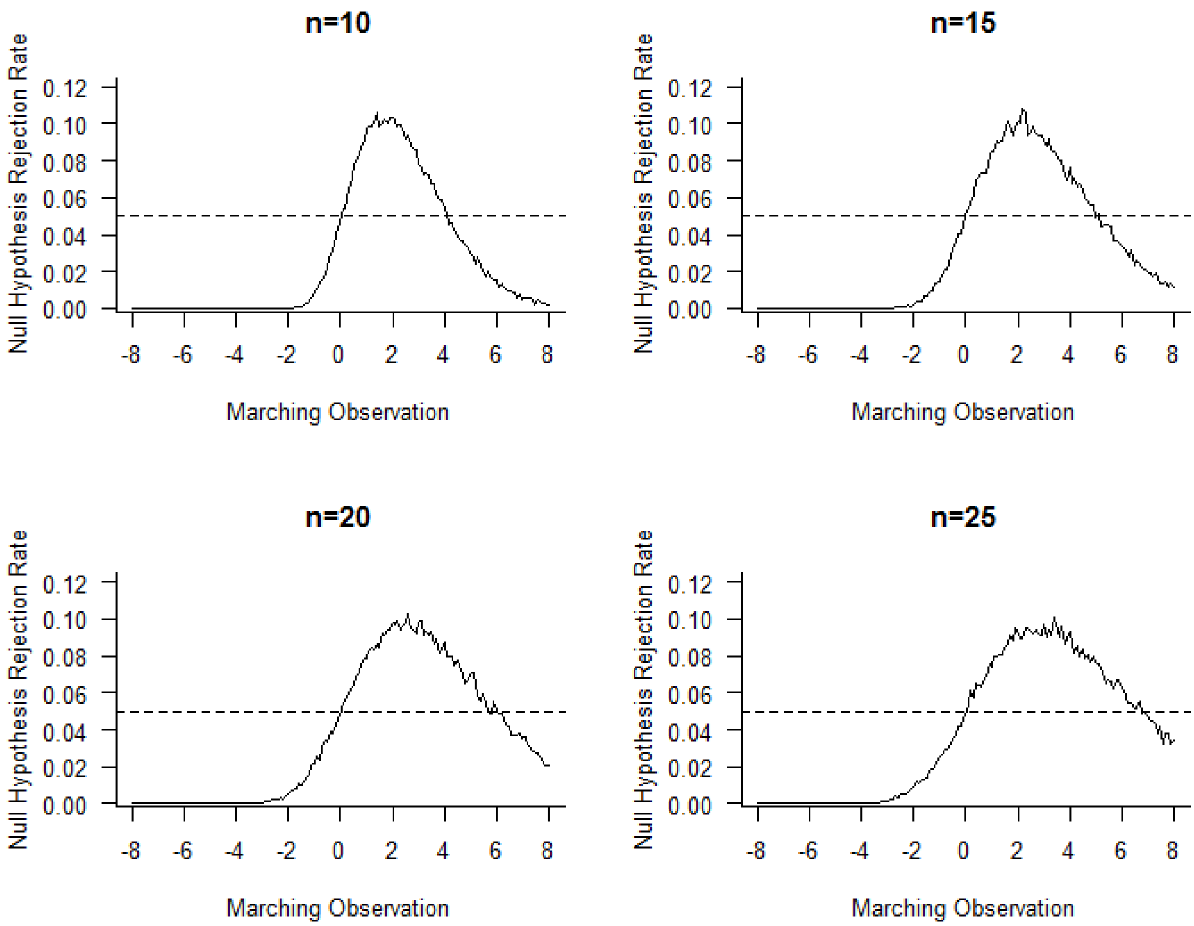


Figure 1. NHRR of the paired samples t-test, $d = 0$, two-sided.

Figure 1 shows that when the value of $x_n = d = 0$, the NHRR is approximately equal to the nominal Type I error rate of 5%. For positive sample means, as the value of $x_n$ starts to increase above zero, the paired samples t-test has an increasingly higher NHRR until a turning point is reached and with a subsequent return to the nominal Type I error rate. Extreme and increasingly larger values of the marching observation, $x_n$, in the direction of the sample effect results in a progressively lower NHRR, with values noticeably lower than the nominal Type I error rate. These effects are replicated in all four sample sizes, but the effects are marginally less noticeable with increasing sample size. Figure 1 also shows that a large value for the marching observation in the opposite direction to the mean of the first n-1 observations, effectively results in a zero value for the NHRR. This effect is consistent with the asymptotic behaviour given in Section 2 and the findings alluded to in the example given in Table 2.

Figure 2 gives the NHRR of Yuen's paired samples t-test and Figure 3 gives the NHRR of the Wilcoxon signed rank sum test, both when $d = 0$.

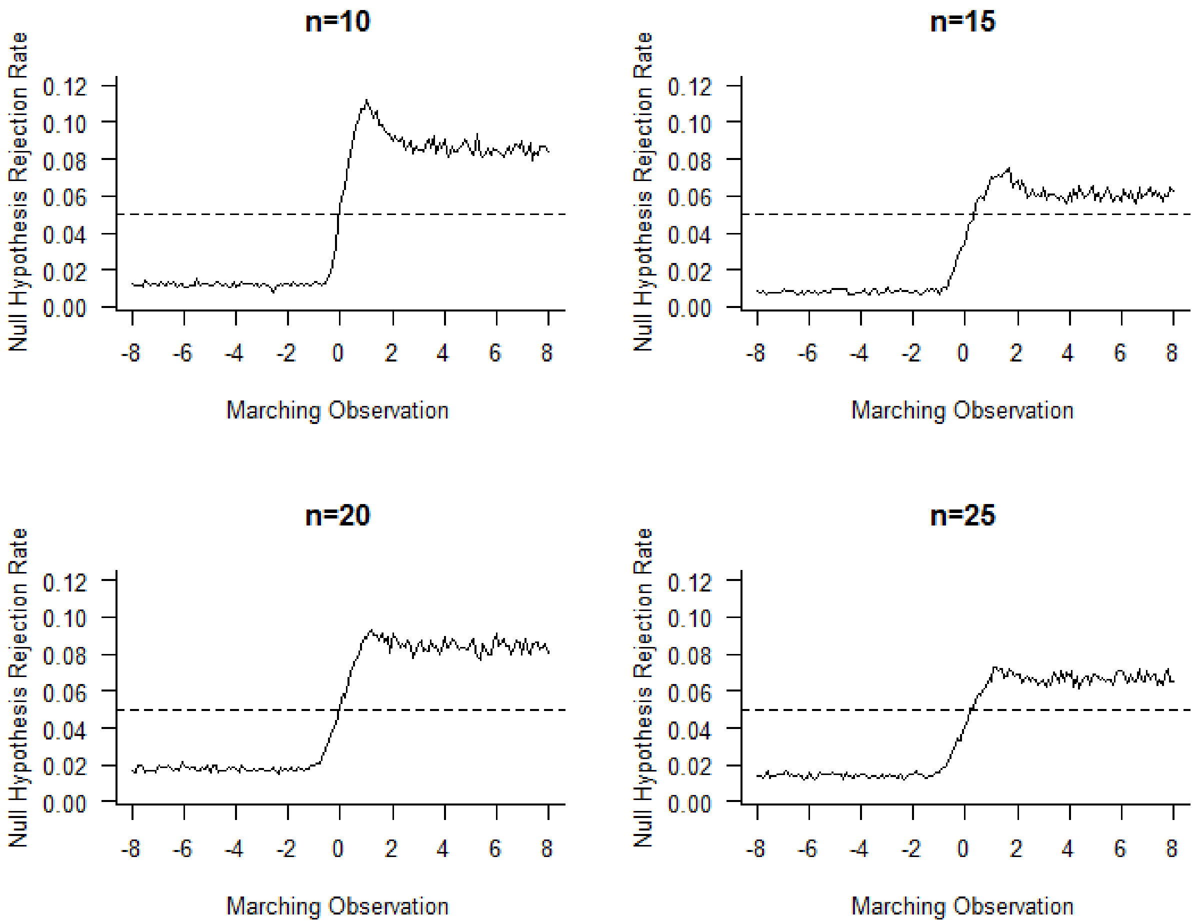


Figure 2. NHRR of Yuen's paired samples t-test, $d = 0$, two-sided.

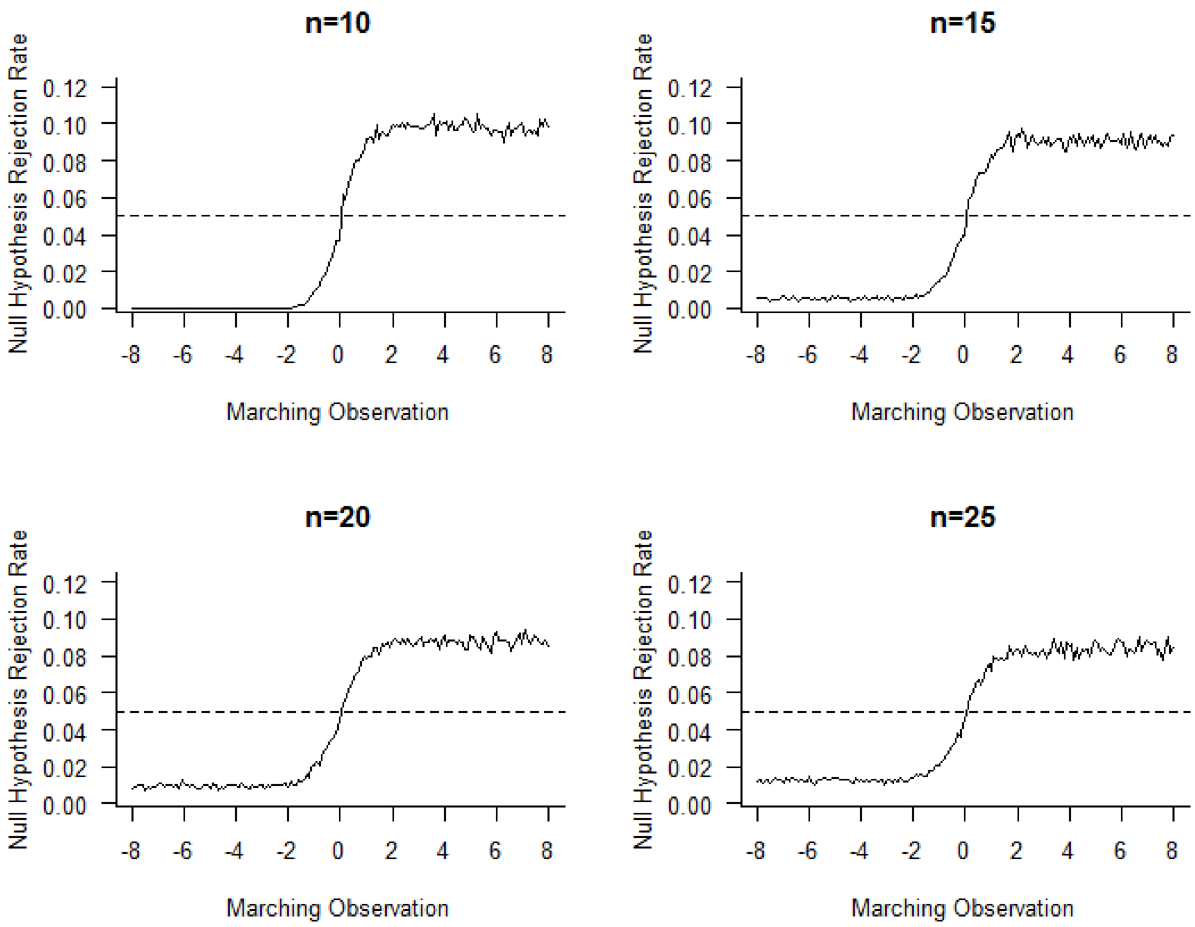


Figure 3. NHRR of the Wilcoxon signed rank sum test, $d = 0$, two-sided.

Figure 2 and Figure 3 show that when $x_n > 0$ and $\bar{x}_{n-1} > 0$, both Yuen's paired samples t-test and the Wilcoxon signed rank sum test result in the null hypothesis being rejected more frequently than the nominal significance level. Conversely, when $x_n < 0$ and $\bar{x}_{n-1} > 0$, both Yuen's paired samples t-test and the Wilcoxon signed rank sum test have a NHRR lower than the nominal significance level. These findings are entirely consistent with expectation for a robust test given the design of the simulation.

For the Wilcoxon signed rank sum test, due to the use of rank values, the test is not greatly affected by the magnitude of the extreme observation. Similarly due to the trimming, Yuen's paired samples t-test is not greatly affected by the magnitude of the extreme observation. The phenomenon of a turning point when $x_n > 0$ is not observed for either the Wilcoxon signed rank sum test or Yuen's paired samples t-test.

Figure 4 gives indicative power of the paired samples t-test, where $d = 0.5$. For a sample of size $n$ = 10 independent Normal deviates with $\mu = 0$ and $\sigma = 1$, the power of the test for the paired samples t-test for testing $H_0 : \mu = 0$ is 0.293. Under the same conditions, the power of the paired samples t-test for $n$ = 15, 20 and 25 is 0.438, 0.565, and 0.670 respectively. These reference lines are added to the graphics for comparative purposes.

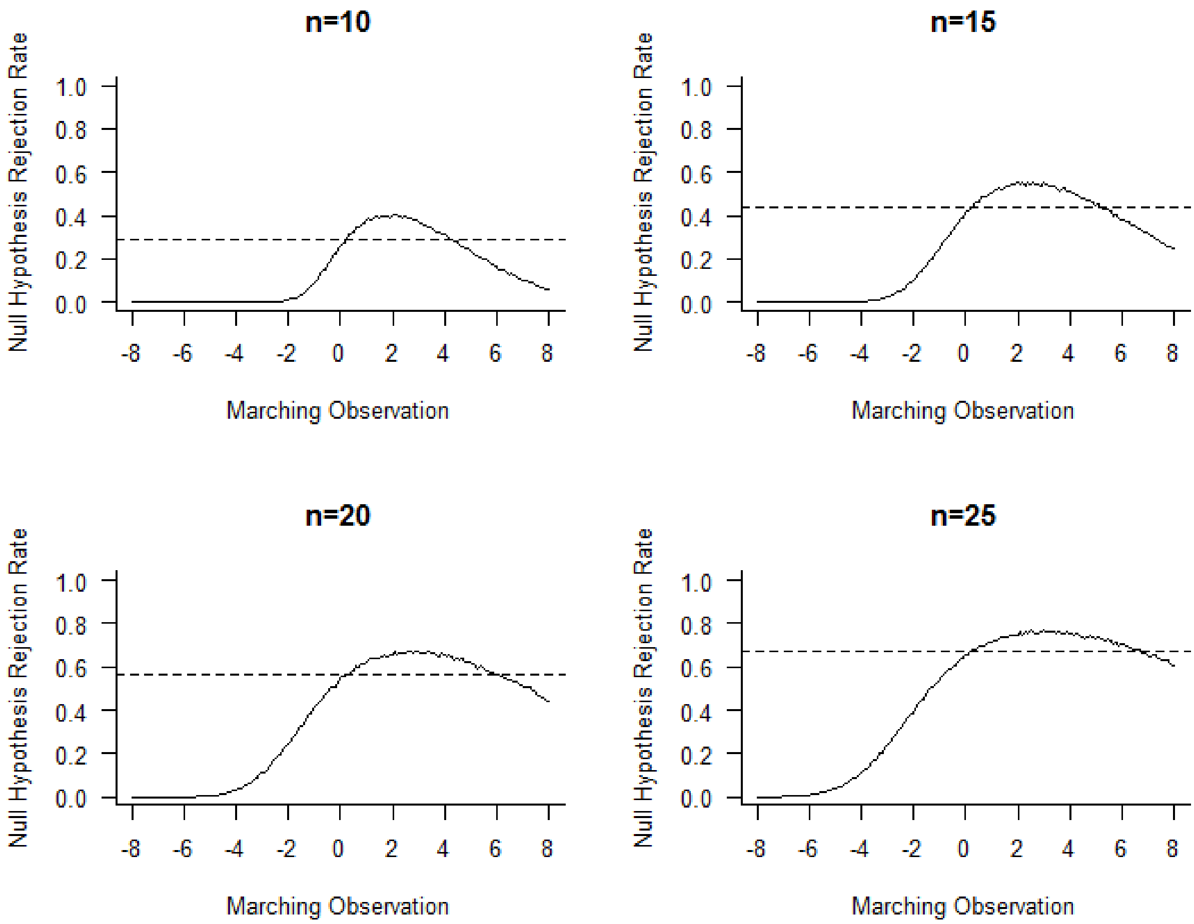


Figure 4. NHRR of the paired samples t-test, $d = 0.5$, two-sided.

Figure 4 shows that for $x_n > d$ = 0.5, increases in $x_n$ are initially associated with an increase in power. This power increase relative to the expected power for each of the sample sizes is clear to see but might not be of great practical consequence. In addition, there is a noticeable turning point at which the power decreases as $x_n$ further increases. For larger sample sizes, the paired samples t-test is relatively more robust to the presence of an extreme observation. For smaller sample sizes, the power reduction when an extreme observation is present is exacerbated. When the marching observation is in the opposite direction to the true effect, an increasingly large negative difference eliminates the effect under the stated conditions.

Figure 5 gives the NHRR of Yuen's paired samples t-test and Figure 6 gives the NHRR of the Wilcoxon signed rank sum test, both when $d = 0.5$. Under the same normality conditions, for $n$ = 10, 15, 20 and 25, the corresponding power for the Wilcoxon signed rank sum test is 0.279, 0.419, 0.543, and 0.648 respectively, and the corresponding power for the Yuen paired samples t-test is 0.263, 0.356, 0.528, and 0.613 respectively. These reference lines are added to the graphic for comparative purposes.

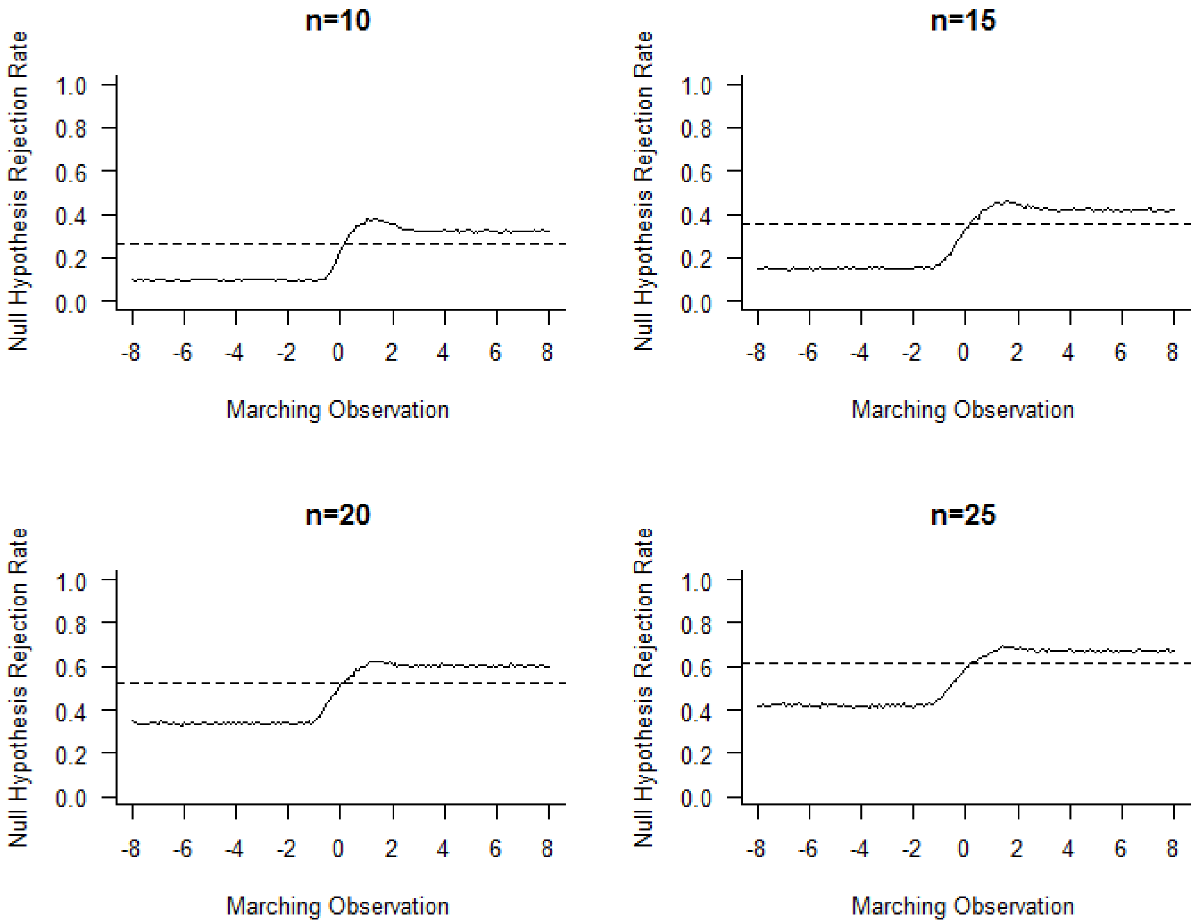


Figure 5. NHRR of Yuen's paired samples t-test, $d = 0.5$, two-sided

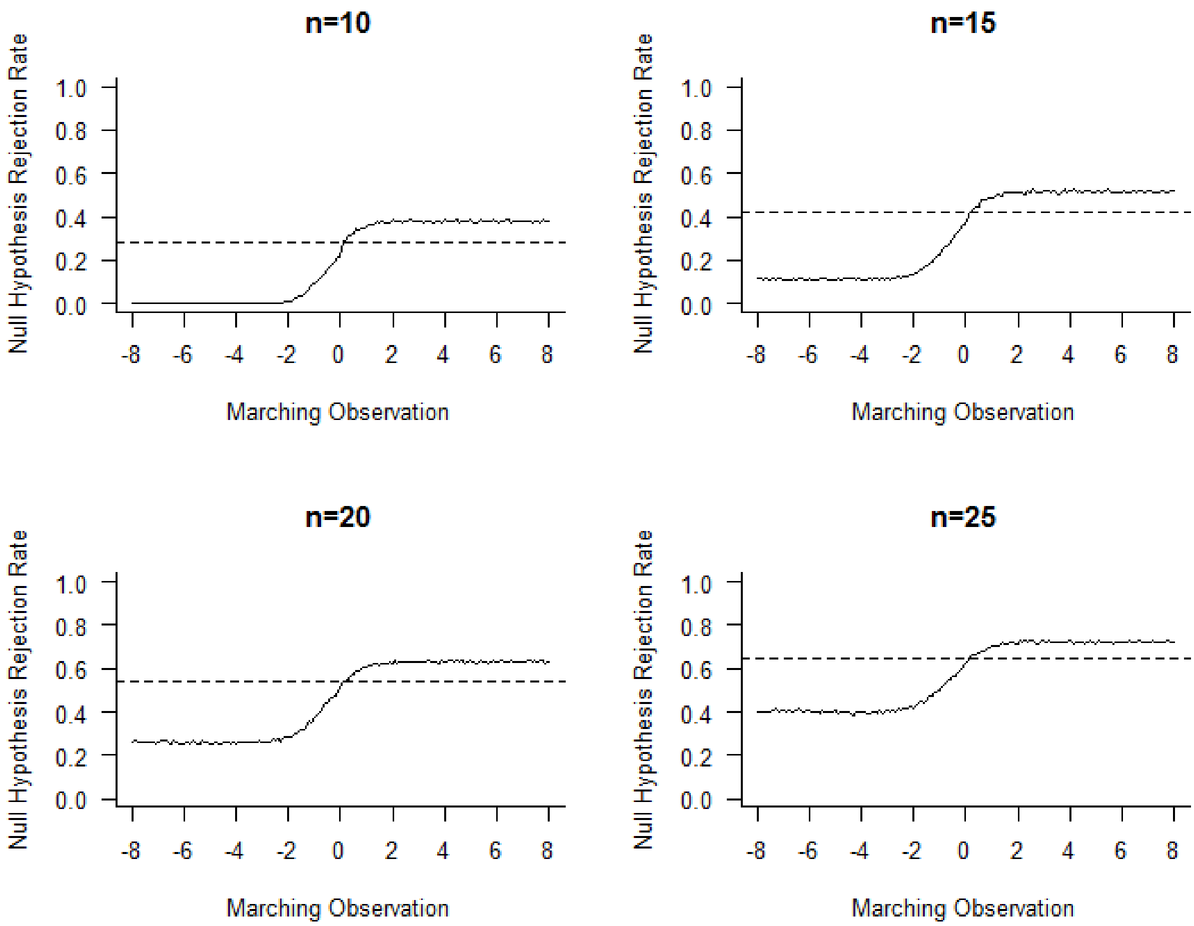


Figure 6. NHRR of the Wilcoxon test, $d = 0.5$, two-sided.

Figure 5 and 6 show that for $x_n > d = 0.5$, increases in $x_n$ are associated with an increase in power relative to the expected power for each of the sample sizes, but the increase might not be of great practical consequence. For small samples, when the marching observation is in the opposite direction to the true effect, an increasingly large negative marching observation reduces the effect and this is seen in the reduced power.

The second simulation set-up is now considered. The condition that the sample mean differences are positive is removed, and a one-sided test using the upper tail of the distribution is performed. Figure 7 shows the impact of the marching observation for each of the three tests when the null hypothesis is true.

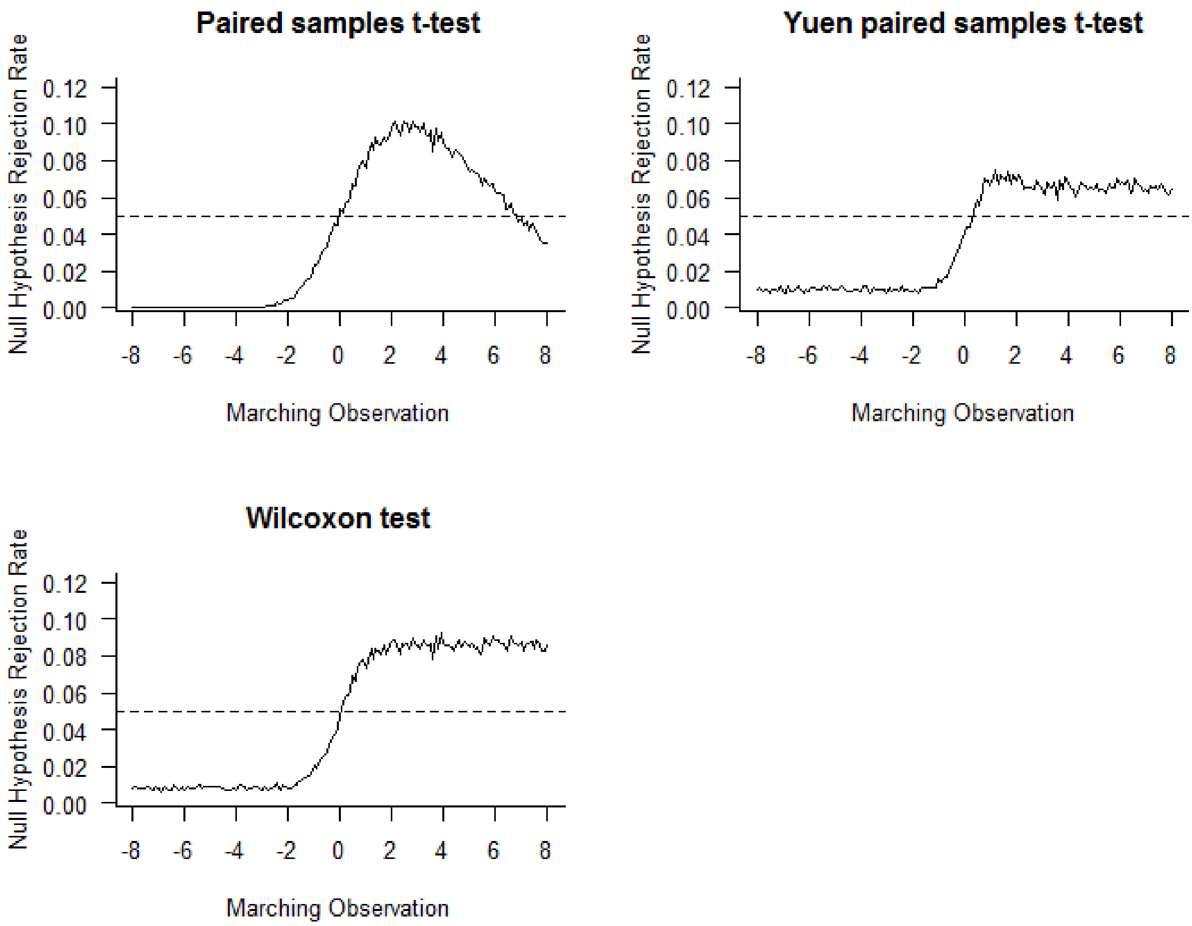


Figure 7 NHRR for each of the three tests when n = 15, $d = 0$, one sided.

Figure 7 demonstrates that the patterns observed and identifiable conclusions for the two-sided tests are the same under these conditions. In fact, the impact of the marching observation in the second simulation set-up is qualitatively similar to the first simulation set-up. For brevity, the remaining graphics under this condition are not displayed.

# 5 Discussion

We have used a systematically increasing marching observation to demonstrate the impact on the Null Hypothesis Rejection Rate (NHRR) for the paired samples t-test, Yuen's paired samples t-test, and the Wilcoxon signed rank sum test. This systematic approach, similar to one-factor at a time experimentation, would lend itself to other similar investigations e.g. two independent samples design, or to other single sample tests such as the single sample variance test, or be extended to investigations involving multiple marching observations. In practice, $x_n$ and the condition $\bar{x}_{n-1} > 0$ may be independent and the condition $\bar{x}_{n-1} > 0$ is imposed to separate potential different behaviours of the tests statistics.

The mathematical exposition in Section 2 indicates that for a two sided paired samples t-test, a large observation either concordant or discordant with the rest of the sample will lead to a non-rejection of the null hypothesis. With the paired samples t-test the inclusion of a very large positive observation $x_n$ into a sample with $\bar{x}_{n-1} > 0$ may in fact severely reduce the probability of rejecting the null hypothesis.

Simulations comprising Normal deviates and in testing a nil-null hypothesis of no location effects have been performed. Stipulation of the condition $\bar{x}_{n-1} > 0$ does not invalidate the two-sided test procedure. However, the inclusion of a single, but often large discrepant observation, does imply that the nil-null hypothesis is not strictly true, hence our use of the terminology of the NHRR (the null hypothesis rejection rate), rather than using the terminology Type I error rate.

For small sample sizes there is a paradox when performing the paired samples t-test that more extreme values of the marching observation in the direction of the sample mean difference result in a greater p-value than a less extreme value of the marching observation.

Under a location shift model, the inclusion of genuinely large positive observation $x_n$ into a sample with $\bar{x}_{n-1} > 0$ should lead to an increase in statistical power in a two-sided test of the nil-null hypothesis. This effect is observed with Yuen's paired samples t-test and with the Wilcoxon signed rank sum test, but it is not consistently observed with the paired samples t-test.

Under a location shift model, the inclusion of a large negative observation $x_n$ into a sample with $\bar{x}_{n-1} > 0$ should lead to a relative decrease in statistical power. This effect is observed with Yuen's paired samples t-test and with the Wilcoxon signed rank sum test, but the effect is most evident, and is sample size dependent, for the paired samples t-test.

In summary, Yuen's paired samples t-test and the Wilcoxon signed rank sum test broadly display properties consistent with being robust statistical tests in the presence of a large outlier. In contrast the paired samples t-test displays behaviour strongly dependent on the magnitude of the outlier. Specifically, for small sample sizes the more extreme the values of the marching observation in the direction of the sample mean difference the greater the p-value compared to a less extreme value of the marching observation.

Zumbo and Jennings (2002), using their novel contamination model, concluded that the paired samples t-test had an inflated Type I error rate with increasing asymmetric contamination, however our marching observation simulations indicate that the effect of a single outlier on this test is dependent on sample size, magnitude and direction of the outlier, and could lead to increases and decreases in the NHRR. It should be noted that the simulations of Zumbo and Jennings (2002) consisted of situations in which the underlying distributions were contaminated with outliers and simultaneously a true null hypothesis is maintained. In contrast our simulations are based on the fulfilment of correct assumptions prior to the inclusion of the marching observation.

Our simulations demonstrate the seemingly paradoxical effect of large outliers on the performance of the paired samples t-test, and although we concur with Zimmerman (2011) that rank based methods do not necessarily eliminate the influence of outliers, the simulations indicate that Yuen's paired samples t-test and the Wilcoxon signed rank sum test have robust behaviour in the presence of a single outlying observation.

In the preparation of this paper, methods for outlier detection in the conditions above were attempted, but we were unable to identify a suitable method. With reference to paired samples, Preece (1982) states that formal procedures for the detection and rejection of outliers are of negligible use for small sample sizes. Further debate and investigation into outlier detection methods offers an area for further research.

## Acknowledgement

The authors thank the reviewers, and the editor, for their generous and insightful comments. Their valuable contributions have resulted in a significantly improved manuscript.

## References

[1] Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods*. **16(2)**, 1-32.

[2] Blair, R. C., & Higgins, J. J. (1985). A comparison of the power of the paired samples rank transform statistic to that of Wilcoxon's signed ranks statistic. *Journal of Educational Statistics*. **10(4),** 368-383.

[3] Box, G. E., & Muller, M. E. (1958). A note on the generation of random normal deviates. *The Annals of Mathematical Statistic*s. **29(2)**, 610-611.

[4] Chaffin, W. W., & Rhiel, G. S. (1993). The effect of skewness and kurtosis on the one-sample t test and the impact of knowledge of the population standard deviation. *Journal of Computation and Simulation*. **46**, 79-90.

[5] Fradette, K., Keselman, H. J., Lix, L., Algina, J., Wilcox, R. (2003). Conventional and Robust Paired and Independent Samples t-tests: Type I Error and Power Rates, *Journal of Modern Applied Statistical Methods*. **2(2)**, 481-496

[6] Gibbons, J. D., & Chakraborti, S. (2011). Nonparametric statistical inference. In *International encyclopedia of statistical science*, Chapter 5, page 196-221, Springer Berlin Heidelberg.

[7] Herrendörfer, G., Rasch, D., & Feige, K. D. (1983). Robustness of statistical methods II. Methods of the one-sample problem, *Biometrical Journal*, **25**, 327 – 343.

[8] Levine, T. R., Weber, R., Hullett, C., Park, H. S., & Lindsey, L. L. M. (2008). A critical assessment of null hypothesis significance testing in quantitative communication research. *Human Communication Research*, *34*(2), 171-187.

[9] Posten, H. O. (1979). The robustness of the one-sample t-test over the Pearson system. *Journal of Statistical Computation and Simulation*, **6**, 133 – 149.

[10] Preece, D. A. (1982). T is for trouble (and textbooks): a critique of some examples of the paired-samples t-test. *The Statistician,* **31(2)**, 169-195.

[11] R Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. www.R-project.org. 2014; version 3.1.3.

[12] Rasch, D., & Guiard, V. (2004). The robustness of parametric statistical methods, *Psychology Science*, **46**, 175 – 208.

[13] Ringwalt, C., Paschall, M. J., Gorman, D., Derzon, J., & Kinlaw, A. (2010). The Use of One- Versus Two-Tailed Tests to Evaluate Prevention Programs. **34(2)**, 135-150.

[14] Wilcox, R. R. (2005). Introduction to robust estimation and hypothesis testing (Statistical Modelling and Decision Science). Chapter 5.9.5, page 195-198, *Academic Press*.

[15] Yuen, K. K. (1974). The two-sample trimmed t for unequal population variances. *Biometrika*. **61**, 165-170.

[16] Zimmerman, D. W. (1997). A note on the interpretation of the paired samples, *Journal of Educational and Behavioral Statistics*. **22(3)**, 349 – 360.

[17] Zimmerman, D. W. (2011). Inheritance of Properties of Normal and Non-Normal Distributions after Transformation of Scores to Ranks. *Psicologica.* **32(1)**, 65-85.

[18] Zumbo B. D., & Jennings, M. J. (2002). The robustness of validity and efficiency of the related samples t-test in the presence of outliers. *Psicológica*, **23(2),** 415-450.