ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Master Thesis Data Science and Marketing Analytics

Robustness of Permutation Tests Compared to Classical Methods: A Study on Permutation test, Student's t-test, maxT, and Bonferroni

Name Student: Pepijn Vonk

Student ID Number: 530164

Supervisor: Jesse Hemerik

Second assessor: Bas Donkers

Date Final Version: 22-08-2024

Abstract

Outliers are a common sight in statistical research or data analysis and they can negatively affect the validity of the results of these studies or analyses. Therefore, using methods or statistical tests that are robust to outliers is important to guarantee validity and reliability. This thesis aims to provide a direct comparison between multiple statistical tests and methods, so researchers can choose the most robust method for their study if their data (potentially) contains outliers. This thesis researches the robustness of classical statistical methods, the Student's t-test and the Bonferroni correction, and permutation methods, the permutation test and maxT method, in the presence of outliers. Using simulations on height and weight data and the Golub dataset, this research evaluates the influence of outliers by comparing results before and after the introduction of outliers. The empirical analysis demonstrates that the permutation methods showed better robustness to outliers due to their flexibility and adaptation of the analyzed data. These findings offer a practical understanding of these methods for researchers in selecting the appropriate statistical method, contributing to more reliable data analysis and statistical studies in various fields.

Inhoud

1. Introduction

In statistical analysis, outliers can have a significant effect on the results and interpretation of hypothesis tests. The error rates can inflate and it could lead to substantial distortions of parameters (Osborne & Overbay, 2004). Hawkins (1980) has defined outliers as follows: "*An Observation which deviates so much from other observation as to arouse suspicions that it was generated by a different mechanism.*" These outliers could pose a challenge when working with statistical measures (Osborne & Overbay, 2004).

Parametric tests, like the Student's t-test, assume that the observations and population follow a certain distribution deviations from this distribution are considered to be outliers. However, when data or samples from the total population do not follow the assumption of a normal distribution or contain outliers, parametric tests could result in biased or inaccurate outcomes.

In contrast, permutation tests, a non-parametric test introduced back in 1925 by Fisher (Berry et al., 2014), do not assume that the observations or population follow any kind of distribution. The different variables of the observations are permutated randomly. This could therefore mean that permutation tests are potentially more robust to outliers.

Marketing and business studies commonly use statistical methods like the Student's t-test and permutation tests. Baidun et al. (2022) uses, among others, the t-test to measure the impact of the marketing mix on customer satisfaction. Eusebio et al. (2006) use the t-test to compare two groups of Spanish firms and their marketing performance. Burk (2006) applied the t-test to A/B split testing, even stating that the results of A/B testing are most often compared using t-tests. Haenlein and Kaplan (2011) described permutation tests and t-tests and analysed their statistical power for marketing research. Tempesta et al. (2010) use permutation tests for relationship identification for market segments. These papers are examples of comparison tests in marketing studies. This indicates that group comparison tests are commonly used for marketing research purposes.

Situations where multiple tests are conducted are also becoming more common. Multiple test problems are not very common in economics, however, multiple tests can be very useful in economic studies. Examples of this are studies conducted by Harvey et al. (2020), List et al. (2019), Romano and Wolf (2005), and Viviano et al. (2021). This indicates that multiple test problems are used in economic research.

Thus, this paper could provide marketing researchers or marketers with useful insights about the most robust tests to compare groups. Marketers could use the insights of this paper to select the most robust or powerful method for their study or analysis. This indicates that the findings of this paper are also

relevant to the marketing field and could help marketers of businesses improve the validity of their analyses.

A research field where multiple test problems are commonly used is biomedical sciences. These problems could include extracted gene expressions and measured phenotype associations (Menyhart et al., 2021). Datasets on these topics usually contain a great number of variables. Finding conclusive evidence could be of essential importance in these studies. If these multiple tests are performed on datasets which contain statistical outliers this could force errors in the results. Common methods for multiple tests are maxT and Bonferroni (Westphal & Zapf, 2024). Bonferroni divides the wanted significance level by the number of tests adjusting the significance level. The maxT method is a form of the permutation test. The distribution of the maximum test statistic is generated under the null hypothesis through data permutations. It is stated that Bonferroni is more conservative than the maxT method (John et al., 2022; Nakagawa, 2004).

Classical statistical methods, such as the Student's t-test, are based on assumptions about the distributions of the data. The Bonferroni correction, a classical statistical method for multiple tests, does not directly rely on an assumption (Cheverud, 2001). These simple and well-understood methods make them useful in many scenarios where the assumptions hold or the data is simple. However, the reliance on parametric assumptions can increase the influence of outliers on the test statistic. Bonferroni is considered very conservative (Noguchi et al., 2019). This could negatively influence the correctness of the tests' outcome when outliers are included in the data. On the contrary, permutation methods, like the permutation test for single comparisons or the maxT method for multiple tests, could offer a more flexible alternative since these only rely on the exchangeability assumption. By reshuffling the data numerous times, permutation tests can result in test statistics and conclusions which are more reliable when outliers are included in the data. The comparison between classical and permutation methods thus highlights a trade-off between the simplicity of classic methods and the flexibility of permutation methods.

1.1. Relevance

Despite the potential advantages of permutation tests in handling outliers, there have been limited studies conducted in which the robustness of permutation tests and Student's t-tests against outliers are compared. While individual studies have explored the robustness of each test separately, a direct comparison between the two methods regarding their ability to withstand the influence of outliers has not been explored extensively.

Furthermore, the robustness of statistical tests in the context of multiple test scenarios also needs attention. The maxT and Bonferroni methods can both be used to address the issue of multiple

comparisons (Hemerik & Goeman, 2017b; Goeman & Solari, 2014). However, when it comes to handling outliers, both methods might react differently. The Bonferroni correction's conservativeness is caused by Bonferroni adjusting the significance level for the number of tests, which can lead to an overly strict threshold (Noguchi et al., 2019). This could, in theory, increase the chance of Type II errors occurring. On the other hand, the maxT method, a permutation-based approach, generates the distribution of the maximum test statistic under the null hypothesis through data permutations. This results in maxT being more flexible and less conservative than Bonferroni, albeit in potential. Some studies have researched the robustness of the Bonferroni method, as Ringland (1983) did in his paper. However, few papers have looked into the influence of outliers on the outcomes of Bonferroni and maxT. A comparison between the two could provide useful insight for future statistical implications.

Since outliers can significantly manipulate the results of statistical tests, it is important to find proof of whether permutation tests or classical methods are more robust to these outliers. This comparison could be essential for helping researchers select the appropriate statistical test technique for analyzing data and conducting statistical studies, particularly when outliers are prevalent in data. This research aims to enhance the general understanding of the robustness of statistical tests against outliers by filling the existing void in the academic literature and, thus, helping improve the reliability and/or validity of statistical analyses for every kind of research that relies on statistical tests.

1.2. Central Research Question

This thesis aims to investigate and compare the robustness of permutation tests, including the maxT method for multiple tests and classical statistical tests, in this case, Student's t-test and the Bonferroni method, when outliers are included in the tested data. The research question of this thesis is: "Are permutation tests and the maxT method more robust to extreme statistical outliers than classical tests, especially the Student's t-test and the Bonferroni correction?"

1.3. Sub Questions

This research question will be further divided into several sub-questions:

1. How do permutation tests and the maxT method work?
2. How do Student's t-tests and the Bonferroni correction work?
3. How robust are permutation tests to extreme statistical outliers and how does this compare to Student's t-tests?
4. How robust is the maxT method to extreme outliers and how does this compare to the Bonferroni method?
5. What happens to the outcome of the different methods when the outliers become more extreme?

6. What happens to the outcome of the different methods when the frequency of outliers increases?

2. Theory

Student's t-tests, like other parametric tests, are based on several critical assumptions about data distribution. These are among others: normality, homogeneity of variances, and independence of observations (Keren & Lewis, 1993). The test statistic is used to calculate the p-value. Therefore, standard normal and t-distribution are an important part of these assumptions (Field, 2018; Tabachnick & Fidell, 2019). These assumptions facilitate valid and reliable test results when the analyzed data comply with the assumptions.

On the other hand, permutation tests are based on almost no assumptions about the distribution of the data (Berry et al., 2014). Instead, permutation tests rely on a distribution of the test statistic generated from the observed data (Good, 2013; Edgington & Onghena, 2007; Phipson & Smyth, 2010). Next, the proportion of permutations that result in a test statistic as extreme as or more extreme than the observed statistic is used to compute the p-value (Ernst, 2004). This characteristic allows permutation tests to be versatile and could thus be applicable in a wide range of scenarios where the assumptions of parametric tests may not hold (Collingridge, 2012).

2.1. Robustness

Permutation tests offer several benefits. These could include robustness to distributional assumptions, flexibility in handling complex data, and better control over error rates. For instance, when the data do not meet the assumptions, permutation tests could provide a reliable alternative to achieve more accurate conclusions (Pesarin & Salmaso, 2010). Due to these properties, permutation tests could be better suited for hypothesis tests in situations where parametric assumptions are not met or are difficult to verify (Ludbrook & Dudley, 1998).

The t-statistic in Student's t-tests represents the difference between sample means that is standardized after being adjusted for the variability within and between groups (Berry et al., 2014). The p-value is calculated based on the t-distribution (Walpole et al., 2006). Since this test relies on assumptions, this could form a problem when these assumptions are violated due to the presence of outliers or non-normal data (Field, 2017). These violations could result in the validity of the test results being compromised, leading to inaccurate conclusions. Means can relatively easily be influenced by outliers or extreme values (Moore et al., 2016). This could form problems for both Student's t-tests and permutation tests that measure the differences in means.

The theoretical arguments stated above suggest that permutation tests may be more robust to extreme statistical outliers compared to Student's t-tests due to their non-parametric nature and reliance on empirical distributions. Permutation tests could be less susceptible to the influence of outliers that violate parametric assumptions (Ludbrook & Dudley, 1998).

2.2. Scenarios

On the other hand, permutation tests are generally considered more robust to outliers due to their non-parametric nature and lack of distributional assumptions (Keller-Mcnulty & Higgins, 1987). There could be possible scenarios in which Student's t-tests may exhibit greater robustness to outliers.

Student's t-tests might be more robust to outliers in larger sample sizes. With larger sample sizes, the distribution of sample means could follow a normal distribution better, even in the presence of outliers (Kwak & Kim, 2017). This is considered to be true under the Central Limit Theorem. Consequently, the t-statistic used in Student's t-tests could become more accurate, leading to better performance.

Another scenario could be if the underlying distribution of the data nearly follows a normal distribution. In this situation, Student's t-tests may be more robust to outliers, provided that the other assumptions of homogeneity of variances are met (Sawilowsky & Blair, 1992) If these assumptions hold, the t-tests could offer better-performing outcomes compared to permutation tests (Hochberg & Tamhane, 1987).

When the variances of the compared groups are approximately equal, Student's t-tests could show robustness to outliers, as the impact of any single outlier is lowered by the homogeneity of variance (Field, 2017). In contrast, permutation tests are sensitive to differences in variances between groups (Ludbrook & Dudley, 1998). Therefore, in cases of equal variances and sample sizes, Student's t-tests may perform better due to the homoscedasticity assumption (Zimmerman, 2004).

Overall, while permutation tests have the advantages of flexibility and minimal assumptions, Student's t-tests could be more efficient under specific conditions of near-normal distribution and equal variances (Edgington & Onghena, 2007). These different scenarios will all be tested in the thesis.

2.3. Multiple Test

The maxT method is a statistical procedure used to control the familywise error rate (FWER) when testing multiple hypotheses. The maxT method focuses on the maximum absolute value of test statistics for multiple comparisons. This statistic is used to determine the significance of the tests (Dudoit, Shaffer, & Boldrick, 2003). By concentrating on the maximum statistic, the maxT method tries to control the overall error rate and provide a correction for multiple tests.

The maxT method is potentially robust to outliers. The maxT method shrinks the impact of outliers on Type I error control through its focus on the maximum absolute value of test statistics. This offers flexibility to handle skewed or tailed distributions that may be created by the presence of outliers (Westfall & Young, 1993).

The Bonferroni method is a straightforward and commonly used approach for controlling the FWER in scenarios where multiple hypotheses are simultaneously tested. Bonferroni specifically divides, as stated before, the significance level (e.g. 0.05) by the number of individual tests. This, albeit, conservatively, controls the Type I Error Rate. However, the influence of outliers on individual tests is not directly tackled (Holm, 1979). Bonferroni's conservativeness is amplified when dealing with a large number of tests (Armstrong, 2014). The maxT method, by contrast, provides a more balanced approach to error rate control (Romano & Wolf, 2005).

The Bonferroni method remains widely used due to its simplicity and ease of implementation. This results in lower computational intensity, especially compared to the maxT method (Armstrong, 2014). The maxT method, permutating data numerous times, requires more computational resources and most often takes more time to be calculated.

This results in a trade-off between simplicity and robustness. The Bonferroni method is easier to apply, being predominantly conservative. The maxT method could offer an approach that is more flexible and potentially less conservative, possibly even in datasets with outliers (Westfall & Young, 1993; Romano & Wolf, 2005).

3. Data

3.1. Dataset Overview

This study aims to empirically research whether permutation tests are more robust to outliers than Student's t-tests. To research this a dataset with weights and heights of 25.000 different humans with the age of 18 years old (SOCR Data Dinov 020108 HeightsWeights - Socr, n.d.). This dataset will further be referred to as the Hong Kong dataset. Weight and height variables are naturally normally distributed (López-Siguero et al., 2008). Therefore they make a useful dataset to use in this research as normal distribution is assumed by the Student's t-test method. The data was used to develop the growth charts that are currently used in Hong Kong. The dataset only contains 3 variables. The first variable indicates the index of the variable. The second variable states the height of the corresponding individual. The height is given in inches, the length measure of the Imperial system. Accordingly, the weight of the observations is stated in pounds. This weight is the third and final variable from the dataset.

To investigate the multiple test methods, the Golub dataset will be used. The Golub dataset is widely used in the field of bioinformatics and computational biology. It was originally used in a study by Golub et al. (1999). This study studied the potential to identify the kind of acute leukaemia patients suffered from. This was done by analysing the gene expressions of the patients. The dataset includes 72 samples or observations. The samples were taken from patients having two different types of leukaemia: 47 samples of acute lymphoblastic leukaemia (ALL) and 25 samples of acute myeloid leukaemia (AML). Each sample contained the expression levels of 7,129 different genes.

3.2. Data Processing and Cleaning of the Hong Kong data

Data cleaning is a critical step to ensure the accuracy and reliability of the analysis. However, the dataset was pre-processed and therefore did not require any cleaning. No missing values were included in the data.

Outliers are the focus of this study. However, if there were outliers in the original data, this would not be useful. The main analysis of this study requires a normal distribution and the possibility of controlling the outliers. Boxplots were drawn up to indicate if any outliers were present in the data. The boxplots in Figures A and B (Appendix) indicate that the original dataset included several outliers for both the Height and Weight variables. These outliers, however, were not deleted from the dataset, since no outliers seem to be an extreme deviation from the rest of the data. Normal distributions naturally contain outliers, which are not extreme (Wilks, 1963). Therefore, the outliers do not form a problem. Additionally, after deleting the outliers both variables did not follow a normal distribution. Thus, it was concluded that the initial outliers would not be deleted and remained in the data.

3.3. Data Exploration of the Hong Kong data

Furthermore, for initial exploration of the data summary statistics were calculated for the Height and Weight variables. These statistics provide a first insight into the tendencies and variability within the data. The summary statistics on the original data are shown in Table 1.

Table 1

Overview of summary statistics of Hong Kong data

Variable	*Min*	*Median*	*Mean*	*Max*	*Standard Deviation*	*Variance*
Height (Inches)	60.28	68.00	67.99	75.15	1.90	3.61
Weight (Pounds)	78.01	127.16	127.08	170.92	11.66	135.98

Since Student's t-tests assume a normal distribution (Berry et al., 2014), checking whether the variables follow a normal distribution is useful. Naturally, height tends to follow a normal distribution for a specific age group (López-Siguero et al., 2008). A normal distribution does not naturally occur for weights, however, this does not indicate anything for the weight variable of this study.

The first method to check for a normal distribution uses a histogram (Das, 2016). It could provide an initial insight into a possible skewness or distribution of the data. Figures 1 and 2 show the histograms with the distributions of the Weight and Height variables respectively. A line indicating a normal distribution's 'bell shape' was included. This was done to help the interpretation.

Figure 1

Histogram of Height variable of Hong Kong data

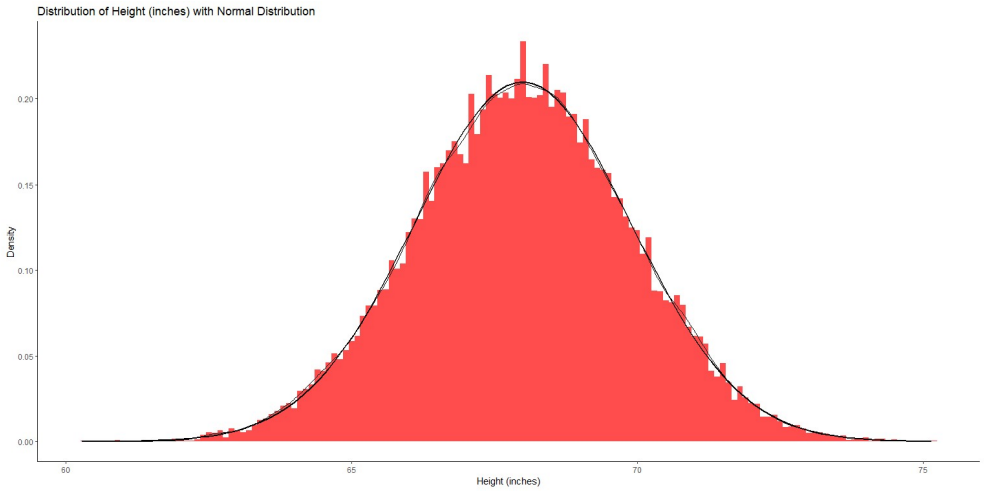
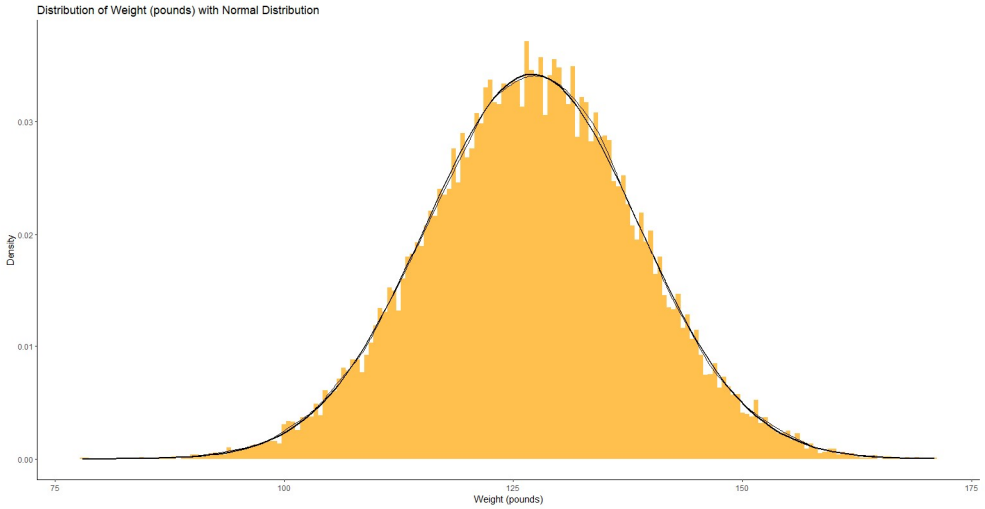


Figure 2

Histogram of Weight variable of Hong Kong data



Both histograms indicate that the Weight and Height variables follow a normal distribution. There is no clear skewness and the bars in the histogram seem to follow the outline of a normal distribution. The second test encompassed a quantile-quantile plot, also known as a Q-Q plot (Das, 2016). A Q-Q plot shows a comparison of two distributions by evaluating quantiles (Almeida et al., 2018; Lee, 2020). The Q-Q plot combines the distribution of the data and compares it to a normal distribution. The resulting plot could be used to indicate whether data follows a normal distribution. If most points follow the central line, this indicates that the observations for the variables could be considered normally distributed (Michael, 1983).

Figures C and D (Appendix) show a similar pattern. Between the 2.5 quantiles, the observations from the Hong Kong data almost perfectly follow the 45-degree line of the Q-Q plot. This indicates that the centre fit of the data follows a normal distribution almost perfectly (Almeida et al., 2018; Lee, 2020). The tails of the distribution are approximately in the accepted range. This means that according to the Q-Q plots almost no observations would be considered outliers (Wilk & Gnanadesikan, 1968).

Since the Q-Q plot and histograms did not provide a conclusive answer, other statistical tests were conducted. These tests included the Shapiro-Wilk, Kolmogorov-Smirnov, and Anderson-Darling test. These tests are empirical distribution tests or check the regression to decide whether the data is normally distributed (Yap & Sim, 2011b).

The Shapiro-Wilk, Kolmogorov-Smirnov, and Anderson-Darling tests are very useful for situations in which the normality of datasets needs to be tested (Stephens, 1974). By calculating test statistic *W* against critical values, the Shapiro-Wilk test checks whether data follows a normal distribution by comparing (Shapiro & Wilk, 1965). However, this test can only be used on datasets on smaller datasets. Therefore, a random sample of 5,000 observations was pulled and used to test whether the overall dataset follows a normal distribution. The test statistic *W* is calculated by the following formula:

$$W = \frac{(\sum_{i=1}^{n} a_i\, x_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Where a_i is the coefficient calculated as: $(a_1, a_2, \ldots, a_n) = \frac{m^T V^{-1}}{C}$

With *m* is the vector and vector norm *C is*: $C = (m^T V^{-1} V^{-1} m)^2$

Kolmogorov-Smirnov and Anderson-Darling for larger datasets are more often used for bigger datasets and therefore were used on the complete dataset. The Kolmogorov-Smirnov test measures the goodness-of-fit between the functions of the empirical distribution and the cumulative distribution of a normal distribution (Stephens, 1974). This test uses the maximum absolute difference, called *D* and is calculated by the following formula: $D_n = \sup(x)\,|F_x(x) - F(x)|$

Where *F* is the distribution function with *n* number of ordered observations X_i: $F_n(x) = \frac{\sum_{i=1}^{n} 1(-\infty)(X_i)}{n}$

Similarly, the Anderson-Darling test computes A^2. A^2 is the test statistic and it is compared against critical values. This test is mostly known for its sensitivity in detecting deviations in distribution tails. A^2 is calculated as follows: $A^2 = -n - S$

Where *S* is calculated as follows: $S = \sum_{i=1}^{n} \frac{2i-1}{n}\left[\ln\left(F(Y_i)\right) + \ln\left(1 - F(Y_{n+1-i})\right)\right]$

F is the cumulative distribution function.

These tests are very important in determining whether data follows normality. Stephens (1974) discusses these methods comprehensively, emphasizing their utility in statistical analyses. Since there is no significant difference or preference for any of these tests, all three tests were conducted on the data to determine whether the data is normally distributed. The test statistics as well as the corresponding p-values for the tests for the Height and Weight variables are shown in Tables 2 and 3. Important to note for these p-values is that in the case of the p-value being smaller than 0.05, the null hypothesis is rejected. The null hypothesis in this case is that (the sample of) the data is normally distributed. So, if the p-value is smaller than 0.05 the data is assumed to not be normally distributed.

Table 2

Overview of test statistics for Height with original data

Test	*Test Statistic*	*P-value*	*Normality assumed*
Shapiro-Wilk	1.000	0.326	Yes
Kolmogrov-Smirnov	0,003	0.979	Yes
Anderson-Darling	0.242	0.771	Yes

Note: The test statistics and p-values are rounded to three decimals for better readability and interpretability.

Table 3

Overview of test statistics for Weight with original data

Test	*Test Statistic*	*P-value*	*Normality assumed*
Shapiro-Wilk	1.000	0.288	Yes
Kolmogrov-Smirnov	0.004	0.808	Yes
Anderson-Darling	0.525	0.181	Yes

Note: The test statistics and p-values are rounded to three decimals for better readability and interpretability.

The findings from the different normality tests all show a similar picture for both the Height and the Weight variables. The Shapiro-Wilks, Kolmogorov-Smirnov, and Anderson-Darling tests all indicate that both the Height and Weight variables follow a normal distribution. This could be important as one of the scenarios that will later be tested is the robustness of Student's t-test and permutation tests when the tested data follow a normal distribution. Since the Student's t-test assumes normality, this could also help improve the validity of the tests when normality is not specifically checked.

3.4. Data Processing and Cleaning of the Golub data

The Golub data contained 72 observations or samples but did contain 7,129 gene probes (T. Golub, 2024). Furthermore, six other variables were included. These variables were: the sample number, the source of the sample, the gender of the patient, the hospital of the sampled patient, a factor indicating the source and gender of the patient, and lastly the type of cancer. Except for the last variable, cancer type, all variables were deleted, as they would not be used in the maxT and Bonferroni evaluations. Initially, the data contained information on 2 different kinds of ALL, ALLt and ALLb. Due to the low amount of observations, these were grouped into a single ALL variable. This also would simplify the Bonferroni and maxT analyses, as only 2 groups would have to be analyzed.

The data was loaded by using the *GolubEsets* package in R. This resulted in the value of the gene probes being normalized (Bolstad et al., 2003). This was not expected to form a problem for further analysis, therefore the normalized data was used. No missing values were present in the data.

3.5. Data Exploration of the Golub data

A goal set for this study was to investigate the potential difference in outlier robustness of maxT and Bonferroni. The independence of the tests is an important assumption for the Bonferroni correction. It is also important for the maxT method. To check this independence, the correlation between the different gene expressions would be of importance. Correlation between variables would also be an important factor to measure, as this could influence the power of both Bonferroni and maxT. Thus, the correlation between the gene probes of the Golub dataset was checked. This involved examining the pairwise relationships between the levels of different gene expression to identify potential causal patterns between two variables (Yule, 1897). This analysis can reveal whether genes show similar patterns across samples. This might suggest relationships between them. Using the Pearson correlation coefficients, the strength and direction of linear relationships between gene pairs could be

measured (Benesty et al., 2009). The Pearson correlation works by analyzing the cross-correlation and the variances of the variables. This results in the following formula: $r = \frac{E(ab)}{\sigma_a \sigma_b}$

The correlation is indicated by the *r*. *E(ab)* is the covariance of the two variables or, in this case, gene probes. The product of the standard deviations of both variables is used to divide this covariance. It is important to note that the indicated correlation between the variables is stronger if *r* approaches 1. A coefficient of 0 would suggest that there is no correlation.

In the context of high-dimensional gene expression data, this step could be crucial for understanding the underlying biological networks and discovering genes that might be related in some way (Langfelder & Horvath, 2008). The Golub dataset has information on over 7 thousand gene probes. This means that the data used is extremely high-dimensional and therefore the correlation evaluation could provide useful insight.

Following this, a correlation matrix was computed to evaluate the pairwise correlations between genes. This matrix was then visualized into a heatmap, which is shown in Figure 3. This heatmap also clustered the genes, but this clustering was not further elaborated upon in this study as, the Bonferroni or MaxT procedures only adjust for Type I errors due to a large number of simultaneous comparisons (Benjamini & Hochberg, 1995; Dudoit, Shaffer, & Boldrick, 2003).

Figure 3

Correlation heatmap of Golub data

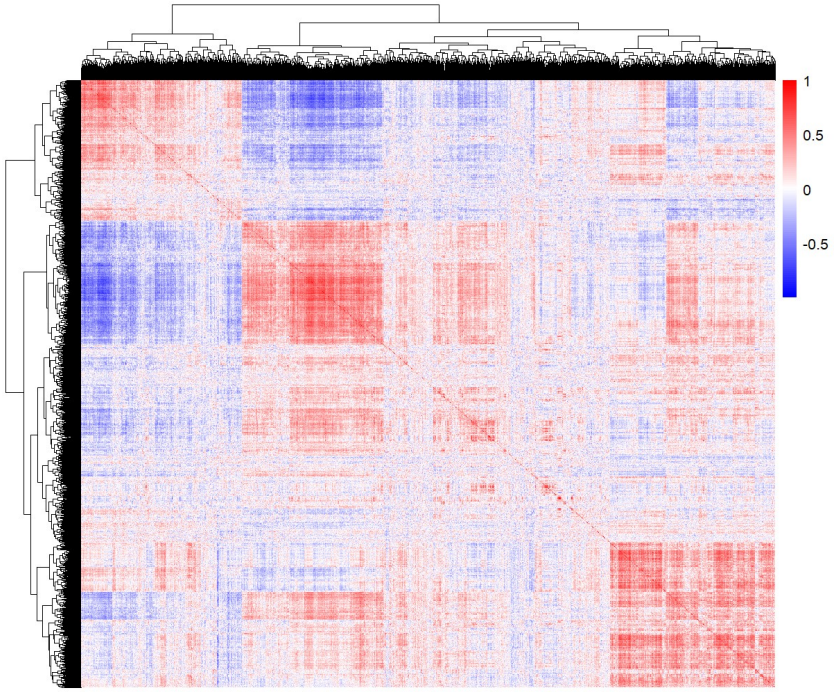


Figure 3 includes the correlation of every individual variable pair. Concluding, all 7,129 gene probes were included in this heatmap. Therefore, the heatmap includes over 50 million combinations. This

results in a heatmap that does not state a lot about individual pairs and correlations. Figure 3, however, offers a general view of different patterns. As is visible, several groups or clusters of gene probes show a strong correlation. 170 pairs even had an absolute correlation coefficient of 0.9 or higher. 446.305 combinations had a Pearson correlation coefficient with an absolute value equal to or higher than 0.5. This suggests that there are some strong relationships between gene probes, albeit that the correlations only apply to a small margin of the data, respectively 0.0007% and 1.8%. This could also indicate that there is no clear independence between all the tests.

To further investigate the independence of the Golub data, the Chi-Square test of independence was conducted on the data. The Chi-Square test measures a possible significant difference between two variables (McHugh, 2013). These two variables, however, should be categorical. Since the gene probe variables are numerical, these were first converted into 3 different groups, so the Chi-Square test could be conducted and a general conclusion about independence could be found. Chi-square compares the observed frequencies against the expected frequencies of the compared variables. The formula of chi-square is as follows: $X^2 = \sum \frac{(O-E)^2}{E}$

The X^2 represents the test statistic. O is the observed frequency, while E is the expected frequency. Using the degrees of freedom and the significance interval of 5% the null hypothesis is rejected or retained. The null hypothesis of the chi-square states that there is no relation between the variables. On the other hand, the alternative hypothesis states that there is a relationship between the variables. Thus, if the null hypothesis is rejected this indicates that there is no independence between the variables.

Since multiple variables were compared at once, the p-value was adjusted by the Bonferroni correction. This resulted in a very stringent p-value since all 7,129 variables were compared to each other. However, the chi-square test was only conducted to get a general overview of any possible dependence on the Golub data. Finally, 57,211 pairs were found to have a significant relationship between them. While these only were 0.23% of all tested pairs, this does indicate that there is some dependence in the data.

Concluding, the correlation and chi-square tests indicate that there are some relationships and correlations between different variables. Therefore, the tests suggest that there is no complete independence between the variables. This could thus influence the power of the Bonferroni correction and maxT method. However, the proportions of the correlation are very small and are not considered to form problems for further study.

4. Methods

The methods chapter will take a deeper dive into the math and statistics behind the researched methods. The methods will be discussed in the following order: Student's t-test, permutation test, Bonferroni correction, and maxT. Furthermore, based on the math and statistical properties of the methods the robustness to outliers will be estimated. After going into the methods separately, the processes of empirically comparing the methods will be explained. This explanation will contain the measures that were used to compare the methods will be elaborated upon.

4.1. Student's t-test

Student's t-tests, also referred to as t-tests, are commonly used statistical tools which check if there is a significant difference between the distributions of two groups, this is done by comparing the means of both groups (Livingston, 2004). T-tests offer a simple way to compare two groups. The null hypothesis of t-tests states that there is no significant difference between the groups being compared. The alternative hypothesis is that there is a difference.

There are two primary types of t-tests:

Independent two-sample t-test: This test is the standard form of the t-test and is based on multiple assumptions. One of these assumptions is independence between the groups.

Paired t-test: This form of t-test is used when the groups are related.

As stated before, t-tests rely on different assumptions to enable valid conclusions based on the test statistics and p-values. These assumptions are what make the t-test simple to use. However, they can be a major limitation to the usefulness, power and even robustness of the tests. The different assumptions that must be met in order to generate valid results with t-tests are (Widerberg, 2019):

Normality: The data should be normally distributed. This is further elaborated in the Data chapter.

Independence: Observations or samples for both groups should be independent. This assumption, however, does not apply to paired t-tests.

Homogeneity of Variance: For independent two-sample t-tests, the variances of the two groups should be the same. T-tests, thus, require homoskedasticity.

Equal sample sizes: If the sample sizes differ relatively much between the two groups, this could have a direct effect on the equality of the variances.

So, independent two-sample t-tests compare the observed means of two groups which are unrelated to each other. The formula for the t-test statistic when the variances of the two groups are assumed to be equal (pooled t-test) is (Teh & Abdul Rahman, 2009): $t = \frac{M_1 - M_2}{S}$

In this formula, M_1 and M_2 are the means of groups 1 and 2. They are divided by the standard error (S) to compute the test statistic *t.* To calculate the standard error, the standard deviation of group 1 (s_1) is divided by the square root of the number of observations in group 1 (N_1). This is then squared. This is then added to the square root of the same calculations for group 2. The standard error is finally derived by square rooting the sum of the measures of groups 1 and 2. This is also illustrated by the following formula: $S = \sqrt{\left(\frac{s_1}{\sqrt{n_1}}\right)^2 + \left(\frac{s_2}{\sqrt{n_2}}\right)^2}$

Interpreting the t-test results requires the degrees of freedom, *n* - 1. With the degrees of freedom and test statistic, using the t-distribution the corresponding p-value can be found. If this p-value is lower than the chosen significance level, 5%, the null hypothesis is rejected. In that case, the alternative hypothesis is accepted and it can be concluded that the groups differ significantly.

Some advantages and limitations of the test should be considered before the simulation results are discussed, this could provide better insights into the underlying components that influence the results of the simulation. It could also help when comparing the results of the difference in the simulations. The advantages of Student's t-test are:

Simplicity and Interpretability: Student's t-tests are straightforward to perform, requiring little computational resources. Interpreting the results is also relatively easy (Ruxton, 2006).

Power with Small Sample Sizes: T-tests can be powerful even with a relatively low number of observations included in the samples, providing valid results (Derrick et al., 2016).

Robust to Deviations from Normality: T-tests are relatively robust to moderate deviations from normality when the sample sizes are relatively large. This is due to the Central Limit Theorem (Lumley et al., 2002).

Naturally, Student's t-tests have limitations too, some of these limitations are:

Assumption Sensitivity: T-tests are based on several assumptions. Violations of these assumptions could lead to inaccurate and invalid results (Zimmerman, 2004).

Limitation with Non-Normal Data: When variables are not normally distributed and the sample sizes are relatively small, non-parametric tests might be a better fit (Gibbons & Chakraborti, 2011).

Concluding, Student's t-tests are a useful and versatile tool in statistical and data analysis. By allowing a comparison of means between groups when several assumptions are met. However, the presence of outliers could cause violations of these assumptions. This could lead to inaccurate or invalid results. Additionally, Student's t-tests heavily rely on means and variances. These statistics are more sensitive to outliers or extreme values than other statistics, like medians (Moore et al., 2016). This could reduce robustness. Outliers can inflate or shrink the t-statistic and p-value due to the shift in variance and means (Wilcox, 2012).

4.2. Permutation Tests

Permutation tests are non-parametric methods, meaning that they are not based on assumptions (Berry et al., 2014). They are used to determine if the distributions are equal between two groups. Their goal is thus similar to the Student's t-test (Hemerik & Goeman, 2017a). This study will solely focus on exact permutation and will disregard any other types of permutation tests. Permutation tests are based on reshuffling the observations or samples several times and evaluating the reshuffled data.

Permutation tests are not based on an underlying distribution, like Student's t-tests. Similar to t-tests the null hypothesis says that the groups do not significantly differ. The alternative hypothesis states that there is a significant difference between the compared groups.

Permutation tests 'permute' or shuffle the labels of all data points. After each permutation, the desired test statistic is obtained. Permutation tests can focus on different test statistics, like differences in means or medians (Good, 2000). The possibility of focusing on medians instead of means could provide better robustness to outliers, as medians are more resistant than means (Moore et al., 2016). Means tend to follow outliers or extreme values towards the skewness of the data. In the further parts of this study, the permutation test measuring the differences in means will be referred to as the 'mean permutation test'. The permutation test measuring the differences in medians will be referred to as the 'median permutation test'. The test statistics are used to determine the distribution of the test statistics. This is done by taking a sample of all possible permutations, as calculating all possible permutations would take an extremely long time, this is called approximate randomization tests. The p-value then is derived from the times that the permutated test statistic is bigger than the observed test statistic. This p-value can then be compared to the significance level, usually 0.05 (Noble, 2009). if the p-value is smaller than the p-value, the null hypothesis is rejected.

Permutation tests are based on only two assumptions:

Exchangeability: The data points must be able to be exchanged for both labels of the data.

Independence: Data observations must be independent of each other. This assumption is similar to the one for Student's t-tests (Ernst, 2004).

The procedure to execute a permutation test consists of 5 different steps:

1. **Definition of test statistic:** A test statistic *T* should be selected. This test statistic needs to be appropriate with regard to the tested hypothesis (Pesarin & Salmaso, 2010). For this study, the test statistics that will be used are the differences in means and medians between groups.
2. **Calculation of observed test statistic:** The test statistic for the observed data T_{obs} is calculated.
3. **Execution of permutations:** The (sampled) data is permuted *n* number of times. The level for *n* that will be used in the simulations for this study is 1.000 permutations. For each permutation the test statistic is calculated, T_i, where *i* is the iteration of the permutation.
4. **Formulation of distribution:** All test statistics of the permutations (T_1, T_2,..., T_n) are compiled to construct the null distribution of the test statistic.
5. **Calculation of p-value:** The p-value is derived from the proportion of the permuted test statistics that are as extreme or more extreme than the observed test statistic. This can also be formulated by the following formula (Ernst, 2004): $p = \frac{\sum_{i=1}^{n} I(T_i \geq T_{obs})}{n}$

   Where *I* is the indicator function measuring if T_i is as extreme or more extreme than the observed test statistic T_{obs}.

The calculated p-value is then compared to the significance level. And depending on the p-value being smaller than the significance level the null hypothesis will be accepted or rejected. And thus the conclusion about the similarities or differences between the groups will be drawn.

Like all other forms of statistical tests, permutations have advantages when compared to other methods, but the advantages will mostly be considered concerning a comparison with Student's t-tests. However, permutation tests have some limitations that need to be considered.

Some advantages of permutation tests are:

Non-parametric: Permutation tests are a non-parametric test, this implies that they do not rely on assumptions about, among others, the distributions. This makes them useful for non-normal data (Good, 2000). This could help the performance of the tests when outliers are included in the data. Permutation tests provide a distribution of test statistics. This discrete distribution offers a better approach than approximately probability values based on certain distributions.

Versatility: Permutation tests can measure different test statistics, like differences in means and medians. This provides a different approach compared to t-tests, which could be useful for certain situations. Medians are less likely to be influenced by outliers than means.

Resistance to extreme values: Appropriate permutation tests can be better resistant to values that can be considered extreme (Mielke & Berry, 2013). This could also apply to outliers that are generated by faults in the data gathering or cases with significantly different values than most other observations.

Some limitations of permutation tests are:

Computational Intensiveness: Permutation tests can require high computational resources as more calculations are executed to compute the p-values. This is especially true for large data sets (Ernst, 2004). Since the Hong Kong data has a lot of observations, the permutation test could require many resources and therefore take relatively longer to be executed.

Approximate randomization: Using a limited number of permutations instead of executing all possible permutations increases the chance of Type I errors. Therefore, comparing the Type I error of Student's t-tests and permutation tests on the same data could provide useful insights. Additionally, the sampling could cause different results to be generated even though identical protocols were followed (Mielke & Berry, 1994).

So, permutation tests could offer a flexible alternative to Student's t-tests. They allow assessments without relying on distributional assumptions which negatively impact the usefulness or validity. Permutation tests are possibly more robust to outliers as they are considered to be resistant to extreme values. However, only using a sample of all possible permutations could limit results. Since permutation tests can focus on different statistical differences between groups, like means and medians, permutation tests could offer versatility when a statistic is skewed due to outliers. Student's t-tests only focus on means, so permutation tests could gain power from using less sensitive statistics as a basis for the evaluations.

4.3. Multiple Test

The methods that have yet been discussed can be applied to determine any possible similarities or differences between 2 groups. Both Student's t-tests and permutation tests can, however, only be applied to test a single variable at a time. In some situations, it can be required to test differences or similarities between groups based on multiple variables. These situations require a form of multiple test (Jafari & Ansari-Pour, 2019). Multiple test allows the analyst to compare groups based on the values of multiple variables. So, the difference between groups can be determined by a wider range of factors.

This could provide useful insights when the evaluated data consists of a relatively bigger number of variables.

Important to note when discussing multiple tests is the increased possibility of Type I errors. Each single statistical test bears upon it the chance to result in a false positive (Lin, 2015). When the null hypothesis is rejected when it should not have been rejected, this is called a false positive. The chance of a false positive occurring is equal to the significance level, mostly 5% (Noble, 2009). This chance is for a single test. So naturally, when multiple tests are conducted simultaneously, the results probably include one or more false positives (Jafari & Ansari-Pour, 2019). This is called the multiple comparisons problem (MCP). For example, if each test is conducted at a significance level of α and the tests are all independent, the chance of the results including at least one Type I error in m number of tests is (Bland & Altman, 1995): $1 - (1 - \alpha)^m$

If α = 0.05 and 20 variables are tested, so m = 20. This results in $(0.95)^{20}$ = 0.36. This is the chance that no Type I error will occur. This is nearly one-third of the original chance of 0.95 of a Type I error occurring. The chance of at least one significant error occurring is then 1 - 0.36 = 0.64. Then the chance of at least one significant Type I error is 0.64. This is almost 13 times higher than the original chance of 0.05.

In order to combat this increase in the probability of false positives, methods were created to lower the chance of false positives. Two examples of these methods are the Bonferroni correction by Dunn (1961) and the maxT method by Westfall and Young (1993). these methods are used to control the familywise error rate (FWER). Multiple test methods are widely used in studies using data on genes as these datasets often consist of a large number of variables (Jafari & Ansari-Pour, 2019).

Important to note is that multiple tests, in principle, are not test methods. Multiple test methods only help with controlling the FWER. So, it is used as an extension of a statistical test, like the t-test or permutation test. Therefore, the Bonferroni and maxT methods were based on the Student's t-test in the simulations part of this study.

4.4. Bonferroni Correction

The Bonferroni correction is a classical statistical method used to correct the MCP (Benjamini & Hochberg, 1995). This is done by adjusting the significance threshold. The Bonferroni correction helps to control the FWER, so the desired level of significance is maintained across multiple tests. Specifically, if m independent tests are conducted, the Bonferroni correction adjusts the significance level for each test to be α'. And α' is calculated by the following formula (Jafari & Ansari-Pour, 2019): $\alpha' = \frac{\alpha}{m}$

Where α is the desired overall significance level, in most situations this would be 5% or 0.05 (Noble, 2009). This means that each hypothesis test is conducted with a more restricted significance level. This restricts the otherwise increased risk of Type I errors that occur with multiple tests.

If the previous example of 20 tests with a significance level of 0.05 is now analyzed with the Bonferroni correction. This would result in the following adjusted p-value: $\alpha' = \frac{0.05}{20} = 0.0025$

So, each test will be conducted and the null hypothesis is only rejected when the p-value is equal or lower than 0.0025 instead of the original 0.05 level. The chance of no Type I error occurring among all tests is: $1 - (1 - \alpha)^m = (1 - 0.0025)^{20} \approx 0.951$

Thus, the chance of at least one Type I error occurring is (1 – 0.951) = 0.049. This is approximately equal to the original 0.05 chance for a single test. Without adjusting the p-value the chance of at least one Type I error was 0.64. Using this adjusted threshold, the Bonferroni correction ensures that the probability of making one or more Type I errors across all tests is maintained at the desired α level.

The Bonferroni correction is straightforward and simple to implement. This simplicity has a downside, Bonferroni is also to be (very) conservative (VanderWeele & Mathur, 2018). This conservative effect is magnified when the number of tests increases as α' will shrink in size. The conservatism and shrinkage of α' potentially lead to a statistical power reduction, which in turn increases the likelihood of Type II errors (false negatives). This is when true effects are missed, so the null hypothesis is accepted while it should have been rejected. Despite this conservativeness, the Bonferroni correction is widely used because of its simplicity and effectiveness in controlling the FWER.

The implementation of the Bonferroni correction only requires three steps:

1. **Identification of the number of tests:** The first step to applying the Bonferroni correction is to determine how many tests will be simultaneously conducted. In the simulations of this study, all gene probes of the Golub dataset will be tested at once, so 7,129 tests (*m*).
2. **Calculation of adjusted significance level:** The desired significance level should be calculated using the stated formula: $\alpha' = \frac{\alpha}{m}$

   So, the desired level of α should be divided by the number of tests.
3. **Application of the adjusted significance with tests:** The acquired significance level α' should next be applied to evaluate the hypotheses of each individual test. This test could be any statistical test that works with a p-value and significance level.

Some advantages of the Bonferroni correction are:

Simplicity: Understanding and implementing the Bonferroni correction is relatively easy, as it requires a single action or calculation to implement over a large number of tests (Armstrong, 2014).

Conservativeness: Bonferroni provides strict control over the Type I error rate. This ensures that the desired level α is not exceeded (Nakagawa, 2004).

Some limitations of the Bonferroni correction are:

Conservativeness: While conservativeness is a strength of Bonferroni correction controlling the Type I error rate, the conservativeness can also have a negative effect when the number of tests increases. This can lead to an increasingly higher risk of Type Type errors. Thus, the chance of overlooking meaningful effects increases (Perneger, 1998).

Independence: Bonferroni assumes that all tests are independent. When there is a correlation between the tests, Bonferroni could be too stringent. This unnecessarily reduces statistical power (Holm, 1979).

In conclusion, the Bonferroni correction is a classic method for handling the multiple comparisons problem in situations with multiple tests, as it ensures control over the overall Type I error rate. While its simplicity and conservativeness form advantages, the method's conservative nature can be a limitation in scenarios with a large number of tests, especially when they are also correlated. The Bonferroni correction does not have a direct influence on the robustness of outliers, as it is a relatively passive method. Bonferroni adjusts the significance level based on the number of tests. Since Bonferroni does not look at values in the data, it does not offer much robustness. If the p-value from the t-test is shrunken due to outliers, this shrinkage could cause the p-value to pass the adjusted significance level. Furthermore, if the p-value is inflated due to the outlier, the Bonferroni conservativeness causes the null hypothesis to be incorrectly accepted.

4.5. MaxT

The maxT method, like Bonferroni, is a technique used to control the FWER in MCP situations (Westfall & Young, 1993). Unlike the simpler Bonferroni correction, which applies a single adjustment to all significance levels of all individual tests by adjusting α for the number of tests, the step-down maxT method uses the data to find the distribution of the test statistics. Therefore, maxT is considered to be more powerful than the Bonferroni correction (Goeman & Solari, 2014). The method adjusts p-values by permutating the data and computing a distribution of the maximum test statistics for each permutation. This approach takes into account the structure of tests. This offers more efficient control over the FWER when compared to simpler methods like the Bonferroni correction (Westfall & Young, 1993).

The maxT method is based on an assumption:

Subset pivotality: The maxT method considers subset pivotality. This means that for each subset of the hypotheses, the distribution would be the same as the distribution of all the hypotheses (Dudoit et al., 2003 & Westfall & Young, 1993). It could be that subset pivotality holds when each individual test only depends on the observations for the variable tested. However the correctness of this statement is disputed (Rempala & Yang, 2013).

The procedure of conducting the maxT method for multiple tests has several steps (Westfall & Young, 1993):

1. **Definition of the test statistic:** The appropriate test statistic hypothesis test should be selected, as stated before for this study the test statistic was the t-statistic derived from the Student's t-test.
2. **Calculation of the observed test statistics:** The test statistic for each hypothesis test, e.g. the t-statistic for every variable/gene probe analyzed, using the original data. These statistics are named $T_1, T_2, \ldots, T_m$, where *m* is the total number of tests or variables.
3. **Generation of permutations:** The labels for the original data are reshuffled or 'permuted' *n* times. For each permutation the test statistics are calculated, resulting in a vector of statistic of $T_1^1, T_2^1, \ldots, T_m^1$.
4. **Computation of the maxT distribution:** For each permutation the maximum test statistic is selected. This can be denoted as the following formula: $T_{max}{}^1 = \max(T_1{}^1, T_2{}^1, \ldots, T_m{}^1)$
   From these maximum test statistics from each permutation, the null distribution of the maxT distribution is created. The distribution would be made up of all the permuted maximum test statistics for all permutations (n): $T_{max}^1, T_{max}^2, \ldots, T_{max}^n$.
5. **Sorting of observed test statistics:** The test statistics that were computed using the original data should be sorted in descending order: $T_1 \geq T_2 \geq \cdots \geq T_m$
6. **Adjustment of p-values:** For every observed test statistic the adjusted p-value is calculated. The adjusted p-value is the proportion of permuted maximum test statistic values (T_{max}^n) that are bigger than the observed test statistics divided by the total number of permutations (*n*). This results in the following formula for the adjusted p-values (Westfall & Young, 1993):
   $$p_i^{adj} = \frac{\sum_{n=1}^{n} I(T_{max}{}^n \geq T_i)}{n}$$
   Where *I( )* is an indicator function that checks whether the permuted test statistic T_{max}^n is bigger than the observed statistic T_i.
7. **Rejection of hypotheses using step-down method:** The adjusted p-values are then evaluated using the original significance level α_1. If the adjusted p-value is smaller than the significance

level, the hypothesis is rejected. The values corresponding to the rejected hypotheses are removed and the new significance level α_2 is calculated. All adjusted p-values that are smaller than α_2 are rejected and their values are removed. This process is repeated until the removal does not lead to new rejections.

The maxT method is considered a powerful tool for controlling the FWER in scenarios that test multiple hypotheses at once. The adjustment is stringent by using permutations and focussing on the maximum test statistics of these permutations. The step-down procedure ensures that the hypotheses are rejected in a downward step-wise manner, ending when stepping down does not result in extra rejections. This helps maintain the overall error rate while providing a clear criterion for statistical significance. This method is a good fit for high-dimensional data. The Golub data used in the simulations has 7,129 gene probes and is thus highly dimensional.

Some advantages of the maxT method are:

Correlation allowance: The maxT method is more powerful than simpler methods like the Bonferroni correction. Its increased power is due to the method considering the correlation structure among tests (Dudoit et al., 2003).

Adaption of data: MaxT uses permutations to compute a distribution of the test statistics. This makes the adjusted p-value and significance level more accurate compared to those derived from the stringent and conservative Bonferroni correction (Westfall & Young, 1993).

Some limitations of the maxT method are:

Computational Intensity: The maxT method requires extensive computational resources. This is particularly true when used for large datasets or when the selected number of permutations is high (Nichols & Hayasaka, 2003). Since, maxT requires more calculations than Bonferroni, and the datasets used for, for example, research into gene expressions are commonly large, this could limit usefulness.

Use of maximum statistics: MaxT relies on the maximum values and this could pose a problem when outliers inflate these maximum statistics (Westfall & Young, 1993). Furthermore, outliers in the original could negatively affect the distributions of the permuted test statistics and thus the results of the statistical analysis.

The maxT method is, potentially, a flexible and powerful tool for controlling the family-wise error rate in multiple hypothesis test scenarios. When test statistics are correlated maxT offers more statistical power than the Bonferroni method. Since the Golub data included some correlated gene expressions this could affect the power and robustness of outliers too when compared to Bonferroni. Since maxT

evaluates the data, the effect of outliers on the original t-statistic is potentially weakened. This is because the same effect applies to the permuted statistics.

4.6. Simulation of Single Comparison Tests

Since the principles of the research methods are discussed, the protocols used to evaluate the robustness of these methods to outliers are laid out. The first simulations were used to compare and evaluate the robustness of Student's t-tests and permutation tests. In the Theory chapter of this study, two scenarios were laid out in which the robustness should be compared. The first scenario that was laid out, stated that the data should follow a normal distribution because this could have a positive effect on the power of Student's tests. The second scenario that was proposed, was a comparison with the data having equal variances between the groups being compared. Simulations for both scenarios were included in the data analysis. However, the first simulations included no special tests. All simulations were conducted for both the Height and Weight variables, in order to evaluate and compare possible differences. Possible differences could provide useful insights. Additionally, since the Height and Weight variables included in the Hong Kong data differed slightly, this could also confirm the hypotheses for different variable types.

The base for all simulations was the same and included the following steps:

Step 1: For all simulations, samples of the observations for the corresponding variable were selected. The selection was conducted randomly. First, from the 25,000 observations of the Hong Kong data 1,000 were randomly selected for group 1. Group 2 consisted of 1,000 observations that were randomly picked from the remaining 24,000 observations, ensuring no overlap between groups (Ripley, 2009).

Step 2: To establish a baseline for the simulation, an initial hypothesis test was conducted on the two groups for both the Student's t-test and the permutation test. The Student's t-test assessed the difference in means between the two groups. Additionally, two permutation tests were performed, evaluating the differences in means and medians. Both permutation tests used 1,000 permutations. The p-value resulting from the three different tests was stored for later comparison after the outliers were added.

Step 3: To examine the robustness of the tests, an extreme outlier was introduced into Group 1. This introduced outlier had a size of 10 times the maximum value found in Group 1. This addition was intended to simulate a typo or error in the data-gathering process resulting in an extreme data point. By adding an extreme value the effect of outliers could be enlarged which could have highlighted the effect.

Step 4: After adding the outlier, the three tests, Student's t-test, mean permutation test, and median permutation test, were conducted again. This time the tests used the data that included the outlier. Both permutation tests used 1,000 permutations again. The p-values of the tests were stored.

Step 5: Next to the change in p-value, the number of times the significance was changed due to the addition of the (extra) outlier was stored as output. This number of changes in significance stated the Type I errors that occurred due to the addition of an (extra) outlier. If the test before the addition of the (extra) outlier did not state a significant test statistic or p-value, but the test after the addition of an (extra) outlier did indicate a significant test statistic or p-value, this would mean that the addition of the outlier had a direct result on the result and therefore the conclusion of the test. In other words, if the test would be robust to outliers the addition of an (extra) outlier should not be able to cause a difference in the result of the test. If the significance of the test result was changed by the addition of an (extra) outlier, this would mean that the test result was a false positive or Type I error.

All simulations tested the robustness in the different scenarios and included 200 iterations. These iterations ensured reliability and generalizability. For each iteration, the steps were conducted and the differences in p-values were stored. This means that the output of the simulation included 200 calculated differences (Efron & Tibshirani, 1994).

After the simulations were run, the difference in p-values could be used to evaluate if there was a significant effect of the outlier on the results of the three tests. To evaluate the potential significance of these differences, the Wilcoxon signed-rank test, which measures differences between two values, was applied This test uses (Wilcoxon, 1945). The formula of the Wilcoxon signed-rank test statistic W is: $W = \min(W^+, W^-)$

Where W^+ is the sum of the positive difference and W^- is the sum of the negative differences.

This test was chosen because it allows researchers to determine if there is a significance in the differences between the two groups (Gibbons & Chakraborti, 2011). The Wilcoxon signed-rank test was used on the differences in p-values between the tests before and after adding the outlier. Thus, the Wilcoxon test assesses whether the introduction of the outlier led to significant changes in the p-values of the respective tests. The results from the simulations highlighted the robustness of the permutation tests compared to the Student's t-test.

The first simulation included no further tests and consisted of only this process. This simulation was, like all other simulations, conducted using both the Weight and Height variables. The second simulation tested the robustness of the tests when the sampled data was normally distributed. Student's t-tests

could be more robust to outliers in larger sample sizes. With data samples including 1,000 observations, the sampled data could follow a normal distribution better (Kwak & Kim, 2017).

In order to test normality, the Shapiro-Wilk tests were performed on both groups 1 and 2 (Shapiro & Wilk, 1965). The Shapiro-Wilk test is further elaborated upon in the data chapter of this study. The normality test was conducted between **Step 1 and Step 2** of the base simulation. When the output from the Shapiro-Wilk test states that one or two groups of the iteration did not follow a normal distribution, the iteration was skipped. This step is crucial as the Student's t-test assumes normality, thus checking whether the analysed data is normally distributed could improve the robustness to outliers. This could even cause a difference in the comparison between the Student's t-test and the permutation tests.

The next simulation compared the robustness of the tests in scenarios when the variances of both groups were equal. The possible equality of variances was tested using Levene's test (Levene, 1961). Levene's test is a widely used method to determine equality between group variances (Brown & Forsythe, 1974). The test statistic *W* is calculated as follows: $W = \frac{(N-k)}{(k-1)} = \frac{\sum_{i=1}^{k} N_i(Z_i - Z_{\dots})^2}{\sum_{i=1}^{k}\sum_{j=1}^{N_i} N_i(Z_{ij} - Z_{\dots})^2}$

Where k is the number of different groups, N is the number of observations, n_i is the observation in group I, Z_i is the mean of the absolute deviation in group I, and Z is the sum of all Z_{ij}.

Whether the outcome of Levene's test is significant, and it can be concluded that there are significant differences between the variances of the group, is determined by evaluating the test statistic *W* against the critical value derived from the F-distribution with $k - 1$ and $N - k$ degrees of freedom. The p-value can be derived from calculating the right-tail probability of the corresponding F-distribution. If the p-value is smaller than the significance level of 0.05, the variances of the groups are not equal. If the output from Levene's test concluded that the variances were not equal the iteration was skipped.

The simulations explained above were all used to answer sub-question 3, which researched the robustness of both methods and the comparison between the two single comparison tests. However, sub-questions 5 and 6 focused on the effect of an increasing number of outliers and an outlier growing in size on the results from the tests. Two different simulations were conducted to test these effects. These simulations were based on the same basis as the other simulations. However, these simulations include 50 iterations of the base simulation which also tested the normality of the data.

The simulation that tested the effect of an increasing amount of outliers, consisted of 50 iterations, where an extra outlier was added every iteration. For every outlier adding iteration, a simulation of 100 iterations was conducted. The 50 outliers were all added to group 1. For each outlier added, the average

p-value for each of the three tests compared of the 100 iterations was stored. Furthermore, the number of changes in significance stated the Type I errors that occurred due to the addition of an (extra) outlier.

This concept was applied the same way when the simulations focused on a growing outlier. But instead of an extra outlier being added, the single outlier grew a factor. The outlier grew exponentially. The size of the single outlier was calculated as follows: $Outlier\ value = initial\ outlier * 2^i$

Where *i* is the iteration of the simulation. The initial was defined as the maximum value of the variables of the original dataset, including 25,000 observations. This simulation was run 51 times. For the first iteration, the value of *i* is zero. So, the outlier was exactly the same as the initial outlier for the iteration. From the initial iteration, the outlier doubled in size.

4.7. Simulation of Multiple Comparison Tests

The multiple comparison tests were compared using the Golub dataset. Important to note is the fact that the data consists of 72 observations which are divided into 2 groups: AML consisting of 47 observations and ALL consisting of 25 observations. This indicates an unequal division between the two groups as the proportion between the two groups is 65%-35%. So, if a single outlier was randomly added to either one of the groups, the results could differ depending on which group the outlier was introduced to. In order to test if there is a difference between the two groups and to validate results, every simulation comparing the maxT and Bonferroni methods was conducted twice. In the first simulation, the outliers were added to the AML group and the second run introduced the outlier to the ALL group.

Contrary to the single comparison tests no specific situations were set out to test, so the simulations for maxT and Bonferroni did not include specific tests about the data. The simulations about maxT and Bonferroni did not sample the data. Instead, the entire Golub dataset was used, so all 72 observations and 7,129 gene probes were used to compare the robustness of the methods. Since no specific scenarios would have been researched, the same simulation was used for all tested applications. These applications are the robustness to a single outlier, the effect of an increasing number of outliers, and the effect of an outlier growing in size. Thus, a base simulation was constructed, it consisted of several steps:

Step 1: First, the baseline p-values were conducted. The Student's t-test was used for this. This test was used because of its simplicity and because it requires relatively little computational resources. The initial test tested the difference between both groups for all gene probes. The values of the gene probes were normalized, so the assumptions of the Student's t-test should hold.

Step 2: Using the Bonferroni correction and maxT method, which used 1000 permutations, the p-values were adjusted. These adjusted p-values were used to find the baseline significant genes. This baseline stated the number of genes that showed a significant difference between the two groups. The adjusted p-values and the number of significant gene expressions were stored for later comparison.

Step 3: After determining the baseline, the outlier was introduced into the specified group. The outlier was added as an extra observation. The gene probe values of a randomly selected observation were multiplied by the outlier factor and the label of the predefined group was added. The outlier factor was set as 10 for the single outlier simulations, but for the increasing number of outliers, the factor was set as 2. For the growing outlier simulation, the factor had other sizes.

Step 4: After adding the outlier, the Student's t-test was conducted again, now using the dataset including the outlier. The adjusted p-values were also redetermined.

Step 5: The differences in the average and median p-values between the baseline and outlier test results were calculated and stored. A possible difference in the number of significant was also calculated and stored. This difference would indicate the Type II errors that were caused by the outlier.

Step 6: The Wilcoxon signed-rank test was used to determine if the differences calculated in **step 5** were significant (Wilcoxon, 1945). The significance level for this test was set at 0.05.

The output of this simulation contained the outlier size and number of outliers, the number of significant genes for both the baseline and final version, the mean and median p-value differences, and the Wilcoxon test result.

This simulation was used to answer sub-question 4. To answer sub-question 5 a simulation was run where the number of outliers was increased. For this simulation, the group to which the outliers were added was not predefined, so each outlier added was given a random label. In total 50 observations containing outliers were added. Each iteration introduced an extra outlier. The factor with which the values of the selected observation were added was set to 2.

To answer sub-question 6 a single outlier was multiplied with a bigger factor every iteration. So, in the first iteration the gene probe values of the selected outlier were multiplied by one, and the iteration after the values of the original observation were multiplied by two. For the last iteration, the values were multiplied by 50, as 50 iterations were conducted. This simulation was once run where the growing outlier was given the AML label and once where the outlier was added to group ALL.

All simulations used in this were run in the statistical computing software package R (R Core Team, 2024). Different R packages were used for the calculations of this study. The packages used included:

‘multtest’ (Pollard et al., 2005), ‘golubEsets’ (T. Golub, 2024), ‘ggplot2’ (Wickham, 2009), ‘reshape2’ (Wickham, 2007), ‘beepr’ (Bååth & Dobbyn, 2024), ‘pheatmap’ (Kolde, 2019), ‘corrplot’ (Wei & Simko, 2019), ‘ggpubr’ (Kassambara, 2023), ‘nortest’ (Gross et al., 2015), ‘stats’ (R Core Team, 2024)., ‘dplyr’ (Wickham et al, 2023), ‘xtable’ (Dahl et al., 2000), ‘car’ (Fox & Weisberg, 2020).

5. Results

The Results chapter presents the findings of the simulation study evaluating the robustness of classical statistical methods: the Student's t-test and the Bonferroni method, against permutation methods: permutation test and maxT method. This chapter will first discuss the single comparison methods, Student's t-tests and permutation tests. Next, the multiple comparison test methods will be discussed.

5.1. Single Comparison Robustness

The first simulation conducted did not consider any additional tests when comparing the Student's t-test, mean permutation test, and median permutation test. So, the data samples used in the simulations could still follow assumptions like normality and variance equality, but this was not verified. A single outlier was added to a group of 1,000 observations. This outlier was the maximum value of group 1 multiplied by 10.

In total 200 iterations were run. To determine the impact of the outlier, the differences in p-values of the tests conducted before and after adding the outliers were calculated for every iteration. The simulation, thus, resulted in a list of 200 differences for all three tests. To determine whether the differences in p-values were significantly bigger than zero, the Wilcoxon signed-rank test was conducted.

Table 4

Overview of results of simulation without test, including Wilcoxon signed-rank test on p-value differences, average difference in p-value and Type I error rate caused by the addition of the outlier for the Student's t-test, mean permutation test and median permutation test

Test	*Variable*	*Wilcoxon P-value*	*Average difference in p-value*	*Type I error rate*	*Test type*
Student's t-test	Height	0.000***	-0.197	0%	No Test
Permutation test, Mean	Height	0.207	0.007	1%	No Test
Permutation test, Median	Height	0.414	-0.002	0%	No Test
Student's t-test	Weight	0.000***	-0.118	0%	No Test
Permutation test, Mean	Weight	0.130	0.005	1.5%	No Test
Permutation test, Median	Weight	0.580	-0.001	0.5%	No Test

Note: The p-values are rounded to three decimals for better readability and interpretability. The stars indicate significance levels: *** indicates $p < 0.001$, ** indicates $p < 0.01$, and * indicates $p < 0.05$. The Type I error rate indicates the proportion of the iterations where a false positive was caused by the introduction of the outlier.

The p-values in Table 4 show that the test results and p-values for the Student's t-test were significantly different after adding the outlier. This is true for both the Height and Weight variables. This indicates that the results of the Student's t-tests were significantly influenced by the outliers. Possible explanations for the effect of the outlier could be that the outlier inflates the mean of group 1 while the mean of group 2 remains unchanged. The extreme outlier could also result in a violation of the assumptions of Student's t-test. However, according to Lumley et al. (2002), normality is not

necessarily needed for large samples, as t-tests could be valid for large samples for any distribution. However, no extra Type I error was caused by the addition of the outlier.

No significant difference was found for the permutation tests, thus, the outlier did not have a significant effect on the p-value. A few Type I errors occurred for the permutation tests, but these are relatively uncommon, occurring in approximately 1% of the iterations. So, if no tests on the sampled data are conducted the permutation tests are more robust to outliers than the Student's t-test.

Building on the initial robustness assessment, the next simulations run included normality tests about the sampled data before the outlier was added. If either group did not follow normality, the iteration was skipped. This resulted in, respectively, 17 and 18 iterations being skipped. This also indicates that some iterations in the simulation without tests contained groups which were not normally distributed.

Table 5

Overview of results of simulation with normality test, including Wilcoxon signed-rank test on p-value differences, average difference in p-value and Type I error rate caused by the addition of the outlier for the Student's t-test, mean permutation test and median permutation test

Test	*Variable*	*Wilcoxon P-value*	*Average difference in p-value*	*Type I error rate*	*Test type*
Student's t-test	Height	0.000***	-0.170	0%	Normality Test
Permutation test, Mean	Height	0.754	0.029	2.2%	Normality Test
Permutation test, Median	Height	0.366	0.002	1.1%	Normality Test
Student's t-test	Weight	0.000***	-0.133	0%	Normality Test
Permutation test, Mean	Weight	0.130	0.001	5.5%	Normality Test
Permutation test, Median	Weight	0.580	0.002	1.6%	Normality Test

Note: The p-values are rounded to three decimals for better readability and interpretability. The stars indicate significance levels: *** indicates $p < 0.001$, ** indicates $p < 0.01$, and * indicates $p < 0.05$. The Type I error rate indicates the proportion of the iterations where a false positive was caused by the introduction of the outlier.

Table 5 shows that the p-values of the Student's t-test were significantly influenced by the addition of a single outlier. Even though the normality assumption holds, before adding the outlier, the Student's t-test is still not more robust to outliers compared to the permutation tests. The mean permutation test, however, shows a relatively high Type I error rate being caused by the addition of an outlier compared to the other tests. Indicating that the test results are also influenced by the addition of the outlier. This is especially true for the Weight variable where Type I errors were triggered in about 5,5% of the conducted iterations. This is a substantial proportion which could indicate weaker robustness to the extreme outliers. For the Height variable, this proportion is approximately 2.2%, which is higher than the other two tests. This will be further investigated in the simulation adding multiple outliers. The median permutation test shows no explicit effect of the outlier, apart from a relatively low Type I error rate. The difference between the Type I errors of the mean and median permutation tests could be explained by medians being less affected by extreme values than means (Moore et al., 2016).

So, when the normality assumption holds for the data, both types of permutation tests, are more robust to outliers than the Student's t-test. But, the mean permutation tests show a tendency to produce more Type I errors.

An additional scenario in which the robustness of the Student's t-test could be better than that of permutation tests could be when the variances of both groups were (approximately) equal (Sawilowsky & Blair, 1992 & Field, 2017). For this simulation, the same principle as the normality test was used, so if Levene's test indicated that the variance between groups was unequal, the iteration was skipped. For the simulation using the Height variable 190 iterations were completed, so in 10 iterations the variances were unequal. The simulation analyzing the Weight variable skipped 7 of the 200 iterations, completing 193 iterations.

Table 6

Overview of results of simulation with variance equality test, including Wilcoxon signed-rank test on p-value differences, average difference in p-value and Type I error rate caused by the addition of the outlier for the Student's t-test, mean permutation test and median permutation test

Test	*Variable*	*Wilcoxon P-value*	*Average difference in p-value*	*Type I error rate*	*Test type*
Student's t-test	Height	0.000***	-0.192	0%	Variance test
Permutation test, Mean	Height	0.050*	-0.009	2.1%	Variance test
Permutation test, Median	Height	0.957	0.000	0%	Variance test
Student's t-test	Weight	0.000***	-0.133	0%	Variance test
Permutation test, Mean	Weight	0.1966	0.005	3.1%	Variance test
Permutation test, Median	Weight	0.1279	-0.002	1.0%	Variance test

Note: The p-values are rounded to three decimals for better readability and interpretability. The stars indicate significance levels: *** indicates $p < 0.001$, ** indicates $p < 0.01$, and * indicates $p < 0.05$. The Type I error rate states the proportion of the iterations where the null hypothesis was rejected after adding the outlier(s) when the null hypothesis was accepted before adding the outlier(s).

The Student's t-test again saw that the added outlier influenced the test statistic and p-value. The Wilcoxon test indicates that for both the Height and Weight variables the difference in -p-value was significant. Additionally, the mean permutation test was significantly influenced by the outlier. This is the only simulation that resulted in a significant difference in p-values for a permutation test. A possible explanation for this could be that adding an outlier has a bigger effect on the mean if the variance between groups is equal (Good, 2005). The addition of an outlier could have a more pronounced effect on the distribution of the permuted means.

However, an interesting statistic is the Type I error rate. For all simulations introducing a single outlier, the Student's t-test did not have any instances where a result was insignificant before the outlier addition but was significant after the outlier addition. The p-value has, however, been significantly affected by the addition of an outlier. This could be due to the Student's t-test resulting in mostly significant p-values before adding the outlier, so the addition would not cause a Type I error. Another

possible explanation could be that the p-values for the Student's t-test were higher for the baseline group. Then a big change could fail to make the p-value cross the significance level threshold.

In conclusion, the Student's t-test p-value is significantly changed by the addition of an outlier. The median permutation test is the most robust to outliers in the case of homogeneity. The mean permutation test is less robust than the median version because, in the Height simulation, the p-values were significantly changed. Furthermore, the mean permutation test is more prone to result in Type I errors than both the median permutation test and the t-test.

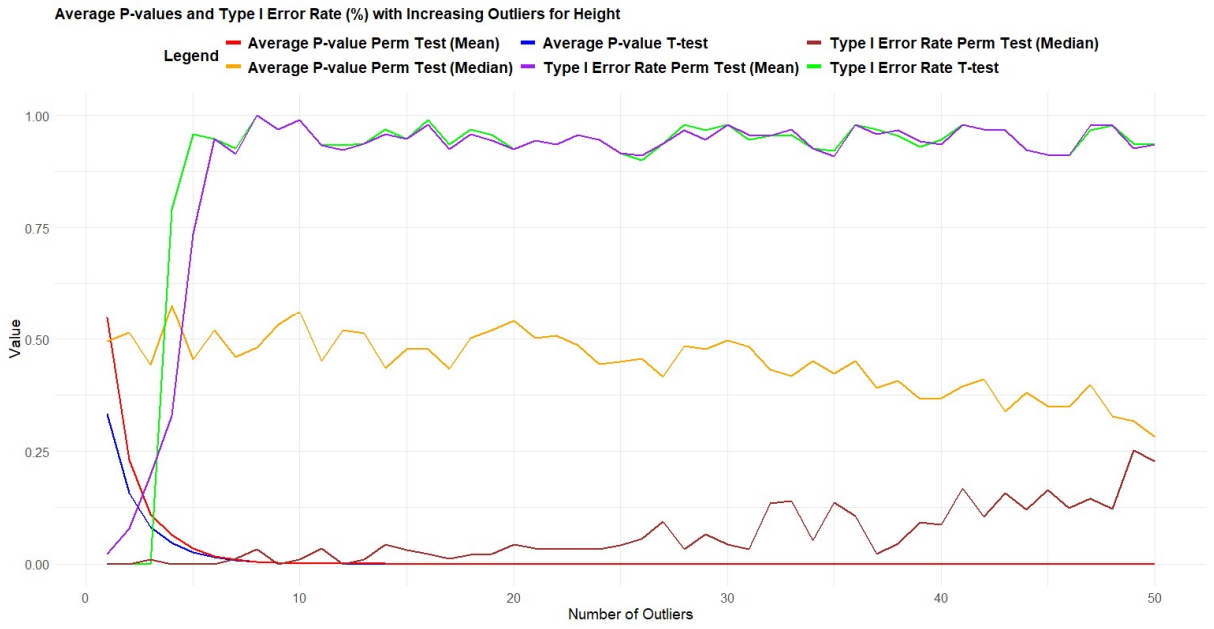
5.2. Single Comparison Robustness with Increasing Number of Outliers

The previous simulations focused on adding a single outlier in some special situations. These simulations showed a clear rank in the robustness of the different tests. The median permutation test was the most robust. And the Student's t-test was the least robust.

This study also aimed to research the effect of adding multiple outliers on the outcome of the tests. The proportion of the outliers of the entire group 1, including the outliers, ranged from about 0.1% to about 4.8%. So, the outliers made up a relatively small part of group 1. Figure 4 shows the outcome of the simulation for Height, the outcome for the Weight variables is shown in Figure 5.

Figure 4

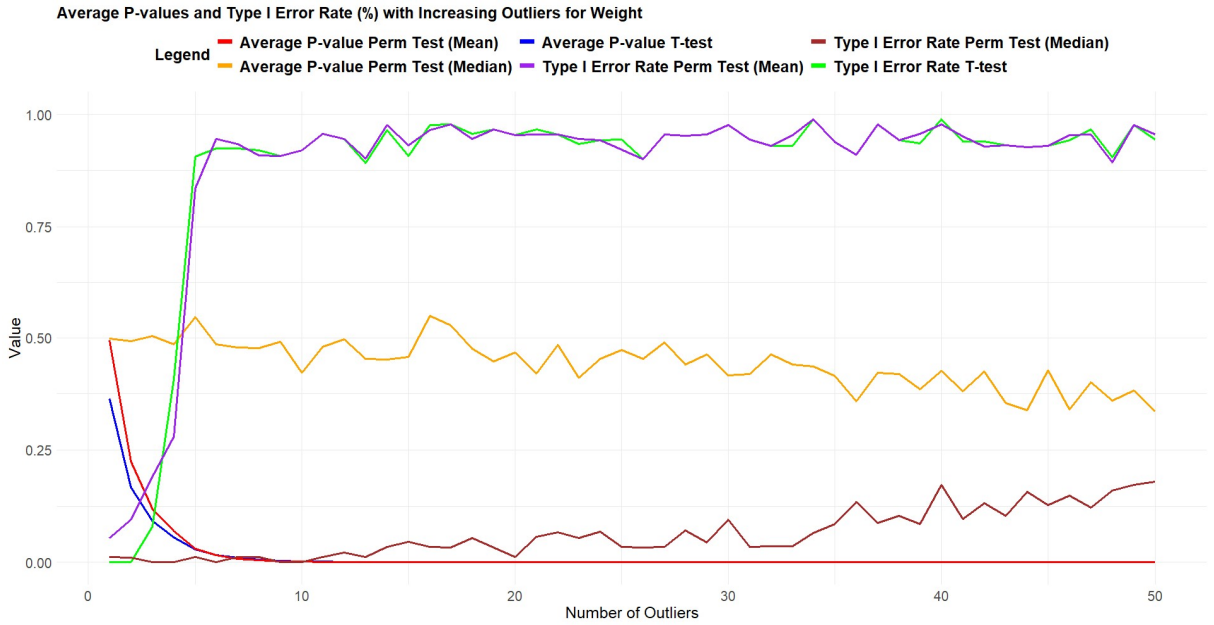
Overview of results of simulation with increasing number of outliers for Height, including average p-values and Type I error rate (%) caused by the addition of the outlier for the Student's t-test, mean permutation test and median permutation test



Note: The Type I error rate states the proportion of the iterations where the null hypothesis was rejected after adding the outlier(s) when the null hypothesis was accepted before adding the outlier(s).

Figure 5

Overview of results of simulation with increasing number of outliers for Weight, including average p-values and Type I error rate (%) caused by the addition of the outlier for the Student's t-test, mean permutation test and median permutation test



Note: The Type I error rate states the proportion of the iterations where the null hypothesis was rejected after adding the outlier(s) when the null hypothesis was accepted before adding the outlier(s).

Both figures show a similar pattern. The Student's t-test and mean permutation test perform similarly to the Type I error and average p-value. The average p-values shrink when extra outliers are added. While the average p-value of the Student's t-test is lower for the first outlier, the p-value of the mean permutation test shrinks faster, equaling zero when the 11th outlier is added. The p-value of the t-test equals zero after the 13th outlier is added. The average p-value equals zero for every extra outlier added. This is exactly the same for both the Weight and Height variables.

A similar pattern can be seen with the Type I error rate. The average proportion of Type I errors caused by the outliers quickly rose and the proportion surpassed the 75%-threshold for both tests for both variables when the 5th outlier was added. The average proportion seems to fluctuate around 90% for the iterations when more than 5 outliers were added.

The median permutation test could be considered the most robust to a single outlier in different scenarios. This conclusion can also be drawn when extra outliers are added. The average p-value slightly shrinks, but this effect is smaller than the effect of the other tests. The lowest average p-value for the median permutation test is 0.283 for Height and 0.336 for Weight. The average p-values stay on level for the first 20 outliers added. The proportions of the Type I errors rose when the number of outliers increased, but the rate at which the proportions rose is significantly lower than those of the other tests. The median permutation test outperforming the other tests could be explained by the lower sensitivity

of medians to outliers. Both the Student's t-test and mean permutation test use means to compute the test statistics and p-values. And means are more sensitive to extreme values than medians (Moore et al., 2016). The slight shrinkage of the p-value could be caused by the outliers being added making group 1 grow from 1000 observations to 1050 observations, so the median of group 1 for the initial iteration and the final iteration could differ. The growth of the number of observations could cause a difference in the medians. Additionally, since the added observations are all outliers, slowly increases the medians of the groups.

Contrary to the simulations introducing a single outlier, multiple outliers do cause Type I errors for the Student's t-test. A possible explanation could be that a single outlier causes a significant shift but the initial p-values were relatively high. Only after multiple outliers are added, do the outliers cause the p-value to cross the significance threshold.

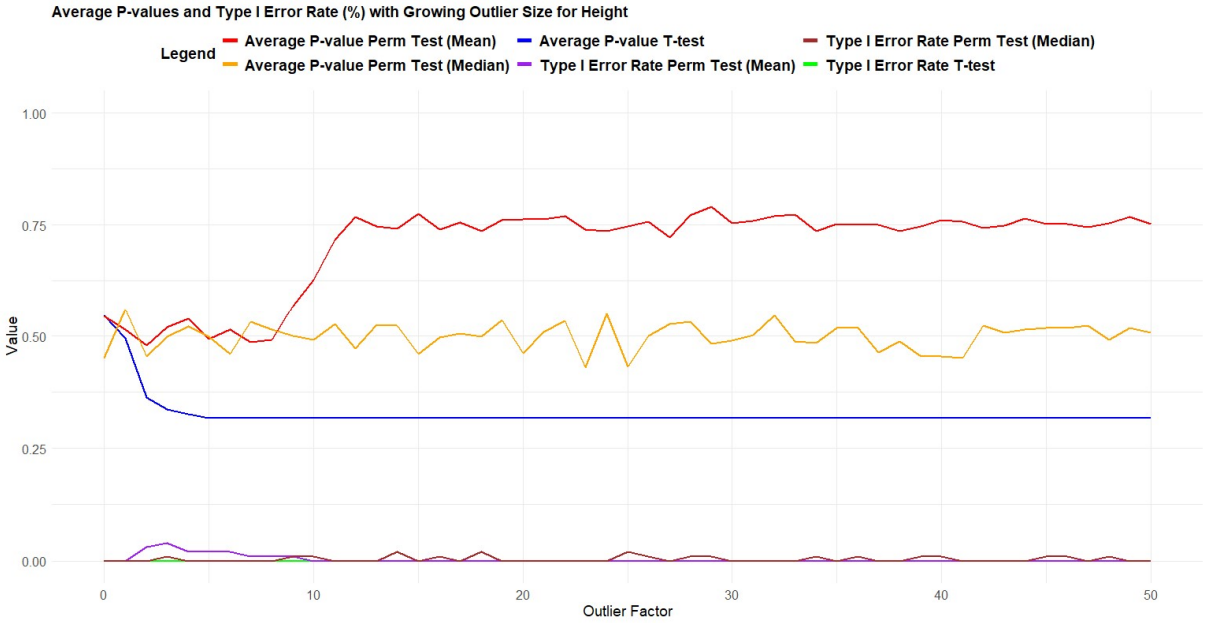
So, if multiple outliers are or could be present in the data a permutation test measuring the differences between medians is the most robust. Both the Student's t-test and mean permutation test perform worse when multiple outliers are present. Both perform almost the same for both the average p-values and Type I errors.

5.3. Single Comparison Robustness with Growing Outlier

Along with test the effect of an increasing number of outliers, this study aims to measure the effect of an outlier that increases in size on the outcome of the Student's t-test and permutation tests. The initial size of the outlier was the maximum value of the original Height and Variables. The next iteration the outlier was doubled in size. This doubling process was repeated 50 times. As a result, the outliers' final sizes were $8e^{16}$ inches and $2e^{17}$ pounds. The outliers had extreme sizes but were less than 0.01% of the observations in group 1.

Figure 6

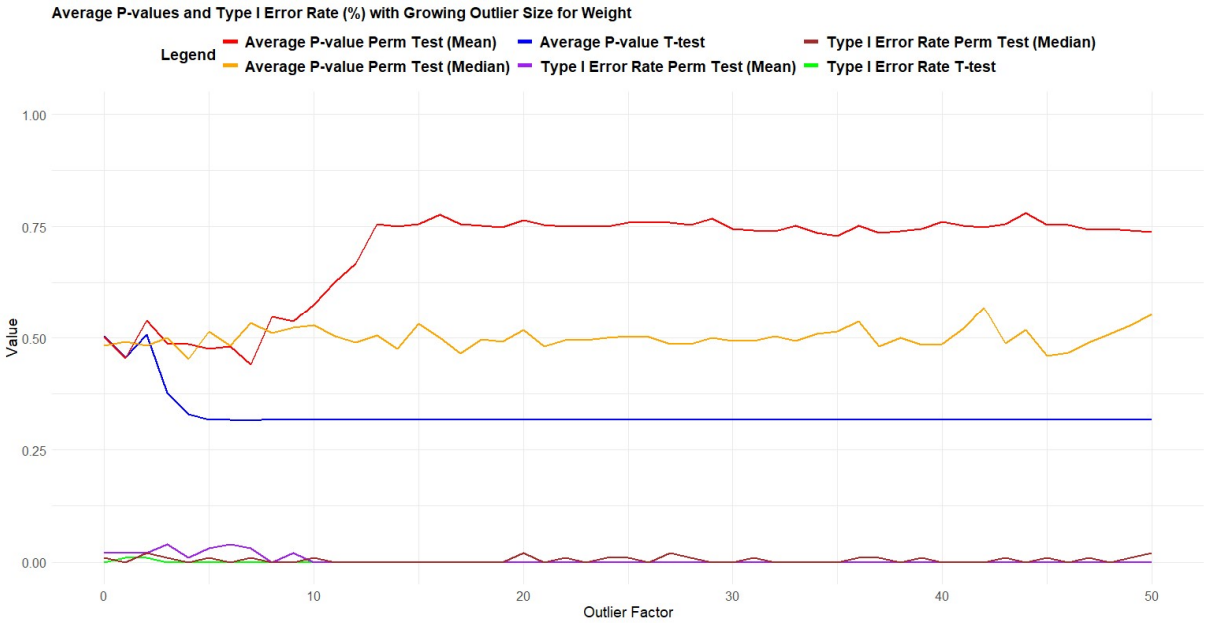
Overview of results of simulation with growing outlier sizes for Height, including average p-values and average Type I error rate (%) caused by the addition of the outlier for the Student's t-test, mean permutation test and median permutation test



Note: The Outlier Factor value states the factor of *i* in the following formula $Outlier\ value = initial\ outlier * 2^i$. The Type I error rate states the proportion of the iterations where the null hypothesis was rejected after adding the outlier(s) when the null hypothesis was accepted before adding the outlier(s).

Figure 7

Overview of results of simulation with growing outlier sizes for Weight, including average p-values and average Type I error rate (%) caused by the addition of the outlier for the Student's t-test, mean permutation test and median permutation test



Note: The Outlier Factor value states the factor of *i* in the following formula $Outlier\ value = initial\ outlier * 2^i$. The Type I error rate states the proportion of the iterations where the null hypothesis was rejected after adding the outlier(s) when the null hypothesis was accepted before adding the outlier(s).

As with the increasing number of outliers, both simulations for Height and Weight show a similar pattern. The first difference with the increasing number of outliers is the fact that the growing outliers

do not cause a high Type I error rate. The Type I error rate stays about the same for the entire simulation, never exceeding 5%. So, a massive outlier has little effect on the validity of the test concerning the Type I error rate.

Where the average p-values of the Student's t-test and mean permutation tests for the simulations introducing multiple outliers followed a similar pattern, the tests show a different pattern for the growing simulations. The average p-value of the t-test shrinks when the size of the outlier grows. For both variables the p-value of the t-tests shrinks and levels off at 0.318 after the outlier was doubled for the 9^{th} time. This could be explained by the outlier inflating the estimated standard deviation of group 1. This causes the *S* or the denominator of the Student's t-test formula to rise significantly. Therefore, the p-value of the t-test shrinks. The levelling off at 0.318 could be a result of the estimated standard deviation and the mean of group 1 both inflating and finally reaching a point where a doubled outlier does not change the values.

The p-value of the mean permutation test increased and settled around 0.750 after 15 iterations. A possible explanation for the fact that the p-values did not rise anymore could be that the outliers are already very extreme and multiplying the outlier more does not cause the mean to shift further towards the outlier.

The median permutation tests seemed to perform the same for the different sizes of the outlier. Contrary to the increasing number of outliers, almost no increase or decrease pattern can be seen. The median permutation test results in an average p-value of around 0.5 for all iterations. This could be explained by the median not being sensitive to extreme values. The simulation adding an increasing number of outliers saw a slight decrease in the average p-values. Since the growing outlier simulation only increases the number of observations by a small margin, the median is less likely to change before and after adding the outlier.

The median permutation test does, however, result in a few Type I errors. The other tests do not result in any Type I errors after they have settled around the observed levels of 0.318 and 0.750 for the t-test and mean permutation test, respectively. So, if the p-values of the tests after the outliers shift towards insignificant levels, fewer or no cases of the test outcomes changing from insignificant to significant after adding the outlier will occur. The Type I error rate decreasing could also suggest that the Type II error rate increases. Since the p-values level off at insignificant levels, the outlier could cause the tests to result in the null hypothesis being accepted when it should have been rejected. However, the Type II error rate was not measured in the simulation. A suggestion for further research could be to research the effect of an increasing outlier on the Type II error rate.

In conclusion, the median permutation performs best when an outlier increasing in size is added. This type of test is not affected by the outlier concerning the average p-value or Type I error rate. The average p-value of the mean permutation test is inflated while the Student's t-test experiences a shrinking p-value. The inflating or shrinking effects level off when the outlier is 512 times its original size. The growing outlier does decrease the Type I error rate but could cause an increase in Type II errors.

5.4. Multiple Comparison Robustness

As well as comparing classical methods and permutation methods for single comparison tests, this study also set a goal to compare the classical and permutation methods for multiple comparison methods. For multiple tests two methods were compared, the Bonferroni correction and the maxT method. Even though the p-values could differ before and after adding the outlier, this is not the most important measurement as the Type II error or change in the number of genes that were found to be significant is a better measure to test robustness.

Since the simulation uses the entire Golub dataset to perform a t-test for all 7,129 gene expressions, a single baseline was established. The result from the initial t-test with a significance level adjusted by the Bonferroni method stated that 143 genes were significant. This means that 143 genes, almost 2% of all gene probes, had a significant difference between the AML and ALL groups. The p-values adjusted by the maxT method used on the outcome of the same t-test resulted in 167, 2.3%, having a significant difference between the ALL and AML samples. This already showed why Bonferroni is considered to be more conservative. In a similar situation, the Bonferroni correction results in fewer gene probes being significant than the maxT method.

Table 7

Overview of results of single outlier robustness simulations, including the number of significant genes before (Baseline) and after outlier addition and Type II error rate

Group	*Significant genes Bonferroni Baseline*	*Significant genes maxT Baseline*	*Significant genes Bonferroni*	*Significant genes maxT*	*Type II error rate Bonferroni*	*Type II error rate maxT*
ALL	143	167	24	59	83.2%	64.7%
AML	143	167	89	141	37.8%	15.6%

Note: The Type II error rate indicates the proportion of differences in the number of significant genes compared to the baseline. These Type II errors were false negatives caused by the introduction of the outlier.

Even though the ALL and AML simulations show different results, they show a similar pattern. The Bonferroni method is most influenced by the addition of an outlier. The Type II error rate of Bonferroni is higher in both simulations. This indicates that the Bonferroni correction is less robust than the maxT method. A negative change in significance could be considered a Type II error as some null hypotheses are accepted when they should be rejected.

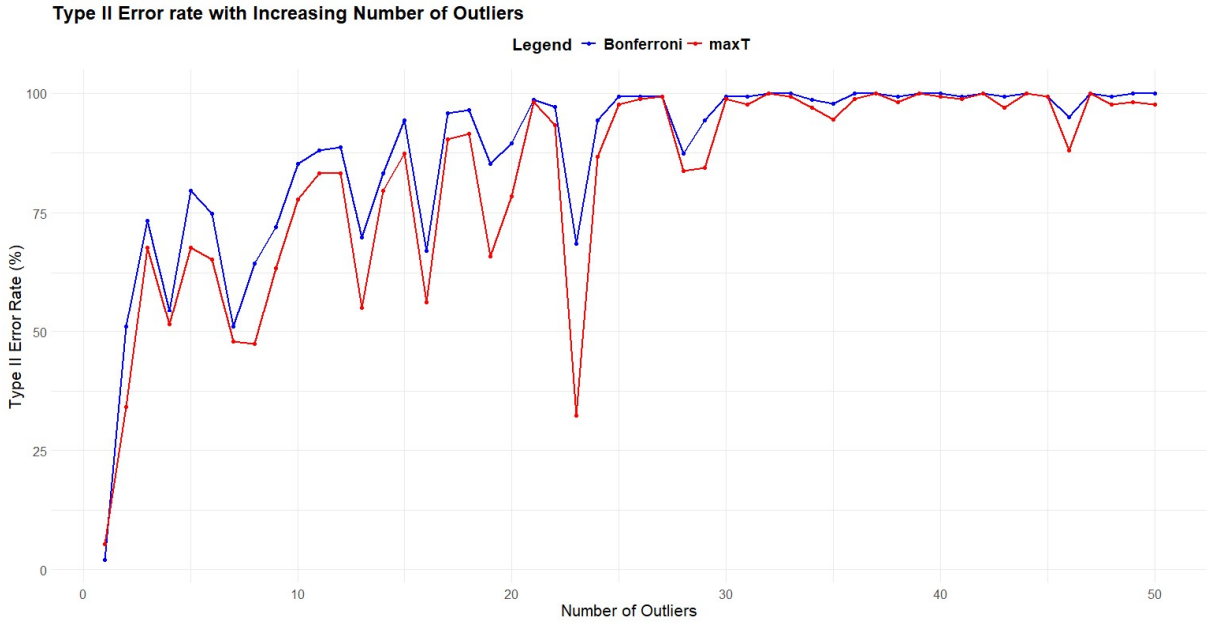
Concluding, the maxT method is more robust to outliers than the Bonferroni method.

5.5. Multiple Comparison Robustness with Increasing Number of Outliers

These simulations aimed to measure the effect of an increasing number of outliers on the outcome of the Bonferroni correction and maxT method. The baseline derived in 5.6, also applied to the simulations that tested the effect of an increasing number of outliers. So, without any outliers, Bonferroni stated that 143 gene probes significantly differed between the cancer types. 167 gene expressions saw a significant difference between the ALL and AML types of cancer according to the maxT method.

Figure 8

Overview of Type II error rate with increasing number of outliers



The first conclusion that can be drawn from Figure 8 is that both methods lack the power to be robust to a large number of outliers. However, this can be explained by the relatively small number of original observations. The data originally contained 72 observations, so every outlier adds about 1,5% to the original observations. The Type II error rate slowly rose after each extra outlier was added. After the 30th outlier was added the Type II error rate levels off around 99-100%. This can be explained by the outliers making up almost 30% of the total number of observations evaluated by the methods. So, both methods are affected by multiple outliers that are present in the data.

To compare the robustness of the Bonferroni and maxT methods, the Type II error rates should be compared. The method with the lowest percentages could be considered the most robust. After 50 iterations, the error rate was equal for both methods 6 times all of these 6 times the Type II error rate for both methods was 100%. These 6 times all occurred after 30 or more outliers were introduced. The Bonferroni correction performed better only 3 times or 6% of the total iterations. The maxT method

performed better in 41 iterations 82% of the time. On average, the Type II error rate of the maxT method is 5 percentage points higher than that of the Bonferroni correction.

A Wilcoxon test was conducted to test whether the difference in significant features between Bonferroni and maxT was significant. Additionally, the differences in Type II error rates for the different iterations were also tested using the Wilcoxon test. Both tests stated that the differences were significantly different.

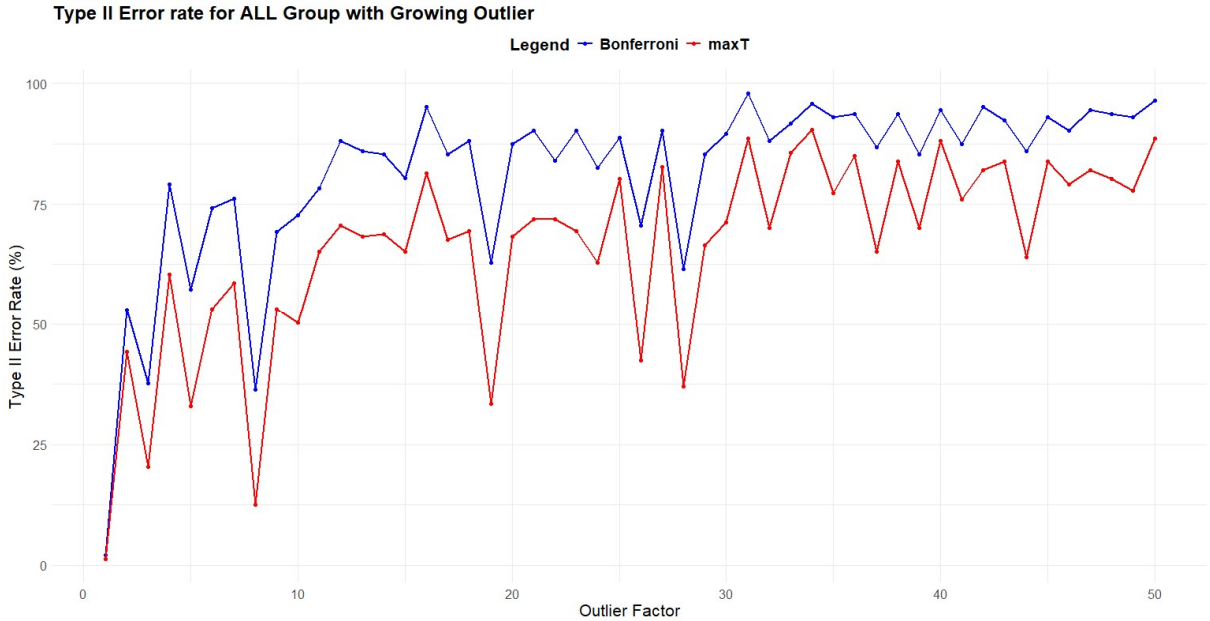
In conclusion, when multiple outliers are introduced to the data the maxT method is significantly more robust than the Bonferroni correction. However, the Type II error rate tends to rise towards 100% if more outliers are present in the data. This is true for both methods.

5.6. Multiple Comparison Robustness with Growing Outlier

Finally, the robustness of multiple comparisons against a single progressively growing outlier is assessed. This simulation was conducted twice. The outlier was once introduced to the ALL group and once was the label AML given. Again the baselines to which the number of significant genes from the simulations will be compared are 143 for the Bonferroni correction and 167 significant genes for the maxT result.

Figure 9

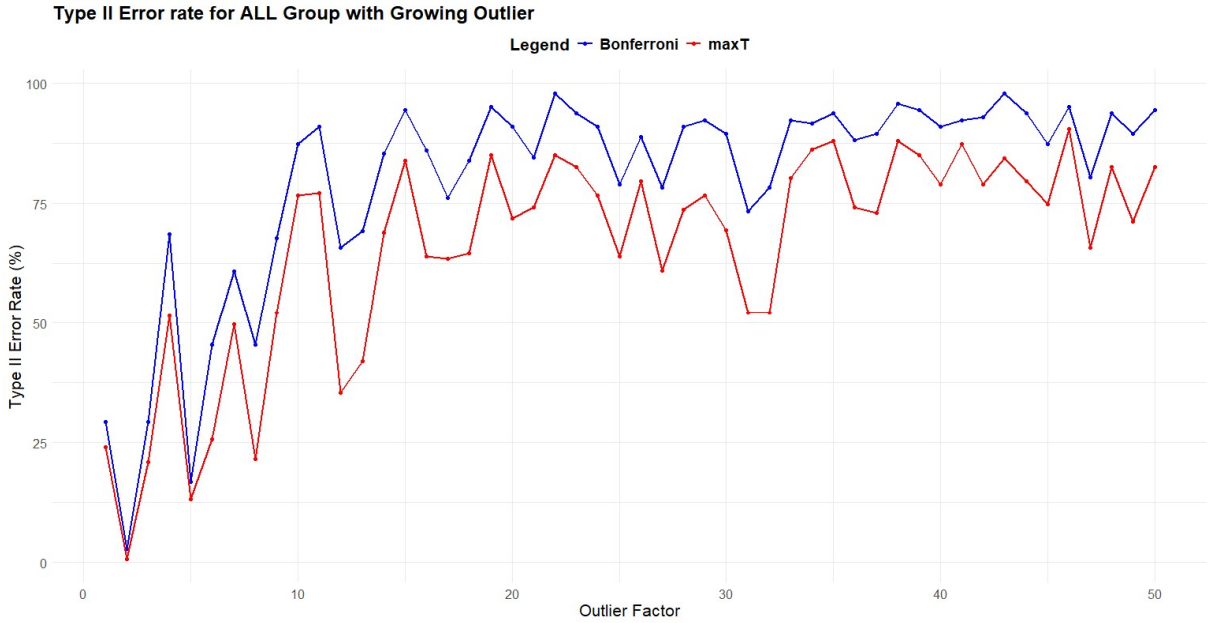
Overview of Type II Error Rates for AML Group with Growing Outlier



Note: The Outlier Factor value states the factor with which the gene expressions of the outlier were multiplied. The Type II error rate indicates the proportion of differences in the number of significant genes compared to the baseline. These Type II errors were false negatives caused by the introduction of the outlier.

Figure 10

Overview of Type II Error Rates for ALL Group with Growing Outlier



Note: The Outlier Factor value states the factor with which the outlier was multiplied. The Type II error rate indicates the proportion of differences in the number of significant genes compared to the baseline. These Type II errors were false negatives caused by the introduction of the outlier.

Figures 9 and 10 show a similar pattern for the Type II error rates for the different outlier factors for both simulations. The growing outlier caused the Type II error rate to rise. Figures 9 and 10 show that the Bonferroni correction had a higher Type II error rate than maxT for every iteration in both simulations. This suggests that the maxT method is more robust to outliers than the Bonferroni correction, even if the outliers are bigger. The maxT method had a Type II error rate that was approximately 14.6 percentage points lower than that of the Bonferroni correction on average. There was a slight difference between the simulations where the outlier was given the label ALL or AML. These results resulted in a Type II error rate of, respectively, 14.0 and 15.1 percentage point difference on average.

Thus, the maxT method is more robust to a bigger outlier than the Bonferroni correction. However, if the outlier gets more extreme values, the Type II error rate rises for both methods.

6. Conclusion and Discussion

This study researched the robustness of permutation tests and the maxT method compared to classical statistical tests, especially the Student's t-test and Bonferroni correction when (extreme) statistical outliers are included in the analysed data. This research directly compares the robustness of these methods against outliers, a comparison not yet made in the existing literature.

The literature review concluded that the permutation tests could be more robust because of their versatility and possibility to adapt to the data. For the single comparison tests, the Student's t-test is limited by the dependence on assumptions and the use of the mean to compute the test statistic. The permutation test that measured the differences between means is more robust than the Student's t-test. However, the permutation test also can be based on the differences between the medians of groups. Medians are less sensitive to extreme values and can thus be more robust. The empirical analysis of this study using weight and height data confirmed these findings. The median permutation test is more robust to outliers than the classical Student's t-test and the mean permutation test. Even in situations where the assumptions of Student's t-test hold before the outlier was introduced the permutation tests were more robust. However, the mean permutation test resulted in the highest Type I error rate after the outlier was added. The Type I error rate of the median permutation test was relatively lower compared to the mean permutation test. Even though the p-value of the Student's t-test was significantly altered by the addition of an outlier, this did not result in an increased Type I error rate.

The empirical analysis also showed that the median permutation test is most robust when multiple outliers are included. The p-values of both the Student's t-test and mean permutation test were eventually shrunken to zero after multiple outliers were added. The Type I error rates of these tests were inflated to 90-100% after the addition of multiple outliers. The p-value and Type I error rate of the median permutation were influenced less by the outliers. So, if multiple outliers are included in the data the use of the median permutation test is recommended as this test would provide the most robust outcome.

The same was concluded for a situation where a bigger outlier was introduced to the data. The outlier influenced the p-values and Type I error rates of the Student's t-test and mean permutation test significantly. However, the growing outlier had a small effect on the Type I error rate, shrinking the rate to zero for the t-test and mean permutation test. This was caused by the p-values of these tests levelling off respectively 0.318 and 0.700. The median permutation test showed almost no effect of the growing outlier. The p-value and Type 1 error rate remained about the same for the different sizes of the outlier.

So, if a single extreme outlier is included in the data the outcome of the median permutation remains valid while the other tests were significantly affected by an extreme outlier.

For multiple comparison methods, the literature indicated that the versatility and data adoption of the maxT method potentially could provide the method more robustness compared to the Bonferroni method. The Bonferroni method could mostly be restricted by its conservativeness. The empirical simulations which analysed the Golub data confirmed the literature findings showing that the maxT had a lower Type II error rate than the Bonferroni correction.

In scenarios where multiple outliers are or a relatively larger outlier is present in the data, the results of the maxT method showed a significantly lower Type II error rate than the results of the Bonferroni correction. However, when the number of outliers increases or an outlier grows in size, the Type II error rate increases towards 100%. This is true for both the maxT and Bonferroni methods.

In conclusion, the permutation-based methods, permutation tests and maxT method, are more robust to outliers than the classical statistic methods, Student's t-test and Bonferroni correction. So, if researchers analyse data or conduct a statistical study and use data that could be outlier-infested, they are advised to use permutation-based methods as they provide more robustness to these outliers.

These findings can be useful for marketers or marketing researchers who have to select a method for their group comparison analyses or studies. Selecting permutation-based methods over classical methods could provide more robust and valid results.

6.1. Discussion

While this study provides insights into the robustness of permutation tests compared to classical statistical methods when outliers are present in the data, recognizing its limitations and potential areas for improvement is an important step.

This study used real-world data but introduced simulated outliers. These outliers could not represent realistic outliers that could be encountered in real data analyses. Additionally, the Golub dataset that was used to compare the multiple comparison methods saw some correlations and dependence between different gene probes. Even though these only applied to a marginal part of the data, these could be considered a violation of the assumptions of the Bonferroni correction.

This study focused on the Type I error of the single comparison methods when an increasing number of outliers were added. The Type I error was shrunken to zero after multiple outliers were introduced. This could have affected the Type II error rate, but this was not measured. So, if researchers use data that could be cluttered with outliers, they should be aware that the Type II error rate could be higher than desired. This study did not measure the statistical power of the single comparison tests but instead

focused on the shift in p-values. Future research could study the influence of outliers on the power of the single comparison methods.

While this study used simulations to test the robustness of multiple methods. These basic simulations introducing a single outlier consisted of 200 iterations. The simulations which added multiple outliers or a growing outlier consisted of 50 times 100 iterations. These simulations used relatively small numbers of iterations due to the computationally intensive methods being used. Due to the long time these simulations took to run, the number of iterations was not increased.

Future studies could focus on other comparisons between the methods analysed or methods not included in this study. Additionally, researching the impact of different types of outliers and distributions on the performance of these tests could be used as a topic of further research. The outliers in this study were mostly bigger versions of existing observations. Researchers could explore the impact of smaller outliers on the outcome of tests.

7. Appendix

7.1. References

Almeida, A., Loy, A., & Hofmann, H. (2018). ggplot2 Compatible Quantile-Quantile Plots in R. The R Journal, 10–2, 248–250. https://svn.r-project.org/Rjournal/trunk/html/_site/archive/2018/RJ-2018-051/RJ-2018-051.pdf

Anderson, M. J., & Robinson, J. (2001). Permutation Tests for Linear Models. Australian & New Zealand Journal of Statistics, 43(1), 75–88. https://doi.org/10.1111/1467-842x.00156

Armstrong, R. A. (2014). When to use the Bonferroni correction. Ophthalmic and Physiological Optics/Ophthalmic & Physiological Optics, 34(5), 502–508. https://doi.org/10.1111/opo.12131

Bååth, R., & Dobbyn, A. (2024). beepr. CRAN. https://cran.r-project.org/web/packages/beepr/beepr.pdf

Baidun, A., Prananta, R., Harahap, M. a. K., & Yusuf, M. (2022). Effect Of Customer Satisfaction, Marketing Mix, And Price In Astana Anyar Market Bandung. Al-Kharaj Journal of Islamic Economic and Business, 4(2). https://doi.org/10.24256/kharaj.v4i2.3583

Barbato, G., Barini, E. M., Genta, G., & Levi, R. (2011). Features and performance of some outlier detection methods. Journal of Applied Statistics, 38(10), 2133–2149. https://doi.org/10.1080/02664763.2010.545119

Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson Correlation Coefficient. In Springer topics in signal processing (pp. 1–4). https://doi.org/10.1007/978-3-642-00296-0_5

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B. Methodological, 57(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Berry, K. J., Johnston, J. E., & Mielke, P. W. (2014). A Chronicle of Permutation Statistical Methods. In Springer eBooks. https://doi.org/10.1007/978-3-319-02744-9

Bland, J. M., & Altman, D. G. (1995). Statistics notes: Multiple significance tests: the Bonferroni method. BMJ, 310(6973), 170. https://doi.org/10.1136/bmj.310.6973.170

Blázquez-García, A., Conde, Á., Mori, U., & Lozano, J. A. (2021). A Review on Outlier/Anomaly Detection in Time Series Data. ACM Computing Surveys, 54(3), 1–33. https://doi.org/10.1145/3444690

Bolstad, B. M., Irizarry, R. A., Astrand, M., & Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics, 19(2), 185–193. https://doi.org/10.1093/bioinformatics/19.2.185

Boytsov, L., Belova, A., & Westfall, P. (2013). Deciding on an adjustment for multiplicity in IR experiments. Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '13). Association for Computing Machinery. https://doi.org/10.1145/2484028.2484034

Brown, M. B., & Forsythe, A. B. (1974). Robust Tests for the Equality of Variances. Journal of the American Statistical Association, 69(346), 364–367. https://doi.org/10.1080/01621459.1974.10482955

Bryc, W. (1995). The Normal Distribution: Characterizations with Applications. http://ci.nii.ac.jp/ncid/BA25001379

Burk, S. (2006). A Better Statistical Method for A/B Testing in Marketing Campaigns [Technical Note]. Marketing Bulletin, 17. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=ab09d3493c8943dda5a1a79304bad79c24b12d9c

Cade, B. S., & Richards, J. D. (1996). Permutation tests for least absolute deviation regression. Biometrics, 52(3), 886. https://doi.org/10.2307/2533050

Chapman, A. R., Lee, D. F., Cai, W., Ma, W., Li, X., Sun, W., & Xie, X. S. (2022). Correlated gene modules uncovered by high-precision single-cell transcriptomics. Proceedings of the National Academy of Sciences of the United States of America, 119(51). https://doi.org/10.1073/pnas.2206938119

Cheverud, J. M. (2001). A simple correction for multiple comparisons in interval mapping genome scans. Heredity, 87(1), 52–58. https://doi.org/10.1046/j.1365-2540.2001.00901.x

Collingridge, D. S. (2012). A Primer on Quantitized Data Analysis and Permutation Testing. Journal of Mixed Methods Research, 7(1), 81–97. https://doi.org/10.1177/1558689812454457

Dahl, D. B., Scott, D., Roosen, C., & Swinton, J. (2000). xtable: Export tables to LaTeX or HTML. ResearchGate. https://www.researchgate.net/publication/275646318_xtable_Export_tables_to_LaTeX_or_HTML

Das, K. R. (2016). A brief review of tests for normality. American Journal of Theoretical and Applied Statistics, 5(1), 5. https://doi.org/10.11648/j.ajtas.20160501.12

Derrick, B., Toher, D., White, P., & University of the West of England, Bristol, England. (2016). Why Welch's test is Type I error robust. Quantitative Methods for Psychology. https://www.tqmp.org/RegularArticles/vol12-1/p030/p030.pdf
Dixon, W. J. (1953). Processing data for outliers. Biometrics, 9(1), 74. https://doi.org/10.2307/3001634
Dudoit, S., Shaffer, J. P., & Boldrick, J. C. (2003). Multiple Hypothesis Testing in Microarray Experiments. Statistical Science, 18(1). https://doi.org/10.1214/ss/1056397487
Dunn, O. J. (1961). Multiple Comparisons among Means. Journal of the American Statistical Association, 56(293), 52–64. https://doi.org/10.1080/01621459.1961.10482090
Edgington, E., & Onghena, P. (2007). Randomization tests. CRC Press.
Efron, B., & Tibshirani, R. (1994). An Introduction to the Bootstrap. In Chapman and Hall/CRC eBooks. https://doi.org/10.1201/9780429246593
Ernst, M. D. (2004a). Permutation methods: a basis for exact inference. Statistical Science, 19(4). https://doi.org/10.1214/088342304000000396
Ernst, M. D. (2004b). Permutation Methods: A Basis for Exact Inference. Statistical Science, 19(4). https://doi.org/10.1214/088342304000000396
Eusebio, R., Andreu, J. L., & Belbeze, M. P. L. (2006). Measures of marketing performance: a comparative study from Spain. International Journal of Contemporary Hospitality Management, 18(2), 145–155. https://doi.org/10.1108/09596110610646691
Field, A. (2017). Discovering statistics using IBM SPSS Statistics: North American Edition. SAGE.
Fox, J., & Weisberg, S. (2020). An R companion to applied regression (Third edition). SAGE Publications.
Gibbons, J. D., & Chakraborti, S. (2010). Nonparametric Statistical Inference. In Chapman and Hall/CRC eBooks. https://doi.org/10.1201/9781439896129
Goeman, J. J., & Solari, A. (2014). Multiple hypothesis testing in genomics. Statistics in Medicine, 33(11), 1946–1978. https://doi.org/10.1002/sim.6082
Golub, T. (2024). exprSets for golub leukemia data [ExperimentData]. https://bioconductor.org/packages/release/data/experiment/manuals/golubEsets/man/golubEsets.pdf
Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., & Lander, E. S. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science, 286(5439), 531–537. https://doi.org/10.1126/science.286.5439.531
Good, P. (2000). Permutation tests. In Springer series in statistics. https://doi.org/10.1007/978-1-4757-3235-1
Good, P. (2005). Permutation, parametric and bootstrap tests of hypotheses. In Springer eBooks. https://doi.org/10.1007/b138696
Good, P. (2013). Permutation tests: A Practical Guide to Resampling Methods for Testing Hypotheses. Springer Science & Business Media.
Gross, J., Ligges, U., & Uwe Ligges. (2015). Tests for Normality [Book]. CRAN. https://cran.r-project.org/web/packages/nortest/nortest.pdf
Gumpinger, A. C., Rieck, B., Grimm, D. G., & Borgwardt, K. (2020). Network-guided search for genetic heterogeneity between gene pairs. Bioinformatics, 37(1), 57–65. https://doi.org/10.1093/bioinformatics/btaa581
Haenlein, M., & Kaplan, A. M. (2011). The Influence of Observed Heterogeneity on Path Coefficient Significance: Technology Acceptance Within the Marketing Discipline. The Journal of Marketing Theory and Practice, 19(2), 153–168. https://doi.org/10.2753/mtp1069-6679190203
Harvey, C. R., Liu, Y., & Saretto, A. (2020). An Evaluation of Alternative Multiple Testing Methods for Finance Applications. The Review of Asset Pricing Studies, 10(2), 199–248. https://doi.org/10.1093/rapstu/raaa003
Hawkins, D. M. (1980). Identification of outliers. In Springer eBooks. https://doi.org/10.1007/978-94-015-3994-4
Hemerik, J., & Goeman, J. J. (2017a). False Discovery Proportion Estimation by Permutations: Confidence for Significance Analysis of Microarrays. Journal of the Royal Statistical Society Series B: Statistical Methodology, 80(1), 137–155. https://doi.org/10.1111/rssb.12238
Hemerik, J., & Goeman, J. J. (2017b). Exact testing with random permutations. TEST, 27(4), 811–825. https://doi.org/10.1007/s11749-017-0571-1
Hochberg, Y., & Tamhane, A. C. (1987). Multiple Comparison Procedures. In Wiley series in probability and statistics. https://doi.org/10.1002/9780470316672
Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. Scandinavian Journal of Statistics, 6, 65–70. https://doi.org/10.2307/4615733

Hsing, T., Attoor, S., & Dougherty, E. (2003). Relation between Permutation-Test P values and Classifier error estimates. Machine Learning, 52, 11–30. https://doi.org/10.1023/A:1023985022691
Jafari, M., & Ansari-Pour, N. (2019). Why, When and How to Adjust Your P Values? DOAJ (DOAJ: Directory of Open Access Journals), 20(4), 604–607. https://doi.org/10.22074/cellj.2019.5992
John, M., Ankenbrand, M. J., Artmann, C., Freudenthal, J. A., Korte, A., & Grimm, D. G. (2022). Efficient permutation-based genome-wide association studies for normal and skewed phenotypic distributions. Bioinformatics, 38(Supplement_2), ii5–ii12. https://doi.org/10.1093/bioinformatics/btac455
Kassambara, A. (2023). Package 'ggpubr.' https://cran.r-project.org/web/packages/ggpubr/ggpubr.pdf
Keller-Mcnulty, A., & Higgins, J. J. (1987). Effect of tail weight and outliers on power and type-i error of robust permutation tests for location. Communications in Statistics: Simulation and Computation, 16(1), 17–35. https://doi.org/10.1080/03610918708812575
Keren, G. (1993). A handbook for data analysis in the Behavioral Sciences: Methodological Issues. Lawrence Erlbaum Assoc Incorporated.
Keren, G., & Lewis, C. (1993). A Handbook for Data Analysis in the Behaviorial Sciences. In Taylor & Francis (Vols. 1–2). Lawrence Erlbaum Associates, Inc., Publishers. https://doi.org/10.4324/9781315799582
Kisoentewari, A. (2022). Evaluating the robustness of permutation-based multiple testing methods [Master Thesis, Universiteit Leiden]. In Student Repository. https://studenttheses.universiteitleiden.nl/handle/1887/3485866?solr_nav%5Bid%5D=950b1455c577a6d06591&solr_nav%5Bpage%5D=4&solr_nav%5Boffset%5D=1
Kolde, R. (2019). Pretty Heatmaps. https://cran.r-project.org/web/packages/pheatmap/pheatmap.pdf
Kwak, S. G., & Kim, J. H. (2017). Central limit theorem: the cornerstone of modern statistics. Korean Journal of Anesthesiology, 70(2), 144. https://doi.org/10.4097/kjae.2017.70.2.144
Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics, 9(1). https://doi.org/10.1186/1471-2105-9-559
Lee, D. K. (2020). Data transformation: a focus on the interpretation. Korean Journal of Anesthesiology, 73(6), 503–508. https://doi.org/10.4097/kja.20137
Lehmann, R., & Lösler, M. (2016). Multiple Outlier Detection: Hypothesis Tests versus Model Selection by Information Criteria. Journal of Surveying Engineering, 142(4). https://doi.org/10.1061/(asce)su.1943-5428.0000189
Levene, H. (1961). Robust tests for equality of variances. Contributions to Probability and Statistics, 279–292. https://ci.nii.ac.jp/naid/10007628681
Lin, Y. (2015, October 19). The multiple comparison problem in GWAS: Bonferroni correction, FDR control, and permutation testing. https://lybird300.github.io/2015/10/19/multiple-test-correction.html
List, J. A., Shaikh, A. M., & Xu, Y. (2019). Multiple hypothesis testing in experimental economics. Experimental Economics, 22(4), 773–793. https://doi.org/10.1007/s10683-018-09597-5
Livingston, E. H. (2004). Who was student and why do we care so much about his t-test? The Journal of Surgical Research, 118(1), 58–65. https://doi.org/10.1016/j.jss.2004.02.003
López-Siguero, J. P., García, J. M. F., De Dios Luna Castillo, J., Molina, J. a. M., Cosano, C. R., & Ortiz, A. J. (2008). Cross-sectional study of height and weight in the population of Andalusia from age 3 to adulthood. BMC Endocrine Disorders, 8(S1). https://doi.org/10.1186/1472-6823-8-s1-s1
Ludbrook, J., & Dudley, H. (1998). Why Permutation Tests Are Superior to t and F Tests in Biomedical Research. the American Statistician, 52(2), 127. https://doi.org/10.2307/2685470
Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The Importance of the Normality Assumption in Large Public Health Data Sets. Annual Review of Public Health, 23(1), 151–169. https://doi.org/10.1146/annurev.publhealth.23.100901.140546
McHugh, M. L. (2013). The Chi-square test of independence. Biochemia Medica, 143–149. https://doi.org/10.11613/bm.2013.018
Menyhart, O., Weltz, B., & Győrffy, B. (2021). MultipleTesting.com: A tool for life science researchers for multiple hypothesis testing correction. PloS One, 16(6), e0245824. https://doi.org/10.1371/journal.pone.0245824
Michael, J. R. (1983). The stabilized probability plot. Biometrika, 70(1), 11–17. https://doi.org/10.1093/biomet/70.1.11
Mielke, P. W., & Berry, K. J. (1994). Permutation Tests for Common Locations Among Samples With Unequal Variances. Journal of Educational Statistics, 19(3), 217–236. https://doi.org/10.3102/10769986019003217
Mielke, P. W. J., & Berry, K. J. (2013). Permutation methods: A Distance Function Approach. Springer Science & Business Media.

Moore, D., McCabe, G., Craig, B., & Alwan, L. (2016). The practice of statistics for business and economics.
mt.maxT function - RDocumentation. (n.d.). https://www.rdocumentation.org/packages/multtest/versions/2.28.0/topics/mt.maxT
Nakagawa, S. (2004). A farewell to Bonferroni: the problems of low statistical power and publication bias. Behavioral Ecology, 15(6), 1044–1045. https://doi.org/10.1093/beheco/arh107
Nichols, T., & Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: a comparative review. Statistical Methods in Medical Research, 12(5), 419–446. https://doi.org/10.1191/0962280203sm341ra
Noble, W. S. (2009). How does multiple testing correction work? Nature Biotechnology, 27(12), 1135–1137. https://doi.org/10.1038/nbt1209-1135
Noguchi, K., Abel, R. S., Marmolejo-Ramos, F., & Konietschke, F. (2019). Nonparametric multiple comparisons. Behavior Research Methods, 52(2), 489–502. https://doi.org/10.3758/s13428-019-01247-9
Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should ALWAYS check for them). Practical Assessment, Research and Evaluation, 9(1), 1–8. https://doi.org/10.7275/qf69-7k43
Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. BMJ. British Medical Journal, 316(7139), 1236–1238. https://doi.org/10.1136/bmj.316.7139.1236
Pesarin, F., & Salmaso, L. (2010). Permutation tests for complex data: Theory, Applications and Software. John Wiley & Sons.
Phipson, B., & Smyth, G. K. (2010). Permutation P-values Should Never Be Zero: Calculating Exact P-values When Permutations Are Randomly Drawn. Statistical Applications in Genetics and Molecular Biology, 9(1). https://doi.org/10.2202/1544-6115.1585
Pollard, K. S., Dudoit, S., & Van Der Laan, M. J. (2005). Multiple Testing Procedures: the multtest Package and Applications to Genomics. In Statistics in the health sciences (pp. 249–271). https://doi.org/10.1007/0-387-29362-0_15
R Core Team. (2024). R: A language and environment for Statistical computing. R Foundation for Statistical Computing, Vienna, Austria. R-project. https://www.R-project.org/
Rajabi, M., Faridrohani, MR., Department of Statistics, Science and Research Branch, Islamic Azad University, Tehran, Iran, & Department of Statistics, Shahid Beheshti University, Tehran, Iran. (2019). The Comparison of Two Multiple Testing Methods for Outliers Detection in Nonparametric Profile Monitoring. Int. J. Industrial Mathematics, 11(3), 9. https://journals.srbiau.ac.ir/article_14669_095a436597bd223706453affd5a9dc9b.pdf
Rempala, G. A., & Yang, Y. (2013). On permutation procedures for strong control in multiple testing with gene expression data. Statistics and Its Interface, 6(1), 79–89. https://doi.org/10.4310/sii.2013.v6.n1.a8
Ringland, J. T. (1983). Robust Multiple Comparisons. Journal of the American Statistical Association, 78(381), 145–151. https://doi.org/10.1080/01621459.1983.10477943
Ripley, B. D. (2009). Stochastic simulation. John Wiley & Sons.
Rohatgi, V. K., Kariya, T., & Sinha, B. K. (1991). Robustness of statistical tests. Technometrics, 33(2), 246. https://doi.org/10.2307/1269062
Romano, J. P., & Wolf, M. (2005a). Stepwise Multiple Testing as Formalized Data Snooping. Econometrica, 73(4), 1237–1282. https://doi.org/10.1111/j.1468-0262.2005.00615.x
Romano, J. P., & Wolf, M. (2005b). Stepwise Multiple Testing as Formalized Data Snooping. Econometrica, 73(4), 1237–1282. https://doi.org/10.1111/j.1468-0262.2005.00615.x
Routledge, R. (1997). P-values from permutation and F-tests. Computational Statistics & Data Analysis, 24(4), 379–386. https://doi.org/10.1016/s0167-9473(96)00073-4
Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test. Behavioral Ecology, 17(4), 688–690. https://doi.org/10.1093/beheco/ark016
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. https://doi.org/10.1037/0033-2909.111.2.352
Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). Biometrika, 52(3–4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591
Silventoinen, K. (2003). DETERMINANTS OF VARIATION IN ADULT BODY HEIGHT. Journal of Biosocial Science, 35(2), 263–285. https://doi.org/10.1017/s0021932003002633
Smiti, A. (2020). A critical overview of outlier detection methods. Computer Science Review, 38, 100306. https://doi.org/10.1016/j.cosrev.2020.100306
SOCR Data Dinov 020108 HeightsWeights - Socr. (n.d.). http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_Dinov_020108_HeightsWeights

Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. Journal of the American Statistical Association, 69(347), 730–737. https://doi.org/10.1080/01621459.1974.10480196
Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences of the United States of America, 100(16), 9440–9445. https://doi.org/10.1073/pnas.1530509100
Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2018). Using multivariate statistics.
Takahashi, S., Hyodo, M., Nishiyama, T., & Pavlenko, T. (2013). MULTIPLE COMPARISON PROCEDURES FOR HIGH-DIMENSIONAL DATA AND THEIR ROBUSTNESS UNDER NON-NORMALITY Journal of the Japanese Society of Computational Statistics, 26(1), 71–82. https://doi.org/10.5183/jjscs.1211001_202
Teh, S. Y., & Abdul Rahman, O. (2009). When does the pooled variance t-test fail? In Academic Journals, African Journal of Mathematics and Computer Science Research (Vols. 2–4, pp. 056–062). Academic Journals. https://academicjournals.org/journal/AJMCSR/article-full-text-pdf/0CC63383714
Tempesta, T., Giancristofaro, R. A., Corain, L., Salmaso, L., Tomasi, D., & Boatto, V. (2010). The importance of landscape in wine quality perception: An integrated approach using choice-based conjoint analysis and combination-based permutation tests. Food Quality and Preference, 21(7), 827–836. https://doi.org/10.1016/j.foodqual.2010.04.007
VanderWeele, T. J., & Mathur, M. B. (2018). SOME DESIRABLE PROPERTIES OF THE BONFERRONI CORRECTION: IS THE BONFERRONI CORRECTION REALLY SO BAD? American Journal of Epidemiology, 188(3), 617–618. https://doi.org/10.1093/aje/kwy250
Vinutha, H. P., Poornima, B., & Sagar, B. M. (2018). Detection of Outliers Using Interquartile Range Technique from Intrusion Dataset. In Advances in intelligent systems and computing (pp. 511–518). https://doi.org/10.1007/978-981-10-7563-6_53
Viviano, D., Wuthrich, K., & Niehaus, P. (2021). (When) should you adjust inferences for multiple hypothesis testing? ResearchGate. https://www.researchgate.net/publication/351119495_When_should_you_adjust_inferences_for_multiple_hypothesis_testing
Walpole, R. E., Myers, R., Myers, S., & Ye, K. (2006). Probability and Statistics for Engineers & Scienti. Pearson. https://brharnetc.edu.in/br/wp-content/uploads/2018/11/21.pdf
Wei, T. and Simko, V. (2019) R Package "CorrPlot" Visualization of a Correlation Matrix (Version 0.92). Available from https://github.com/taiyun/corrplot. (n.d.). https://scirp.org/reference/referencespapers?referenceid=3067218
Welch, W. J. (1990). Construction of permutation tests. Journal of the American Statistical Association, 85(411), 693–698. https://doi.org/10.1080/01621459.1990.10474929
Westfall, P. H., Tobias, R. D., & Wolfinger, R. D. (2011). Multiple comparisons and multiple tests using SAS, second edition. SAS Institute.
Westfall, P. H., & Young, S. S. (1993). Resampling-Based multiple testing: Examples and Methods for p-Value Adjustment. John Wiley & Sons.
Westphal, M., & Zapf, A. (2024). Statistical inference for diagnostic test accuracy studies with multiple comparisons. Statistical Methods in Medical Research, 33(4), 669–680. https://doi.org/10.1177/09622802241236933
Wickham, H. (2007). Reshaping Data with thereshapePackage. Journal of Statistical Software, 21(12). https://doi.org/10.18637/jss.v021.i12
Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis. Springer Science & Business Media.
Wickham, H., François, R., Henry, L. and Müller, K. (2023) Dplyr: A Grammar of Data Manipulation. R Package Version 1.1.4, <https://CRAN.R-project.org/package=dplyr>. (n.d.). https://www.scirp.org/reference/referencespapers?referenceid=2761682
Widerberg, C. (2019). The Two-Sample t-test and the Influence of Outliers : - A simulation study on how the type I error rate is impacted by outliers of different magnitude. DIVA. https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1284567&dswid=9516
Wilcox, R. R. (2012). Introduction to robust estimation and hypothesis testing. Academic Press.
Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. Biometrics Bulletin, 1(6), 80–83. https://sci2s.ugr.es/keel/pdf/algorithm/articulo/wilcoxon1945.pdf
Wilk, M. B., & Gnanadesikan, R. (1968). Probability plotting methods for the analysis for the analysis of data. Biometrika, 55(1), 1–17. https://doi.org/10.1093/biomet/55.1.1
Wilks, S. S. (1963). Multivariate Statistical Outliers. Sankhyā: The Indian Journal of Statistics, Series a (1961-2002), 25(4). https://www.jstor.org/stable/25049292

Winkler, A. M., Ridgway, G. R., Webster, M., Smith, S. M., & Nichols, T. E. (2014). Permutation inference for the general linear model. NeuroImage, 92, 381–397. https://doi.org/10.1016/j.neuroimage.2014.01.060

Yap, B. W., & Sim, C. H. (2011). Comparisons of various types of normality tests. Statistical Computation and Simulation/Journal of Statistical Computation and Simulation, 81(12), 2141–2155. https://doi.org/10.1080/00949655.2010.520163

Yule, G. U. (1897). On the Theory of Correlation. Journal of the Royal Statistical Society, 60(4), 812. https://doi.org/10.2307/2979746

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. British Journal of Mathematical & Statistical Psychology/British Journal of Mathematical and Statistical Psychology, 57(1), 173–181. https://doi.org/10.1348/000711004849222

Zumbo, B. D., & Jennings, M. (2002). The robustness of validity and efficiency of the related samples T-Test in the presence of outliers. DOAJ (DOAJ: Directory of Open Access Journals). https://doaj.org/article/603f2c32f5354a2fb1a85b0a7bf7aa40

7.2. Appendix A: Figures and Tables

Figure A

Boxplot of Height variable from the Hong Kong data

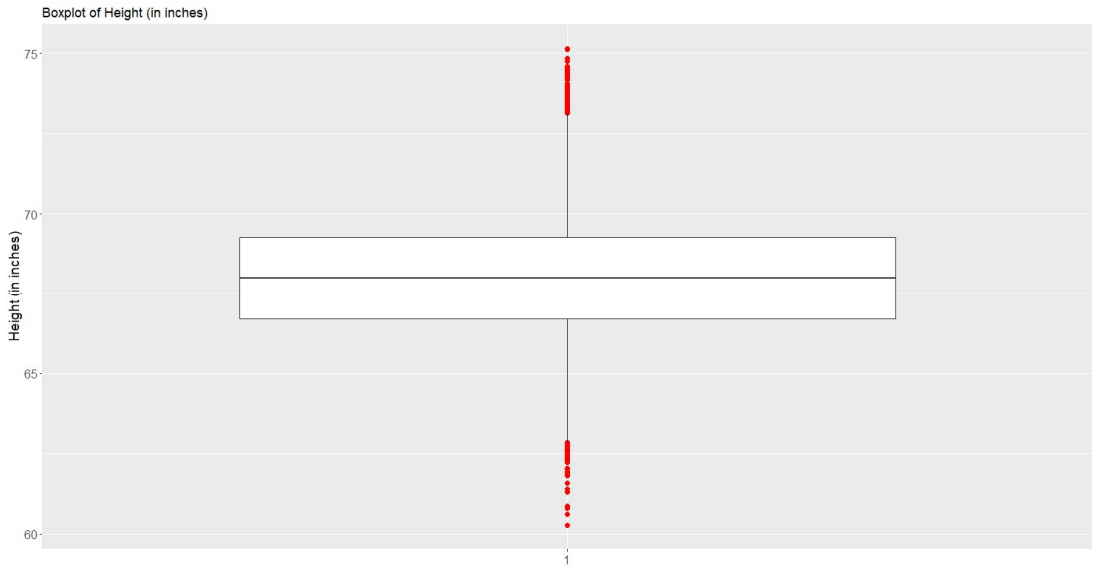


Figure B

Boxplot of Weight variable from the Hong Kong data

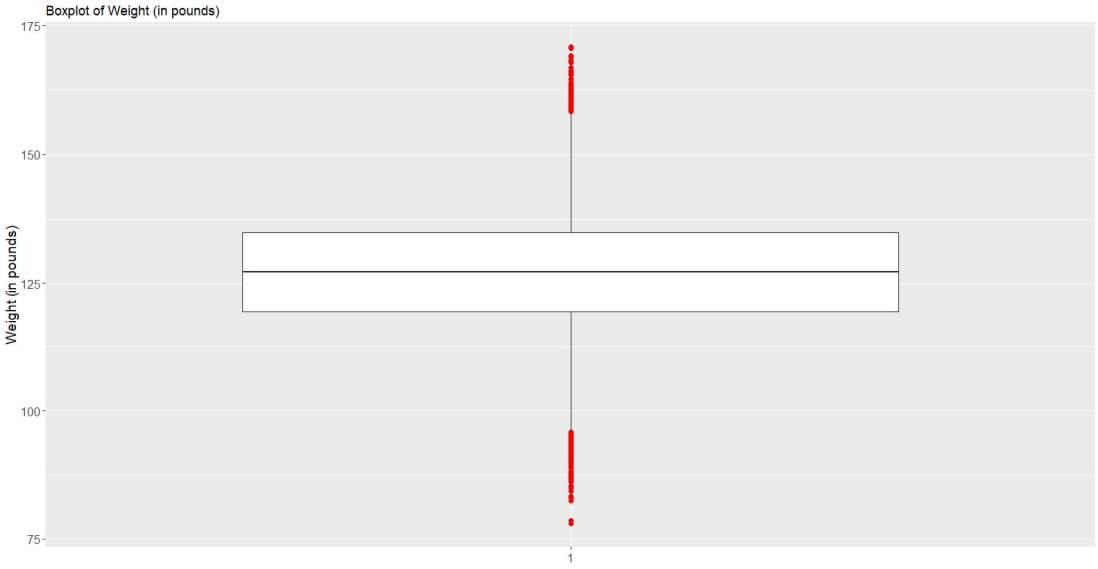


Figure C

Q-Q plot of Height of Hong Kong data

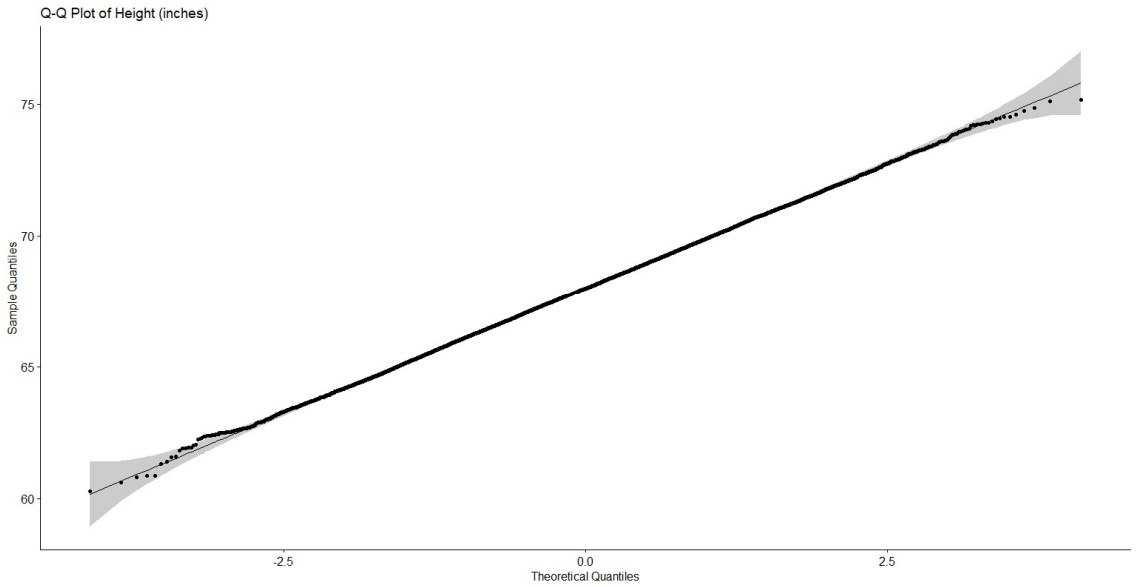


Figure D

Q-Q plot of Weight of Hong Kong data

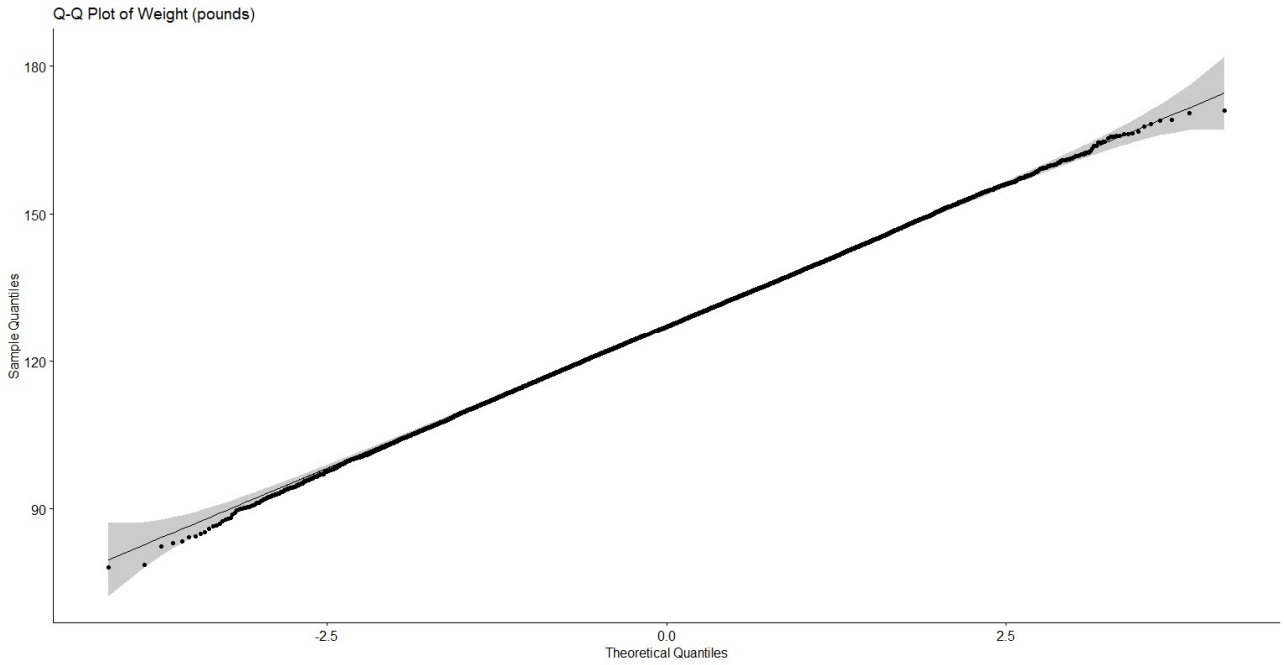


Table A

Overview of results of simulation for single comparison with increasing number of outliers for Height, including Wilcoxon signed-rank test on p-value differences, average difference in p-value and Type I error rate caused by the addition of the outlier for the Student's t-test, mean permutation test and median permutation test

Number of outliers	*Average p-value t-test*	*Average p-value permutation test, mean*	*Average p-value permutation test, median*	*Type I error t-test*	*Type I error permutation test, mean*	*Type I error permutation test, median*
1	0.333	0.547	0.495	0.0%	2.2%	0.0%
2	0.157	0.232	0.515	0.0%	7.8%	0.0%
3	0.082	0.110	0.443	0.0%	19.8%	1.0%
4	0.046	0.065	0.576	79.1%	33.0%	0.0%
5	0.026	0.035	0.454	95.8%	73.7%	0.0%

6	0.014	0.016	0.520	94.7%	94.7%	0.0%
7	0.008	0.009	0.460	92.6%	91.5%	1.1%
8	0.005	0.005	0.482	100%	100%	3.2%
9	0.003	0.002	0.533	96.8%	96.8%	0.0%
10	0.002	0.001	0.562	99.0%	99.0%	1.0%
11	0.001	0.000	0.451	93.3%	93.3%	3.4%
12	0.001	0.000	0.519	93.4%	92.3%	0.0%
13	0.000	0.000	0.513	93.7%	93.7%	1.1%
14	0.000	0.000	0.436	96.8%	95.7%	4.3%
15	0.000	0.000	0.478	94.7%	94.7%	3.2%
16	0.000	0.000	0.477	98.9%	97.8%	2.2%
17	0.000	0.000	0.434	93.5%	92.4%	1.1%
18	0.000	0.000	0.503	96.8%	95.8%	2.1%
19	0.000	0.000	0.520	95.6%	94.4%	2.2%
20	0.000	0.000	0.540	92.4%	92.4%	4.3%
21	0.000	0.000	0.503	94.3%	94.3%	3.4%
22	0.000	0.000	0.508	93.5%	93.5%	3.2%
23	0.000	0.000	0.486	95.6%	95.6%	3.3%
24	0.000	0.000	0.444	94.6%	94.6%	3.3%
25	0.000	0.000	0.450	91.6%	91.6%	4.2%
26	0.000	0.000	0.457	90.0%	91.1%	5.6%
27	0.000	0.000	0.417	93.8%	93.8%	9.4%
28	0.000	0.000	0.485	97.8%	96.7%	3.3%
29	0.000	0.000	0.478	96.7%	94.6%	6.5%
30	0.000	0.000	0.497	97.8%	97.8%	4.3%
31	0.000	0.000	0.484	94.6%	95.7%	3.2%
32	0.000	0.000	0.432	95.5%	95.5%	13.5%
33	0.000	0.000	0.418	95.7%	96.8%	14.0%
34	0.000	0.000	0.452	92.7%	92.7%	5.2%
35	0.000	0.000	0.424	92.0%	90.9%	13.6%
36	0.000	0.000	0.452	97.9%	97.9%	10.6%
37	0.000	0.000	0.391	96.8%	95.7%	2.1%
38	0.000	0.000	0.408	95.5%	96.6%	4.5%
39	0.000	0.000	0.367	93.0%	94.2%	9.3%
40	0.000	0.000	0.369	94.6%	93.5%	8.7%
41	0.000	0.000	0.395	97.9%	97.9%	16.8%
42	0.000	0.000	0.411	96.9%	96.9%	10.4%
43	0.000	0.000	0.340	96.6%	96.6%	15.7%
44	0.000	0.000	0.381	92.3%	92.3%	12.1%
45	0.000	0.000	0.352	91.2%	91.2%	16.5%
46	0.000	0.000	0.349	91.0%	91.0%	12.4%
47	0.000	0.000	*0.398*	96.6%	97.8%	14.6%
48	0.000	0.000	0.329	97.8%	97.8%	12.2%
49	0.000	0.000	*0.318*	93.7%	92.6%	25.3%
50	0.000	0.000	0.283	93.5%	93.5%	22.8%

Note: The average p-values are rounded to three decimals for better readability and interpretability. The Type I error values are percentages.

Table B

Overview of results of simulation for single comparison with increasing number of outliers for Weight, including Wilcoxon signed-rank test on p-value differences, average difference in p-value and Type I error rate caused by the addition of the outlier for the Student's t-test, mean permutation test and median permutation test

Number of outliers	*Average p-value t-test*	*Average p-value permutation test. mean*	*Average p-value permutation test. median*	*Type I error t-test*	*Type I error permutation test. mean*	*Type I error permutation test. median*
1	0.364	0.494	0.499	0.0%	5.4%	1.1%
2	0.167	0.225	0.493	0.0%	9.5%	1.1%
3	0.091	0.118	0.504	7.9%	19.1%	0.0%
4	0.055	0.070	0.486	40.9%	28.0%	0.0%
5	0.028	0.030	0.546	90.6%	83.5%	1.2%
6	0.016	0.015	0.486	92.5%	94.6%	0.0%
7	0.010	0.008	0.478	92.4%	93.5%	1.1%
8	0.006	0.004	0.477	92.0%	90.9%	1.1%
9	0.003	0.002	0.492	90.8%	90.8%	0.0%
10	0.002	0.001	0.422	92.0%	92.0%	0.0%
11	0.001	0.000	0.479	95.7%	95.7%	1.1%
12	0.001	0.000	0.496	94.6%	94.6%	2.2%
13	0.000	0.000	0.453	89.1%	90.2%	1.1%
14	0.000	0.000	0.452	96.6%	97.7%	3.4%
15	0.000	0.000	0.457	90.8%	93.1%	4.6%
16	0.000	0.000	0.549	97.7%	96.6%	3.4%
17	0.000	0.000	0.528	97.8%	97.8%	3.3%
18	0.000	0.000	0.476	95.7%	94.6%	5.4%
19	0.000	0.000	0.448	96.7%	96.7%	3.3%
20	0.000	0.000	0.468	95.4%	95.4%	1.1%
21	0.000	0.000	0.420	96.6%	95.5%	5.6%
22	0.000	0.000	0.485	95.6%	95.6%	6.7%
23	0.000	0.000	0.411	93.5%	94.6%	5.4%
24	0.000	0.000	0.453	94.3%	94.3%	6.8%
25	0.000	0.000	0.473	94.4%	92.1%	3.4%
26	0.000	0.000	0.453	90.1%	90.1%	3.3%
27	0.000	0.000	0.490	95.5%	95.5%	3.4%
28	0.000	0.000	0.440	95.3%	95.3%	7.1%
29	0.000	0.000	0.463	95.6%	95.6%	4.4%
30	0.000	0.000	0.417	97.6%	97.6%	9.4%
31	0.000	0.000	0.419	94.4%	94.4%	3.4%
32	0.000	0.000	0.463	92.9%	92.9%	3.5%
33	0.000	0.000	0.440	93.0%	95.3%	3.5%
34	0.000	0.000	0.436	98.9%	98.9%	6.5%
35	0.000	0.000	0.416	93.9%	93.9%	8.5%
36	0.000	0.000	0.359	91.0%	91.0%	13.5%

37	0.000	0.000	0.423	97.8%	97.8%	8.8%
38	0.000	0.000	0.419	94.3%	94.3%	10.3%
39	0.000	0.000	0.385	93.6%	95.7%	8.5%
40	0.000	0.000	0.426	98.9%	97.8%	17.2%
41	0.000	0.000	0.382	94.0%	95.2%	9.6%
42	0.000	0.000	0.424	94.0%	92.9%	13.1%
43	0.000	0.000	0.355	93.1%	93.1%	10.3%
44	0.000	0.000	0.339	92.8%	92.8%	15.7%
45	0.000	0.000	0.428	93.0%	93.0%	12.8%
46	0.000	0.000	0.340	94.3%	95.5%	14.8%
47	0.000	0.000	0.400	96.7%	95.6%	12.2%
48	0.000	0.000	0.359	90.4%	89.4%	16.0%
49	0.000	0.000	0.382	97.7%	97.7%	17.2%
50	0.000	0.000	0.336	94.4%	95.5%	18.0%

Note: The average p-values are rounded to three decimals for better readability and interpretability. The Type I error values are percentages.

Table C

Overview of results of simulation for single comparison with growing outlier sizes for Height, including Wilcoxon signed-rank test on p-value differences, average difference in p-value and Type I error rate caused by the addition of the outlier for the Student's t-test, mean permutation test and median permutation test

Outlier size factor	*Average p-value t-test*	*Average p-value permutation test. mean*	*Average p-value permutation test. median*	*Type I error t-test*	*Type I error permutation test. mean*	*Type I error permutation test. median*
0	0.546	0.545	0.452	0.0%	0.0%	0.0%
1	0.495	0.514	0.560	0.0%	0.0%	0.0%
2	0.363	0.480	0.456	0.0%	3.0%	0.0%
3	0.337	0.521	0.499	0.0%	4.0%	1.0%
4	0.326	0.539	0.522	0.0%	2.0%	0.0%
5	0.317	0.493	0.498	0.0%	2.0%	0.0%
6	0.318	0.515	0.460	0.0%	2.0%	0.0%
7	0.317	0.486	0.532	0.0%	1.0%	0.0%
8	0.317	0.491	0.515	0.0%	1.0%	0.0%
9	0.318	0.569	0.500	0.0%	1.0%	1.0%
10	0.318	0.627	0.492	0.0%	0.0%	1.0%
11	0.318	0.717	0.526	0.0%	0.0%	0.0%
12	0.318	0.768	0.473	0.0%	0.0%	0.0%
13	0.318	0.746	0.525	0.0%	0.0%	0.0%
14	0.318	0.742	0.523	0.0%	0.0%	2.0%
15	0.318	0.775	0.459	0.0%	0.0%	0.0%
16	0.318	0.741	0.498	0.0%	0.0%	1.0%
17	0.318	0.756	0.506	0.0%	0.0%	0.0%
18	0.318	0.736	0.499	0.0%	0.0%	2.0%
19	0.318	0.761	0.536	0.0%	0.0%	0.0%
20	0.318	0.763	0.461	0.0%	0.0%	0.0%
21	0.318	0.763	0.509	0.0%	0.0%	0.0%

22	0.318	0.770	0.534	0.0%	0.0%	0.0%
23	0.318	0.740	0.431	0.0%	0.0%	0.0%
24	0.318	0.736	0.549	0.0%	0.0%	0.0%
25	0.318	0.747	0.432	0.0%	0.0%	2.0%
26	0.318	0.757	0.500	0.0%	0.0%	1.0%
27	0.318	0.722	0.526	0.0%	0.0%	0.0%
28	0.318	0.772	0.533	0.0%	0.0%	1.0%
29	0.318	0.791	0.482	0.0%	0.0%	1.0%
30	0.318	0.753	0.490	0.0%	0.0%	0.0%
31	0.318	0.760	0.503	0.0%	0.0%	0.0%
32	0.318	0.769	0.546	0.0%	0.0%	0.0%
33	0.318	0.773	0.488	0.0%	0.0%	0.0%
34	0.318	0.736	0.486	0.0%	0.0%	1.0%
35	0.318	0.752	0.518	0.0%	0.0%	0.0%
36	0.318	0.751	0.517	0.0%	0.0%	1.0%
37	0.318	0.751	0.464	0.0%	0.0%	0.0%
38	0.318	0.737	0.488	0.0%	0.0%	0.0%
39	0.318	0.747	0.455	0.0%	0.0%	1.0%
40	0.318	0.762	0.454	0.0%	0.0%	1.0%
41	0.318	0.758	0.451	0.0%	0.0%	0.0%
42	0.318	0.743	0.524	0.0%	0.0%	0.0%
43	0.318	0.749	0.508	0.0%	0.0%	0.0%
44	0.318	0.765	0.515	0.0%	0.0%	0.0%
45	0.318	0.753	0.518	0.0%	0.0%	1.0%
46	0.318	0.752	0.517	0.0%	0.0%	1.0%
47	0.318	0.745	0.523	0.0%	0.0%	0.0%
48	0.318	0.755	0.491	0.0%	0.0%	1.0%
49	0.318	0.768	0.518	0.0%	0.0%	0.0%
50	0.318	0.753	0.508	0.0%	0.0%	0.0%

Note: The average p-values are rounded to three decimals for better readability and interpretability. The Type I error values are percentages.

Table D

Overview of results of simulation for single comparison with growing outlier sizes for Weight, including Wilcoxon signed-rank test on p-value differences, average difference in p-value and Type I error rate caused by the addition of the outlier for the Student's t-test, mean permutation test and median permutation test

Outlier size factor	*Average p-value t-test*	*Average p-value permutation test. mean*	*Average p-value permutation test. median*	*Type I error t-test*	*Type I error permutation test. mean*	*Type I error permutation test. median*
0	0.504	0.502	0.483	0.0%	2.0%	1.0%
1	0.456	0.455	0.492	1.0%	2.0%	0.0%
2	0.507	0.539	0.483	1.0%	2.0%	2.0%
3	0.377	0.489	0.500	0.0%	4.0%	1.0%
4	0.331	0.486	0.454	0.0%	1.0%	0.0%
5	0.318	0.475	0.515	0.0%	3.0%	1.0%
6	0.318	0.481	0.482	0.0%	4.0%	0.0%

7	0.315	0.440	0.534	0.0%	3.0%	1.0%
8	0.319	0.548	0.511	0.0%	0.0%	0.0%
9	0.318	0.537	0.524	0.0%	2.0%	0.0%
10	0.318	0.576	0.529	0.0%	0.0%	1.0%
11	0.318	0.627	0.505	0.0%	0.0%	0.0%
12	0.318	0.667	0.490	0.0%	0.0%	0.0%
13	0.318	0.756	0.505	0.0%	0.0%	0.0%
14	0.318	0.750	0.476	0.0%	0.0%	0.0%
15	0.318	0.756	0.532	0.0%	0.0%	0.0%
16	0.318	0.777	0.501	0.0%	0.0%	0.0%
17	0.318	0.756	0.465	0.0%	0.0%	0.0%
18	0.318	0.752	0.496	0.0%	0.0%	0.0%
19	0.318	0.749	0.493	0.0%	0.0%	0.0%
20	0.318	0.764	0.518	0.0%	0.0%	2.0%
21	0.318	0.754	0.481	0.0%	0.0%	0.0%
22	0.318	0.750	0.496	0.0%	0.0%	1.0%
23	0.318	0.753	0.496	0.0%	0.0%	0.0%
24	0.318	0.751	0.501	0.0%	0.0%	1.0%
25	0.318	0.759	0.505	0.0%	0.0%	1.0%
26	0.318	0.761	0.503	0.0%	0.0%	0.0%
27	0.318	0.760	0.487	0.0%	0.0%	2.0%
28	0.318	0.755	0.486	0.0%	0.0%	1.0%
29	0.318	0.769	0.500	0.0%	0.0%	0.0%
30	0.318	0.746	0.493	0.0%	0.0%	0.0%
31	0.318	0.741	0.493	0.0%	0.0%	1.0%
32	0.318	0.741	0.503	0.0%	0.0%	0.0%
33	0.318	0.752	0.493	0.0%	0.0%	0.0%
34	0.318	0.736	0.509	0.0%	0.0%	0.0%
35	0.318	0.729	0.515	0.0%	0.0%	0.0%
36	0.318	0.753	0.537	0.0%	0.0%	1.0%
37	0.318	0.736	0.482	0.0%	0.0%	1.0%
38	0.318	0.739	0.501	0.0%	0.0%	0.0%
39	0.318	0.745	0.485	0.0%	0.0%	1.0%
40	0.318	0.760	0.487	0.0%	0.0%	0.0%
41	0.318	0.753	0.521	0.0%	0.0%	0.0%
42	0.318	0.749	0.569	0.0%	0.0%	0.0%
43	0.318	0.756	0.488	0.0%	0.0%	1.0%
44	0.318	0.780	0.518	0.0%	0.0%	0.0%
45	0.318	0.754	0.460	0.0%	0.0%	1.0%
46	0.318	0.754	0.467	0.0%	0.0%	0.0%
47	0.318	0.743	0.490	0.0%	0.0%	1.0%
48	0.318	0.746	0.510	0.0%	0.0%	0.0%
49	0.318	0.741	0.528	0.0%	0.0%	1.0%
50	0.318	0.738	0.553	0.0%	0.0%	2.0%

Note: The average p-values are rounded to three decimals for better readability and interpretability. The Type I error values are percentages.

Table E

Overview of results of single outlier robustness simulations with increasing number of outliers, including the number of significant genes after outlier addition and Type II error rate

Number of outliers	*Significant genes Bonferroni*	*Significant genes maxT*	*Type II error rate Bonferroni*	*Type II error rate maxT*
1	3	9	2.1%	5.4%
2	73	57	51.0%	34.1%
3	105	113	73.4%	67.7%
4	78	86	54.5%	51.5%
5	114	113	79.7%	67.7%
6	107	109	74.8%	65.3%
7	73	80	51.0%	47.9%
8	92	79	64.3%	47.3%
9	103	106	72.0%	63.5%
10	122	130	85.3%	77.8%
11	126	139	88.1%	83.2%
12	127	139	88.8%	83.2%
13	100	92	69.9%	55.1%
14	119	133	83.2%	79.6%
15	135	146	94.4%	87.4%
16	96	94	67.1%	56.2%
17	137	151	95.8%	90.4%
18	138	153	96.5%	91.6%
19	122	110	85.3%	65.9%
20	128	131	89.5%	78.4%
21	141	164	98.6%	98.2%
22	139	156	97.2%	93.4%
23	98	54	68.5%	32.3%
24	135	145	94.4%	86.8%
25	142	163	99.3%	97.6%
26	142	165	99.3%	98.8%
27	142	166	99.3%	99.4%
28	125	140	87.4%	83.8%
29	135	141	94.4%	84.4%
30	142	165	99.3%	98.8%
31	142	163	99.3%	97.6%
32	143	167	100.0%	100.0%
33	143	166	100.0%	99.4%
34	141	162	98.6%	97.0%
35	140	158	97.9%	94.6%
36	143	165	100.0%	98.8%
37	143	167	100.0%	100.0%
38	142	164	99.3%	98.2%
39	143	167	100.0%	100.0%
40	143	166	100.0%	99.4%
41	142	165	99.3%	98.8%

42	143	167	100.0%	100.0%
43	142	162	99.3%	97.0%
44	143	167	100.0%	100.0%
45	142	166	99.3%	99.4%
46	136	147	95.1%	88.0%
47	143	167	100.0%	100.0%
48	142	163	99.3%	97.6%
49	143	164	100.0%	98.2%
50	143	163	100.0%	97.6%

Note: The Type II error rate indicates the proportion of the genes that were found to be significant compared to the baseline. These Type II errors were false negatives caused by the introduction of the outlier.

Table F

Overview of results of single outlier robustness simulations with single outlier growing in size for AML, including the number of significant genes after outlier addition and Type II error rate

Outlier factor	*Significant genes Bonferroni*	*Significant genes maxT*	*Type II error rate Bonferroni*	*Type II error rate maxT*
1	140	169	2.1%	1.2%
2	67	93	53.1%	44.3%
3	89	133	37.8%	20.4%
4	30	66	79.0%	60.5%
5	61	112	57.3%	32.9%
6	37	78	74.1%	53.3%
7	34	69	76.2%	58.7%
8	91	146	36.4%	12.6%
9	44	78	69.2%	53.3%
10	39	83	72.7%	50.3%
11	31	58	78.3%	65.3%
12	17	49	88.1%	70.7%
13	20	53	86.0%	68.3%
14	21	52	85.3%	68.9%
15	28	58	80.4%	65.3%
16	7	31	95.1%	81.4%
17	21	54	85.3%	67.7%
18	17	51	88.1%	69.5%
19	53	111	62.9%	33.5%
20	18	53	87.4%	68.3%
21	14	47	90.2%	71.9%
22	23	47	83.9%	71.9%
23	14	51	90.2%	69.5%
24	25	62	82.5%	62.9%
25	16	33	88.8%	80.2%
26	42	96	70.6%	42.5%
27	14	29	90.2%	82.6%
28	55	105	61.5%	37.1%
29	21	56	85.3%	66.5%

30	15	48	89.5%	71.3%
31	3	19	97.9%	88.6%
32	17	50	88.1%	70.1%
33	12	24	91.6%	85.6%
34	6	16	95.8%	90.4%
35	10	38	93.0%	77.2%
36	9	25	93.7%	85.0%
37	19	58	86.7%	65.3%
38	9	27	93.7%	83.8%
39	21	50	85.3%	70.1%
40	8	20	94.4%	88.0%
41	18	40	87.4%	76.0%
42	7	30	95.1%	82.0%
43	11	27	92.3%	83.8%
44	20	60	86.0%	64.1%
45	10	27	93.0%	83.8%
46	14	35	90.2%	79.0%
47	8	30	94.4%	82.0%
48	9	33	93.7%	80.2%
49	10	37	93.0%	77.8%
50	5	19	96.5%	88.6%

Note: The Type II error rate indicates the proportion of the genes that were found to be significant compared to the baseline. These Type II errors were false negatives caused by the introduction of the outlier.

Table G

Overview of results of single outlier robustness simulations with single outlier growing in size for ALL, including the number of significant genes after outlier addition and Type II error rate

Outlier factor	*Significant genes Bonferroni*	*Significant genes maxT*	*Type II error rate Bonferroni*	*Type II error rate maxT*
1	101	127	29.4%	24.0%
2	139	168	2.8%	0.6%
3	101	132	29.4%	21.0%
4	45	81	68.5%	51.5%
5	119	145	16.8%	13.2%
6	78	124	45.5%	25.7%
7	56	84	60.8%	49.7%
8	78	131	45.5%	21.6%
9	46	80	67.8%	52.1%
10	18	39	87.4%	76.6%
11	13	38	90.9%	77.2%
12	49	108	65.7%	35.3%
13	44	97	69.2%	41.9%
14	21	52	85.3%	68.9%
15	8	27	94.4%	83.8%
16	20	60	86.0%	64.1%
17	34	61	76.2%	63.5%

18	23	59	83.9%	64.7%
19	7	25	95.1%	85.0%
20	13	47	90.9%	71.9%
21	22	43	84.6%	74.3%
22	3	25	97.9%	85.0%
23	9	29	93.7%	82.6%
24	13	39	90.9%	76.6%
25	30	60	79.0%	64.1%
26	16	34	88.8%	79.6%
27	31	65	78.3%	61.1%
28	13	44	90.9%	73.7%
29	11	39	92.3%	76.6%
30	15	51	89.5%	69.5%
31	38	80	73.4%	52.1%
32	31	80	78.3%	52.1%
33	11	33	92.3%	80.2%
34	12	23	91.6%	86.2%
35	9	20	93.7%	88.0%
36	17	43	88.1%	74.3%
37	15	45	89.5%	73.1%
38	6	20	95.8%	88.0%
39	8	25	94.4%	85.0%
40	13	35	90.9%	79.0%
41	11	21	92.3%	87.4%
42	10	35	93.0%	79.0%
43	3	26	97.9%	84.4%
44	9	34	93.7%	79.6%
45	18	42	87.4%	74.9%
46	7	16	95.1%	90.4%
47	28	57	80.4%	65.9%
48	9	29	93.7%	82.6%
49	15	48	89.5%	71.3%
50	8	29	94.4%	82.6%

Note: The Type II error rate indicates the proportion of the genes that were found to be significant compared to the baseline. These Type II errors were false negatives caused by the introduction of the outlier.