Universiteit Leiden

Psychologie
Faculteit der Sociale Wetenschappen

Violation of Homogeneity of Variances: A Comparison Between Welch T Test and the Permutation Test

Ruchella Kock

Bachelor thesis Psychology
Institute of Psychology
Faculty of Social and Behavioural Sciences – Leiden University
Date: 3 – 27 – 2019
Bachelorproject number: 34
Student number: 1815458
First examiner: Julian Karch

Contents

Introduction

In psychological research parametric tests are used more often than non-parametric tests (Edgington, 1974; Goodwin & Goodwin, 1985; Skidmore & Thompson, 2010). The t-test is one of these parametric tests that is widely used in research. It is used to statistically test the differences between means. There are many different types of t-test s such as the Student t test, Welch t-test and Yuen t-test (Student, 1908; Welch, 1947; Yuen, 1974). The tests differ in terms of the assumptions they make. The three central assumptions of the t-test are independence, homogeneity of variances and normality. The permutation test is a non-parametric tests also used to compare means. It was first discussed by Fisher (1937). In the permutation test all possible permutations are calculated for a sample, this forms a test distribution. The null hypothesis can be tested under this distribution (Howell, 2009). All statistical tests may lead to wrong conclusions, we distinguish between two types of errors. The type I error and type II error. Type I error is when H_0 is rejected when it should not have been. Type II error is when H_0 is not rejected when it should have been. We want both of these errors to be as low as possible. As the type II error decreases the power of a test increases. The power is the probability that H_0 is rejected when H_1 is true (Howell, 2009).

In this thesis we will compare the Welch t-test with the permutation test in terms of type I error and power. More specifically we will focus on the comparison between the two tests when the assumption of homogeneity is violated. Welch t-test was chosen because it provides more reliable type I error rates when the assumption of homogeneity of variance is not met. Compared to Student's t-test, Welch's t-test loses some statistical power. However, this loss is very small. (Delacre, Lakens, & Leys, 2017).

In the following two sections the assumptions of the two tests are discussed in depth. This is followed by a review of existing literature comparing the two tests. Then the research questions and hypothesis are discussed. Followed by a description of the simulation study and the results of the study. Finally we will end with a discussion of the findings and

conclude with which test performs better when the assumption of homogeneity is violated.

Assumptions of the t test

The t-test has a number of assumptions. First of all, it assumes independent errors. This means that the residuals should not be able to be predicted above chance. This assumption would be violated if within a group, one participant can influence another one. This assumption cannot be tested. It must be controlled for based on the design of the study. Secondly, it assumes that the the sampling distribution is normal. Many other parametric tests have this assumption. The reason why the normal distribution was chosen, can be explained by the central limit theorem (CLT). The CLT says that when starting with random and independent samples, the distribution of sample means will be approximately normally distributed if the sample size is large enough. As a simple example, say we flip a coin N number of times. Then we repeat the procedure many times. If N is large enough then our sampling distribution will be normal. The question remains as to what is large enough. In general it depends on what the sample looks like. However, it is believed that if $N \geq 30$ the sampling distribution of the mean is normal. Thus, if the sample size is greater than or equal to 30 then this assumption is fulfilled. The distribution of the data can be visually tested with a Q-Q/P-P plot. However, more often statistical tests are used namely, the Kolmogorov–Smirnov-test, the Shapiro-Wilk's W test or Skewness and Kurtosis can be checked. (Howell, 2009).

Another assumption is that there are no outliers. As the means of the groups are compared an outlier can greatly skew the mean which can lead to incorrect conclusions. There are many different ways to detect and deal with outliers. Some tests that can be used are Grubbs' test and Dixon's test. For a more detailed analysis on outliers and testing read Quesenberry and David (1961) or David and Paulson (1965).

Finally, there is the assumption of homogeneity of variances. Variance (σ^2) referes to the way the scores are distributed around the mean. Homogeneity of variances means that the variances across groups are considered equal. This assumption is important because if the

scores in the one group were spread differently, compared to the second group before any treatment was given, then these groups are no longer comparable (Salkind, 2010). The null hypothesis when testing this assumption is $H_0 : \sigma_1^2 = \sigma_2^2$ or $H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$. It is most commonly tested using the levene's test (Schultz, 1985). However, other tests include Hartley's F-max, Cochran's and Barlett's test (Conover, Johnson, & Johnson, 1981).
There are many studies that show that the t-test is robust against violations of the assumptions (e.g., Sawilowsky & Blair, 1992; Bradley, 1978). In general it is believed that the t-test is robust against non-normality if the samples size is greater or equal to 30. The t-test is believed to be robust against violation of the assumption of homogeneity if the group sizes are approximately equal. However, when the assumption of homogeneity is violated, the Welch t-test can be used. It is robust against this violation (Howell, 2009; Delacre et al., 2017).

Assumptions of the permutation test

There are two kinds of probability models namely the randomization model and the population model. In the randomization model, the subjects are randomly assigned to a condition. In the population model subjects are randomly sampled from a population (Ernst, 2004). The name permutation test is often used to refer to both the randomization model and population model because in many cases they can be equivalent to each other. The two tests are also referred to as randomization test and permutation test (Nichols & Holmes, 2002).
The randomization and permutation tests assume exchangeability. This assumption has different implications for the tests. One implication is the stable unit treatment value assumption (SUTVA). In Rubin (1980) he explained the idea of SUTVA. In an experiment, subjects/units i can be exposed to treatment j. Therefore, Y_{i_j} is the observed effect of unit i in treatment j. In this experiment each unit is only part of one treatment group at a time. Thus, Y_{i_1} and Y_{i_2} cannot be observed at the same time, we have to make inferences about the value that was not observed. The effect of treatment 1 on unit i should be independent

of the effect on other units in any treatment group, otherwise SUTVA will be violated. In the randomization model exchangeability should be a given because participants are randomly assigned to the groups and should therefore be thought of as interchangeable. However, this cannot always be assumed for the population model.

For the population model the random assignment was not possible therefore exchangeability cannot be directly assumed. Thus, there is the assumption that the distributions of the two groups have approximately the same shape (Nichols & Holmes, 2002).

These permutation tests are non-parametric tests. However, there are parametric permutation tests such as the permutation *t*-test s which then takes the assumptions of the *t*-test (Toothaker, Wisconsin Univ., & for Cognitive Learning., 1972; Mendeş & Akkartal, 2010).

To conclude, there is a subtle difference between the two tests in terms of who the population is. In this thesis the randomization model is used. Thus, the assumption of exchangeability is met. The randomization model is chosen because the population model is often not used in psychological studies. Convenient sampling is used instead (Fife, 2013). In the randomization model convenient sampling can be used as long as the participants are randomly distributed between groups. In the population model convenient sampling cannot be used because the researchers are not sampling from the population.

Literature review

In psychology the *t*-test is often used more than permutation tests (e.g., Goodwin & Goodwin, 1985). However, often times we do not know the extent to which the assumptions are met in the population (Hunter & May, 1993). Compared to the t test, the permutation test has less assumptions. It assumes exchangeability. However, the randomization test automatically meets this assumption. Thus, the question of how the two tests compare in terms of type I error and power remains. In this section we review some literature comparing the two tests.

Toothaker et al. (1972) wrote a dissertation on comparing the permutation t-test with Student's t-test and the Mann Whitney U test. He performed a simulation study using sample sizes ranging from 2 to 5. The study concluded that the permutation t-test does not outperform Student's t-test and Mann Whitney U test and the latter two should be preferred when comparing means.

Ludbrook and Dudley (1998) compared permutation tests with t-test and F-test in Biomedical Research. They found that researchers in this field often choose an F-test or t-test instead of a permutation test even if the assumptions are not met. They conclude that exact permutation or randomization tests should be preferred in biomedical research.

Hughes (2010) conducted a simulation study, she compared the two sample t-test with the two sample exact permutation test. She used 6 non-normal distributions, tested at 3 different significance levels and the sample sizes ranged from 2 to 6. She concluded that the permutation test should be preferred, especially if power is very important for a study.

Most relevant to this thesis is the simulation study performed by Mendeş and Akkartal (2010). They compared the ANOVA F-test and Welch t-test with the permutation F-test and the permutation Welch t test. They used 3 different distributions, 5 different group sizes ranging from 5 to 15 and 3 different group variances namely, equal variances ($\sigma_1^2 = 1, \sigma_2^2 = 1, \sigma_3^2 = 1$), a small deviation ($\sigma_1^2 = 1, \sigma_2^2 = 1, \sigma_3^2 = 4$) and a larger deviation ($\sigma_1^2 = 1, \sigma_2^2 = 1, \sigma_3^2 = 9$). By comparing these groups they observed the effects of non-normality and heterogeneity. They concluded that when the assumption of homogeneity and normality is violated, the permutation F-test should be used. When the assumption of normality is violated but equal variances are assumed then the permutation Welch t-test should be used.

In all these papers very small sample sizes were used, the largest group size being 15. This is not representative of current psychological studies. We can see in the study from Kühberger, Fritz, and Scherndl (2014) that only 14.9% of studies had a sample size of 15 or smaller. Moreover, only one study compared the two tests when the homogeneity

assumption is violated. Mendeş and Akkartal (2010) also looked at the effect of different group sizes. However, the largest deviation between groups was 10. Larger deviations when comparing the two tests have not been studied.

This thesis aims to observe the effects on both tests when the assumption of homogeneity is violated. This assumption is not widely explored. Most literature observe the effect of non normality (e.g., Hughes, 2010; Weber & Sawilowsky, 2009). Moreover, the t-test is robust against violation of homogeneity of variances when the group sizes are approximately equal (Howell, 2009). Therefore, there should be more research on the effect of different group sizes. We look at small deviations as well as large deviations between group sizes. The goal of this thesis is to provide relevant results that can be used in psychological research. To achieve this goal sample sizes that are often used in psychology will be chosen and the randomization test which is more common in psychological research will be used.

Research questions

To be able to compare the Welch t-test with the permutation test when the variances are not equal, we look at different circumstances that affect the tests. The tests are compared in terms of type I error and power. In this section, the questions and sub-questions and hypothesis are discussed.

The research question of this thesis is: How does the permutation test compare to the Welch t-test under violation of the assumption of homogeneity of variances? To answer this question the following sub-questions will be explored.

- How does the permutation test compare to Welch t-test under no violation of the assumption of homogeneity of variances?
- What is the effect of sample size on the performance of the permutation test under violation of the assumption of homogeneity of variances?
- What is the effect of sample size on the performance of the Welch t-test under violation of the assumption of homogeneity of variances?

- What is the effect of unequal group sizes on the performance of the permutation test under violation of the assumption of homogeneity of variances?
- What is the effect of unequal group sizes on the performance of the Welch t-test under violation of the assumption of homogeneity of variances?

We hypothesize that due to the robustness of the Welch t-test against violation of homogeneity (Delacre et al., 2017), the Welch t-test outperforms the permutation test.

Methods

To compare the t-test and permutation test, a simulation study is conducted using the programming language R (R Core Team, 2018). The type I error and power of the Welch t-test and permutation test are computed and compared against each other.

Sample sizes

To select sample sizes that are relevant to psychology the data provided by Kühberger et al. (2014) was used. They investigated whether effect size is independent from sample size in psychological research. To do this they randomly sampled 1000 articles from 22 different psychological disciplines. All these articles were published in 2007. They excluded many articles because they did not meet their criteria. The final data set contained 531 articles. From these articles 529 sample sizes were reported.

From these articles we selected 3 different sample sizes. First of all, a small sample size used in psychology, namely $N = 10$. Despite this being a small sample size, it was reported in 13 articles. Less than 10% of the articles have a sample size that is smaller than 10 (8.9%). Secondly, a medium sample size that is commonly used was chosen. This is $N = 60$. It was reported 11 times and almost half of the sample sizes are smaller than or equal to 60 (47.6%). The third sample size is an extremely large sample size, namely $N = 1000$. Only 10% of the sample sizes that were reported were larger than 1000.

To determine the effect of group ratios 8 different ratios were chosen. The first group sizes are equal this is called condition 1. The other groups have a downwards or upwards

deviation. The second group sizes have a small deviation of 25% (condition 2a and 2b). The third group sizes (condition 3a and 3b) deviate with 50% and condition 4a and 4b have a large deviation of 75%. This creates 24 different conditions each with slightly different group sizes (see Table 1).

Simulation

Data was simulated with a normal distribution. The mean was 0 and the standard deviation was 1. When testing for type I error both means were kept equal. If the p-value of either test is smaller than $\alpha = 0.05$, then there is a type I error. The power is calculated with 1 - type II error. Type II error was simulated with 0 as the mean of one group and 1 as the mean of the other group. This is a mean that deviates by one standard deviation. If the p-value of either test is larger than $\alpha = 0.05$, then there is a type II error.

To simulate the violation of the assumption of homogeneity of variances, the standard deviation (σ) was altered. This is because variance is the squared standard deviation (σ^2). When there is no violation the variances of both groups are equal ($\sigma_1^2 = 1 : \sigma_2^2 = 1$). However, when the assumption is violated, the two variances are not equal. Six different deviations were chosen to simulate this. A downwards and upwards deviation of 25%, 50%, 75% and an upwards deviation of 300% (see Table 2). Each condition was performed with all seven deviations. The simulation was essentially performed 7 * 8 times for all 3 sample sizes.

We also varied the effect size (ES). The ES is the standardized mean difference between two groups (Coe, 2002). If there is a strong effect, the ES will be large. This means that the probability that the statistical test is significant is also large. Therefore, different effect sizes have different implications. In this thesis Cohen's three benchmark effect sizes were chosen, namely a small ES of 0.2, a medium ES of 0.5 and a large ES of 0.8 (Cohen, 2013). Each simulation was repeated 10000 times. The code is included in the appendix. The number of type I and type II errors that occurred in 10000 were recorded for each condition. The data was simulated using `rnorm()`. The Welch t-test was performed using

the `t.test()` formula in R with the argument `var.equal` set to False. The permutation test was performed using the library *perm* (Fay & Shaw, 2010). The Monte Carlo sampling technique was used during the permutation test. Ideally all permutations are performed in a permutation test. However, with larger sample sizes the number of permutations become very large. Therefore, the Monte Carlo sampling technique should be used. This technique randomly chooses test statistics from the permutation distribution. From this random sample the *p*-value for the permutation test can be calculated (Ernst, 2004; Hastings, 1970). We can conclude that one test outperforms another when the type I and type II error for one test is smaller than or equal to 0.05 and the type I and type II error for the other test is larger than 0.05. If both tests have a type I and type II error smaller or equal to 0.05, then we look at how big the difference is between the two tests. The test with the smaller errors outperforms the other.

References

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*(2), 144-152. doi: 10.1111/j.2044-8317.1978.tb00581.x

Coe, R. (2002). It's the effect size, stupid: What effect size is and why it is important.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.

Conover, W. J., Johnson, M. E., & Johnson, M. M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics*, *23*(4), 351-361. doi: 10.1080/00401706.1981.10487680

David, H., & Paulson, A. (1965). The performance of several tests for outliers. *Biometrika*, *52(3/4)*(3/4), 429–436. doi: 10.2307/2333695

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use welch's t-test instead of student's t-test. *International Review of Social Psychology*, *30*(1), 92–101. doi: http://doi.org/10.5334/irsp.82

Edgington, E. S. (1974). A new tabulation of statistical procedures used in apa journals. *American Psychologist*, *29*(1), 25-26. doi: http://dx.doi.org/10.1037/h0035846

Ernst, M. D. (2004, 11). Permutation methods: A basis for exact inference. *Statist. Sci.*, *19*(4), 676–685. doi: 10.1214/088342304000000396

Fay, M. P., & Shaw, P. A. (2010). Exact and asymptotic weighted logrank tests for interval censored data: The interval R package. *Journal of Statistical Software*, *36*(2), 1–34. Retrieved from `http://www.jstatsoft.org/v36/i02/`

Fife, D. A. (2013). *The achilles heel of psychology: How convenience sampling affects parameter estimates*. The University of Oklahoma.

Fisher, R. A. (1937). *The design of experiments*. Edinburgh, London: Oliver And Boyd.

Goodwin, L. D., & Goodwin, W. L. (1985). An analysis of statistical techniques used in the journal of educational psychology, 1979-1983. *Educational Psychologist*, *20*(1), 13-21. doi: 10.1207/s15326985ep2001_3

Hastings, W. K. (1970, 04). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97-109. doi: 10.1093/biomet/57.1.97

Howell, D. C. (2009). *Statistical methods for psychology*. Cengage Learning.

Hughes, M. B. (2010). *Comparison of size/level and power of two-sample t-test to that of two-sample exact permutation test under various population shapes and small sample sizes*. Stephen F. Austin State University.

Hunter, M. A., & May, R. B. (1993). Some myths concerning parametric and nonparametric tests. *Canadian Psychology/Psychologie canadienne*, *34*(4), 384–389. doi: http://dx.doi.org/10.1037/h0078860

Kühberger, A., Fritz, A., & Scherndl, T. (2014, 09). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *PLOS ONE*, *9*(9), 1-8. doi: 10.1371/journal.pone.0105825

Ludbrook, J., & Dudley, H. (1998). Why permutation tests are superior to t and f tests in biomedical research. *The American Statistician*, *52*(2), 127–132. doi: 10.1080/00031305.1998.10480551

Mendeş, M., & Akkartal, E. (2010). Comparison of anova f and welch tests with their respective permutation versions in terms of type i error rates and test power. *Kafkas Univ Vet Fak Derg*, *16*(5), 711–716. doi: 10.9775/kvfd.2009.1507

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping*, *15*(1), 1–25. doi: https://doi.org/10.1002/hbm.1058

Quesenberry, C. P., & David, H. A. (1961). Some tests for outliers. *Biometrika*, *48(3/4)*(3/4), 379–390. doi: 10.2307/2332759

R Core Team. (2018). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, *75*(371), 591–593. doi: 10.2307/2287653

Salkind, N. J. (2010). *Encyclopedia of research design* (Vol. 1). Sage.

Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and type ii error properties of the t test to departures from population normality. *Psychological bulletin*, *111*(2), 352-360. doi: http://dx.doi.org/10.1037/0033-2909.111.2.352

Schultz, B. B. (1985, 12). Levene's Test for Relative Variation. *Systematic Biology*, *34*(4), 449-456. doi: 10.1093/sysbio/34.4.449

Skidmore, S. T., & Thompson, B. (2010). Statistical techniques used in published articles: A historical review of reviews. *Educational and Psychological Measurement*, *70*(5), 777-795. doi: 10.1177/0013164410379320

Student. (1908). The probable error of a mean. *Biometrika*, *6*(1), 1–25. doi: doi:10.2307/2331554

Toothaker, L. E., Wisconsin Univ., M. R., & for Cognitive Learning., D. C. (1972). *An empirical investigation of the permutation t-test as compared to student's t-test and the mann-whitney u-test.* Washington, D.C.: ERIC Clearinghouse.

Weber, M., & Sawilowsky, S. (2009). Comparative power of the independent t, permutation t, and wilcoxontests. *Journal of Modern Applied Statistical Methods*, *8*(1), 3. doi: 10.22237/jmasm/1241136120

Welch, B. L. (1947). The generalization of 'student's' problem when several different population variances are involved. *Biometrika*, *34(1/2)*(1/2), 28–35. doi: doi:10.2307/2332510

Yuen, K. (1974, 04). The two-sample trimmed t for unequal population variances. *Biometrika*, *61*(1), 165-170. doi: 10.1093/biomet/61.1.165

Table 1

Methods

Sample Size	Group Ratios
Small N = 10	
Condition 1	$N_1 = 10 : N_2 = 10$
Condition 2a	$N_1 = 10 : N_2 = 8$
Condition 2b	$N_1 = 10 : N_2 = 13$
Condition 3a	$N_1 = 10 : N_2 = 5$
Condition 3b	$N_1 = 10 : N_2 = 15$
Condition 4a	$N_1 = 10 : N_2 = 3$
Condition 4b	$N_1 = 10 : N_2 = 18$
Medium N = 60	
Condition 1	$N_1 = 60 : N_2 = 60$
Condition 2a	$N_1 = 60 : N_2 = 45$
Condition 2b	$N_1 = 60 : N_2 = 75$
Condition 3a	$N_1 = 60 : N_2 = 30$
Condition 3b	$N_1 = 60 : N_2 = 90$
Condition 4a	$N_1 = 60 : N_2 = 15$
Condition 4b	$N_1 = 60 : N_2 = 105$
Large N = 1000	
Condition 1	$N_1 = 1000 : N_2 = 1000$
Condition 2a	$N_1 = 1000 : N_2 = 750$
Condition 2b	$N_1 = 1000 : N_2 = 1250$
Condition 3a	$N_1 = 1000 : N_2 = 500$
Condition 3b	$N_1 = 1000 : N_2 = 1500$
Condition 4a	$N_1 = 1000 : N_2 = 250$
Condition 4b	$N_1 = 1000 : N_2 = 1750$

Table 2

Standard Deviations

Standard Deviation
$\sigma_1 = 1 : \sigma_2 = 1$
$\sigma_1 = 1 : \sigma_2 = 0.75$
$\sigma_1 = 1 : \sigma_2 = 1.25$
$\sigma_1 = 1 : \sigma_2 = 0.50$
$\sigma_1 = 1 : \sigma_2 = 1.50$
$\sigma_1 = 1 : \sigma_2 = 0.25$
$\sigma_1 = 1 : \sigma_2 = 1.75$
$\sigma_1 = 1 : \sigma_2 = 3$