Exploring the role of permutations in the detection of errors and anomalies in statistical analysis

Master's thesis

Ludwig-Maximilians-Universität München
Department of Statistics

Anna Maria Orzelek

Supervised by Prof. Dr. Anna-Laure Boulesteix

May 23th, 2024

Abstract

This thesis explores how permutation checks can help detect errors and anomalies in statistical analysis. Permutations, which involve randomly reassigning the values of the outcome variable, can uncover these issues when the expected errors are unknown. This method breaks any relationship between the explanatory variables and the outcome, allowing for an error check before the actual analysis. Here the focus is on binary outcomes, where permutations are straightforward. The process involves conducting the planned analysis on permuted data multiple times and comparing the results. If the results are similar across different permutations, it suggests there might be problems in the analysis, allowing further investigation before proceeding with the original data. This thesis employs permutation checks on simulated and real-world data, assessing a bias of the chi-square and the Cochran-Mantel-Haenszel test along with Yate's correction for continuity. Practical considerations are also discussed, such as the effect of unevenly distributed outcome variables and the number of permutations needed to detect errors. The findings show that permutations are a useful pre-analysis check to improve the accuracy of statistical conclusions. Future research could expand the use of permutations to non-binary and continuous outcomes.

Contents

List of Figures

List of Tables

1 Introduction

When following a statistical workflow and conducting analyses, it is almost impossible to rule out any potential error that could occur in the process. In fact, the famous paper from Ioannidis (2005) claims that most of the published research findings are wrong. As one of the sources of that phenomenon, he mentions bias that could be introduced for example by measurement errors or overlooked problems in the analysis method. As a consequence, the results may be over-optimistic or even incorrect. If one knows which errors or biases may be expected, it is possible to control or test for them, as there are many specialized methods for that. Think of Principal Components Analysis, helping to detect outliers (Gewers et al. 2021). How should one proceed when wanting to make sure the analysis results are correct but not knowing which errors to expect and thus lacking in methods to apply? In that case, permuting the outcome variable is a tool to help detect errors and anomalies in the analysis process. In this context, permuting means randomly assigning the value of the outcome and thus breaking a possible relation between the explanatory variables and the outcome of interest. The workflow follows this pattern: first, the planned analysis is conducted on permuted data. Several repetitions of the permutation and analysis process have to be done, and the results of the permutations are compared. If similar or identical results occur across multiple permutations, it may indicate that there are underlying issues. That way it is already possible to uncover potential errors and pitfalls before even conducting the analysis on the original data. It is expected that after the permutations, the analysis results in no relationship between the outcome and the explanatory variables. If nevertheless the results across different permutations are similar, it can be interpreted as a possible error during the data analysis process. The origins of this error can then be investigated, and the issues resolved before the final analysis is conducted.

As this thesis will show, the errors can have various sources, such as undetected outliers in the data or a bias in a statistical method. The focus will be on binary outcomes, as permuting them is straightforward and unambiguous. For continuous outcomes, the optimal permutation method is less clear. Note that the data the original analysis is conducted with, before the permutations, will be referred to as “original data”, whereas the data used for the analysis after the permutations will be called “permuted data”. The results will be color-coded, with the analysis results on the original data being visualized in orange and the results on the permuted data visualized in green. Furthermore, all results in the tables are rounded to one decimal place.

The idea to employ permutations in the context of error detection originates from two previous publications: “Random forest Gini importance favours SNPs with large minor allele frequency: impact, sources and recommendations” by Boulesteix et al. (2012) and “Over-Optimism in Gene Set Analyses: how does the choice of methods and tools influence the detection of enriched gene sets?” by Wünsch (2022). In both cases, random permutations of the outcome variable are used to identify errors along the analysis process, which have an impact on the results. The paper by Boulesteix et al. (2012) delves into the ranking of genetic polymorphisms through the commonly used Gini variance importance measure (Gini VIM). Through random permutations of the outcome variable, it is shown that the Gini VIM systematically favours common polymorphism variants. When unaware of this bias, common polymorphisms are falsely interpreted as playing an important role in predicting the outcome phenotype, while the predictive importance of uncommon polymorphisms is overlooked. In this case, permutations reveal a bias in a statistical method that is widely used for important research findings. Passing over such pitfalls can seriously undermine the reliability and validity of studies.

This thesis will highlight the role of permutations in the detection of errors and anomalies in the statistical analysis process by showing several use cases and application fields of the permutation checks. To begin this thesis, in Chapter 2 the previously mentioned paper by Wünsch (2022) will be elaborated further, breaking down the discovered error into easier examples, thanks to which the use of the permutations will be introduced. In Chapter 3, the focus will be on the paper of Boulesteix et al. (2012). The bias which is described thanks to the permutations will be investigated further in different use cases. Again, the role of permutations in finding the bias will be explored, using simulated data as well as real-world data. Based on the discovered bias in the previous chapter, in Chapter 4 the permutation checks will be applied to the Cochran-Mantel-Haenszel test in order to investigate if the bias extends to it. Furthermore, Chapter 5 will cover some practical considerations about the application of the permutation process, considering especially the case of unevenly distributed outcome variables and the needed number of permutations. In Chapter 6, the final part of this thesis will give a summary of the main results and aspects of the role of permutations, as well as an outlook.

The results were obtained using version 4.3.3 of the Software R (R Core Team 2021). To create the result plots, the following packages were used: `ggplot2` (H. Wickham, Chang, and M. H. Wickham 2016), `gridExtra` (B. Auguie, Antonov, and M. B. Auguie 2017) and `reshape2` (H. Wickham 2020).

2 Introductory examples

2.1 t-test and outlier

The first introductory example is based on the thesis by Wünsch (2022), which explores the variety of gene set analysis, a widely-used method for interpreting high-throughput gene expression data. It explains the multiple methodologies and tools available for conducting gene set analysis while highlighting the common practice of optimizing parameters in these tools, leading to over-optimistic results that lack validity. Those parameter adjustments can artificially inflate the number of detected differentially enriched gene sets. As part of the evaluation, the outcome variable is randomly permuted 5 times to assess the tool's qualities. It is expected that after those random permutations, no gene sets should be systematically identified as differentially enriched since the relationship between the gene sets and the outcome variable is disrupted. It shows that, when using one specific group of tools, the over-representation analysis, multiple gene sets are identified as differentially enriched across several of the random phenotype permutations. After further investigations of the gene sets, this finding reveals count outliers of a small number of genes that highly influenced the analysis results.

In this case, the permutations of the outcome variable help to find an undiscovered outlier. In the following, the work of Wünsch (2022) is translated to a more general setting. The aim is to highlight the usability of the permutation checks by showing how they can reveal an outlier that influences the results of a two-sample t-test. Two standard normally distributed samples are generated $X_1, X_2 \sim \mathcal{N}(0,1)$ with 100 observations in both of the samples. In the second sample, one of the observations is replaced by an outlier value of 25. This data generating process is done 1000 times and a two-sample t-test is performed for each of those simulations. The test hypotheses are $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$ with μ_1 and μ_2 being the mean values of the samples X_1 and X_2 respectively. The test statistic is given by:

$$T = \frac{\mu_2 - \mu_1}{\sqrt{\frac{1}{n}\frac{(n-1)(S_1^2+S_2^2)}{2n-2}}}, \tag{1}$$

with S_1^2 and S_2^2 being the sample variances (Manfei et al. 2017). Note that since the sample sizes are equal, the formula is simplified and n is used as common sample size instead of n_1 and n_2 . The significance threshold for the p-values is set to $\alpha = 5\%$.

Next, the permutation process starts: by randomly permuting the sample membership of the $n = 200$ observations between the samples X_1 and X_2, the group-belonging of the observations is disrupted and the random permutations are created. The permuted samples have the same sample sizes as before and sampling is done without replacement. The outlier observation that was in sample X_2 before, is now randomly in one of the samples. Using those two permuted samples, the t-test is repeated. This permutation and test scheme is repeated 10 times for every simulation. The results are finally plotted in terms of p-values.

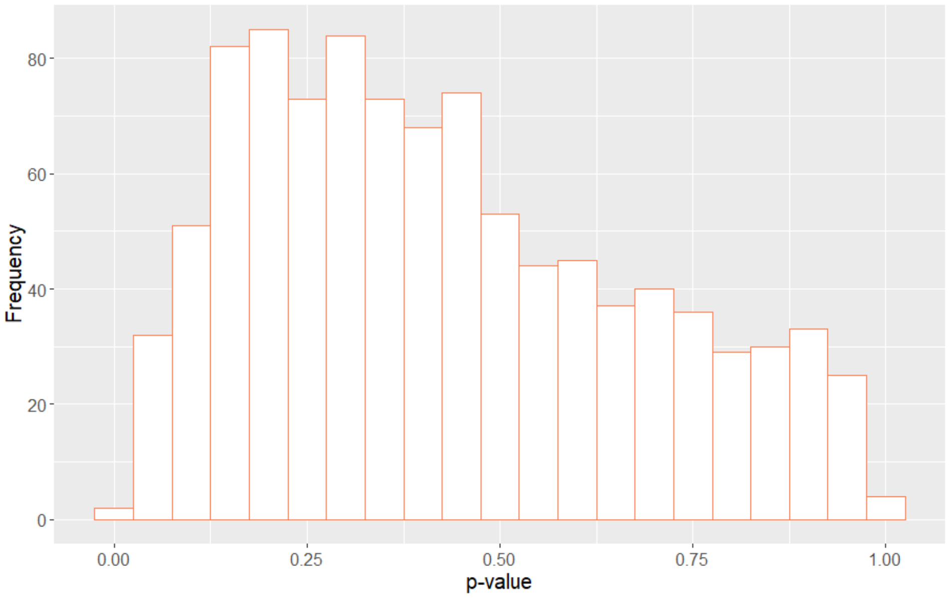

Figure 1: Absolute frequency distribution of the p-values of 1000 two-sample t-tests. The samples have 100 observations each and are normally distributed $X_1, X_2 \sim \mathcal{N}(0, 1)$ with an outlier value of 25 in the second sample.

Figure 1 shows the absolute frequency distribution of the 1000 p-values for the two-sample t-test on the original data with an outlier in one of the samples. As anticipated, due to the outlier, the p-values are not uniformly distributed as it would be expected under the null hypothesis (Fodor, Tickle, and Richardson 2007). The distribution is right-skewed with a mean of approximately 0.43 and a median of approximately 0.39. In practice, if the outlier remains undiscovered, there is a higher possibility to falsely assume a difference in the means for the two samples, and thus falsely rejecting the null hypothesis.

The absolute frequency distribution of the p-values after 10 permutation and test procedures of each simulation is depicted in Figure 2, looking very similar to the previous distribution. Note that the y-axes of the two plots are not identical, as for every single simulation and testing iteration of the original data, 10 permutations and testing iterations are done. That way, the permutations reveal the outlier observation, by showing that the p-values are not uniformly distributed even when a possible difference of the means is taken out of the data.

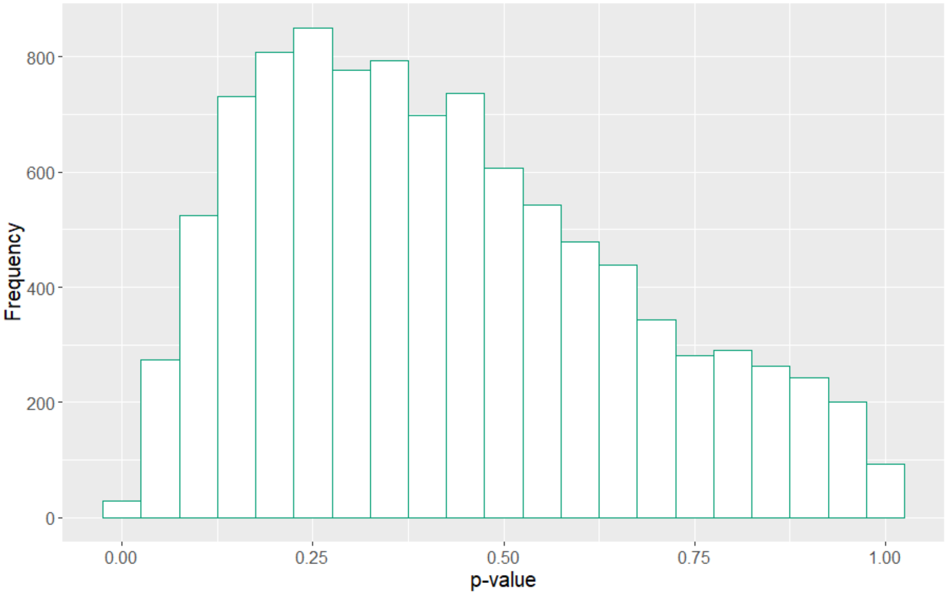

Figure 2: Absolute frequency distribution of the p-values of two-sample t-tests. 1000 simulations of two samples $X_1, X_2 \sim \mathcal{N}(0,1)$ with an outlier value of 25 in the second sample are done, then the samples are permuted 10 times. A two-sample t-test is performed on each permutation, yielding 10000 p-values.

When the permutation and testing process is done on two samples without any outliers, the p-values are uniformly distributed as it would be expected under the null hypothesis, seen in Figure 3. This confirms that the permutation process does not influence the results in a way that an effect or bias is created. The skewed distribution in Figure 2 is only a result of the outlier observation in one of the samples.

Even though in this simple scenario an outlier would be discovered before testing, it shows an application field for the permutation process. In some research fields with high-dimensional data, such as gene set analysis, not every explanatory variable, every gene, can be inspected with regard to outliers before the analysis. In those cases, permutations can be used to check if the results are influenced by outliers, as seen in Wünsch (2022).

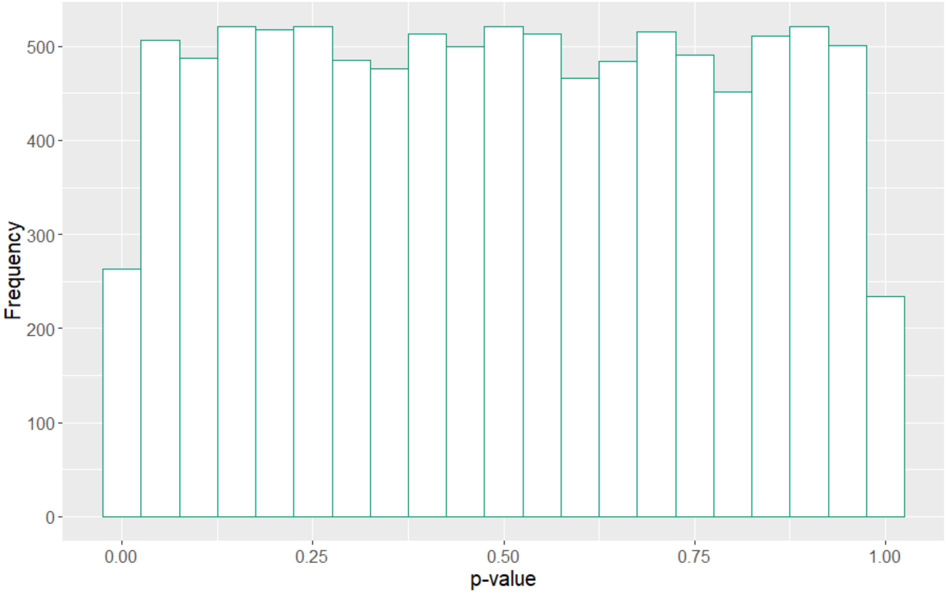

Figure 3: Absolute frequency distribution of the p-values of two-sample t-tests. 1000 simulations of two samples $X_1, X_2 \sim \mathcal{N}(0,1)$ are done, then the samples are permuted 10 times. A two-sample t-test is performed on each permutation, yielding 10000 p-values. Note: the first and last bin are smaller than the others because of the plot coding and not because of true differences in the p-value frequencies there.

2.2 t-test and bootstrap

The following example is similar to the previous one as the conducted analysis is also a t-test. Here the goal is to illustrate how the permutation checks help to reveal an increase in the type I error of the t-test that is induced by bootstrap-sampling. Bootstrapping is a resampling technique commonly used to estimate statistics more precisely by repeatedly sampling with replacement from the observed data (Davison and Hinkley 1997).

Again, two standard normally distributed samples $X_1, X_2 \sim \mathcal{N}(0,1)$ with 100 observations each are generated, those are the original samples. Based on those two samples, bootstrapping is performed, resulting in two new samples on which a t-test is applied. In this step, no permutations are involved. As previously, the simulation, bootstrapping and testing procedure is done 1000 times. The results can be observed in Figure 4, showing the absolute frequency distribution of the p-values for the two-sample t-test after bootstrap sampling. The right-skewed distribution indicates an increased type I error. With 17% of the p-values being smaller or equal to 0.05, even though the original samples have identical means, too many of the t-tests yield significant p-values.

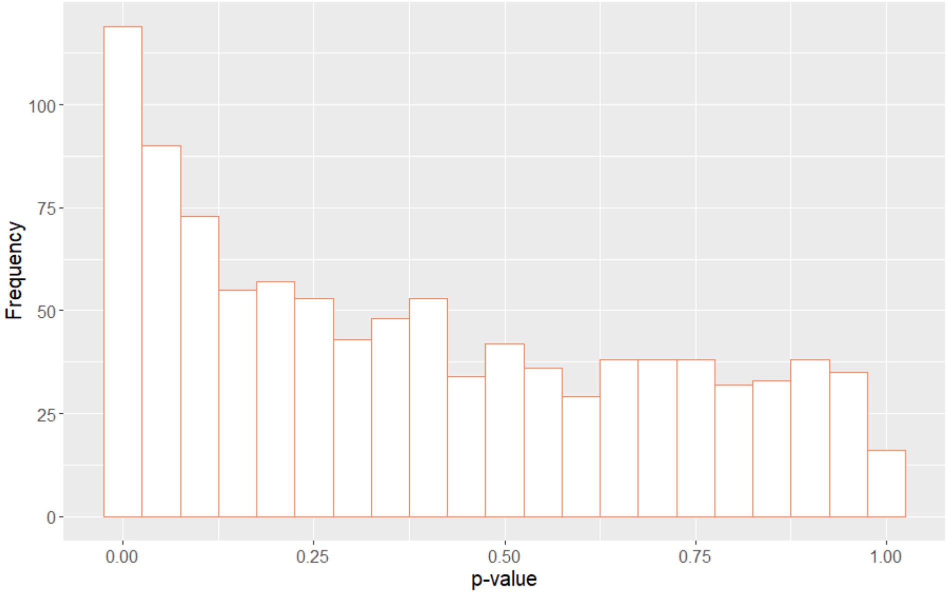

Figure 4: Absolute frequency distribution of the p-values for 1000 two-sample t-test. The samples have 100 observations each and are normally distributed $X_1, X_2 \sim \mathcal{N}(0,1)$. Bootstrap sampling is performed before testing.

When unsure if in such a scenario the significant p-values are a result of a true effect or not, one can again employ permutations. By permuting the original samples, then performing bootstrapping and testing 10 times as before for each of the 1000 simulations, any possible difference in the means of the original samples gets erased and the p-values should be uniformly distributed. Figure 5 displays the distribution of the p-values after this procedure. It can be seen that even after permuting, the distribution is still right-skewed, indicating that the increase in significant p-values results from bootstrapping. A possible reason for this is a change in the sample means that arises through the bootstrapping procedure, resulting in unequal means of the samples which are then detected by the t-test.

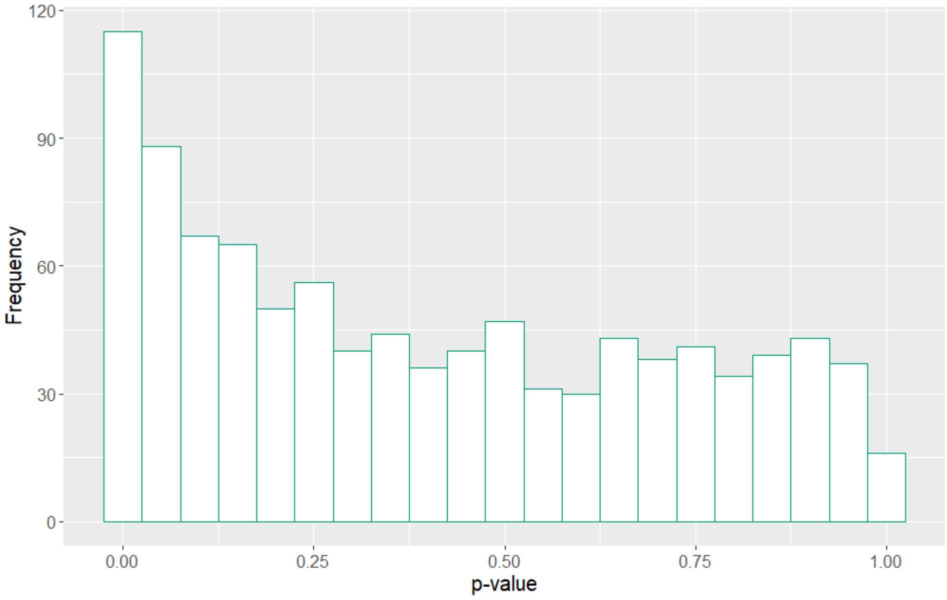

Figure 5: Absolute frequency distribution of the p-values of two-sample t-tests. 1000 simulations of two samples $X_1, X_2 \sim \mathcal{N}(0,1)$ are done, then the samples are permuted 10 times. Bootstrap sampling and a two-sample t-test is performed on each permutation, yielding 10000 p-values.

Even though the scenarios are not very complex, the two previous examples were able to demonstrate simple use cases of the permutation checks, highlighting their ability to uncover errors during the statistical analysis. Next the work of Boulesteix et al. (2012) is elaborated further, applying the permutation checks to the chi-square test.

3 Application of the permutation checks to assess a bias of the chi-square test

3.1 Background and aim

This chapter aims to show a more complex application of the permutation process by picking up on the work of Boulesteix et al. (2012). The paper inspects a bias originating from the Gini variable importance measure (Gini VIM) that is calculated during random forest algorithms. The bias is revealed through the use case of genetic data where the connection between genetic variants and a phenotype of interest is inspected. The single nucleotide polymorphisms (SNPs), which are a type of genetic variation, are the explanatory variables. They are categorical, each SNP can have three different variants. As some variants are very rare and may not occur at all, the variables have sometimes only two instead of three groups. The phenotype of interest, being the outcome variable, is binary. Through simulations and permutation checks, it is shown that the Gini VIM favours evenly distributed groups of the explanatory variable.

Furthermore, Boulesteix et al. (2012) refer to a publication by Grabmeier and Lambe (2007) showing an equality between the Gini impurity measure and Pearson's chi-square statistic for binary classification variables. This equality implies that the bias is not only present when using the Gini VIM but also in the application of the chi-square test. Boulesteix et al. elaborate on the bias of the chi-square statistic, stating that it is also present with binary and three-categorical predictors outside of the genetic framework. Based on this finding, the goal of this chapter is to inspect if the bias created by the Gini VIM in favour of evenly distributed explanatory variables can also be observed when performing a chi-square test. To achieve this, the bias will be inspected on simulated datasets that always contain four different explanatory variables with varying group proportions as well as on the real-world HapMap dataset, all that with the help of permutation checks. It is important to note that besides the bias originating from unbalanced groups, another bias is expected: when ranking categorical variables, random forest VIMs favour those variables that have more categories. This bias is commonly known, as seen in Strobl et al. (2007).

3.2 Data simulation and description of the HapMap dataset

To inspect the presence and extent of the bias of the chi-square teststatistic, different datasets are simulated in order to cover multiple scenarios: to begin with, the explanatory variables are 2-categorical, extending to 3-categorical variables in a second time. Dependencies between the explanatory variables and the outcome are also added with the intention to show that the permutation checks are still able to reveal the bias. This section will give an explanation of the simulation process for the different datasets as well as a description of the HapMap data.

The aim of this chapter is to inspect whether the chi-square statistic, and thus the chi-square test favours evenly distributed explanatory variables or not. To achieve this, the simulated data always contains four different explanatory variables with varying group proportions. To inspect the bias on two-categorical explanatory variables, four binary variables are created, X_i with $i = 1, ..., 4$, using the `rbinom` function. Note that if not specified otherwise, it always applies that $i = 1, ..., 4$. To simulate variables with different group proportions, the `prob` parameter of the function is adapted, using four different probabilities per dataset, one for each of the explanatory variables respectively:

$$\boldsymbol{X.proba} = (x_1.proba,\ x_2.proba,\ x_3.proba,\ x_4.proba). \tag{2}$$

This parameter defines the probability $x_i.proba = P(X_i = 1)$, allowing to create very unbalanced as well as balanced group proportions. The probabilities $x_i.proba$ are set in an increasing order such that X_1 has very unevenly distributed group counts while the group counts of X_4 are balanced:

$$\boldsymbol{X.proba} = (0.05,\ 0.15,\ 0.3,\ 0.5). \tag{3}$$

That way the explanatory variables follow a binomial distribution $X_i \sim Bin(n, x_i.proba)$ with $i = 1, ..., 4$ and n being the sample size of choice.

In a second time, the binary outcome variable Y is generated, using `rbinom` as well. This step allows to add dependencies between the explanatory variables X_i and Y through a logistic regression model. For this purpose, a vector $\boldsymbol{\beta} = (\beta_1,\ \beta_2,\ \beta_3,\ \beta_4)$ is defined, containing coefficients for said model. A predictor matrix is created by binding a column of "1" for the intercept with the predictor variables. Then the linear predictor is calculated by multiplying the predictor matrix with the coefficient vector $\boldsymbol{\beta}$:

$$\eta = 1 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4. \tag{4}$$

Then the logistic function `plogis` is applied to the linear predictor, looking as follows:

$$P(y = 1) = \frac{1}{1 + e^{-\eta}}. \tag{5}$$

Finaly Y is sampled from a binomial distribution using `rbinom`, with the probability of success being equal to the probability generated by the logistic function, meaning $Y \sim Bin(n, P(y = 1))$. This ensures that Y is a binary

variable and follows the regression coefficients specified in $\boldsymbol{\beta}$.
An exemplary result of this data-generating process can be seen in Table 1, showing the group count distribution of the outcome variable as well as the four explanatory variables for 1000 simulated observations. The latter have the following probabilities of success: $\boldsymbol{X.proba} = (0.05, 0.15, 0.3, 0.5)$ and for Y the probability of success is set to 0.5 as no dependencies are involved here.

Table 1: Absolute group counts for the outcome Y and the four explanatory variables X_1, ..., X_4 with probabilities $\boldsymbol{X.proba} = (0.05, 0.15, 0.3, 0.5)$

	Y	X_1	X_2	X_3	X_4
0	511	947	861	693	494
1	489	53	139	307	506

The table shows how the different probabilities of success for the X_i influence the count distributions of the groups. While X_1 for example is very unevenly distributed with only 53 out of 1000 observations belonging to group "1", for X_4 the groups are almost of equal size. This way to generate the data allows to compare the results of the chi-square test between explanatory variables with very different group distributions.

Next, the bias has to be inspected on three-categorical explanatory variables. The groups being "0", "1" and "2" now have the following probabilities:

$$P(x_i = 0) = 1 - (x_{i1}.proba + x_{i2}.proba), \tag{6}$$

$$P(x_i = 1) = x_{i2}.proba, \tag{7}$$

$$P(x_i = 2) = x_{i1}.proba. \tag{8}$$

Again variable-specific group probabilities $\boldsymbol{X.proba} = (x_1.proba$, $x_2.proba$, $x_3.proba$, $x_4.proba)$ are defined, with the difference that the $x_i.proba$ are no longer scalar values but ordered pairs of probabilities:

$$x_i.proba = (x_{i1}.proba, x_{i2}.proba). \tag{9}$$

While the simulation of the outcome Y stays the same, the data-generating process of the X_i is adapted for that purpose. As the four explanatory variables have now 3 instead of 2 categories, the X_i are not drawn from a binomial distribution anymore but are sampled using the R function `sample`. Depending on the choices for the probabilities, it is possible to have only 2 groups for one or several of the variables, being able to recreate the scenario of the paper by Boulesteix et al. (2012).

In total five different data and testing situations will be simulated, each of them consisting of 10000 simulations. The datasets always contain 1000 observations, and the $\boldsymbol{\beta}$-vector is set to $\boldsymbol{\beta} = (-0.5, 0.5, -0.5, 0.5)$. That way all the X_i all have an effect of the same magnitude on Y. The simulations differ in the settings for the exploratory variables and in the application of Yate's correction:

1. Simulation setting: two-categorical explanatory variables with $\boldsymbol{X.proba} = (0.05, 0.15, 0.3, 05)$ and Yate's correction is applied when performing the chi-square tests.

2. Simulation setting: two-categorical explanatory variables with $\boldsymbol{X.proba} = (0.05, 0.15, 0.3, 05)$ and Yate's correction is not applied when performing the chi-square tests.

3. Simulation setting: three-categorical explanatory variables with $\boldsymbol{X.proba} = ((0.05, 0.1), (0.1, 0.2), (0.15, 0.3), (0.33, 0.33))$.

4. Simulation setting: three-categorical explanatory variables with $\boldsymbol{X.proba} = ((0.003, 0.1), (0.1, 0.2), (0.15, 0.3), (0.33, 0.33))$ and Yate's correction is applied when performing the chi-square tests.

5. Simulation setting: three-categorical explanatory variables with $\boldsymbol{X.proba} = ((0.003, 0.1), (0.1, 0.2), (0.15, 0.3), (0.33, 0.33))$ and Yate's correction is not applied when performing the chi-square tests.

Please note that a different simulation approach has also been tested, where the datasets only contain one explanatory variable and the outcome is thus generated solely based on the dependency with that variable. This was done in order to make sure that the observed results in Section 3.4 are not influenced by the present simulation process where Y is generated based on all four explanatory variables simultaneously. The results using the datasets with only one explanatory variable were identical to the results presented here, the simulation process did not have any influence.

The application of the permutation checks to real-world data will be made on a subset of the HapMap dataset. The dataset consists of 350 individuals. 9670 different SNPs, which are the explanatory variables, are measured. Those polymorphisms are categorical variables, 366 having two categories and 9394 having three categories. The subject's ethnicity, a binary variable, is the outcome. The dataset is available along with the code of Boulesteix et al. (2012), from the website `http://www.ibe.med.uni-muenchen.de/organisation/mitarbeiter/020_professuren/boulesteix/ginibias`.

3.3 Methods and estimands

In the paper by Boulesteix et al. (2012), the bias in favour of evenly distributed explanatory variables is observed on the Gini VIM. The Gini impurity measure is used in decision trees to help the algorithm on how to split the data at each node to maximize the homogeneity of the resulting subsets. It measures the probability of incorrectly classifying a randomly chosen element, the goal is to minimize this probability. The Gini VIM is calculated based on the Gini impurity by considering the decrease in Gini impurity reached by a particular split. This decrease is weighted by the proportion of samples reaching the current node, the total importance of a variable is then calculated by summing up the impurity decreases across all decision nodes where that variable is used for splitting. That way both are linked as a decreasing Gini impurity measure leads to an increase of the Gini VIM. Furthermore, Grabmeier and Lambe (2007) state the equality between the Gini impurity measure and Pearson's chi-square statistic in binary splits. It is shown that a small Gini impurity implies a large chi-square statistic.

In this chapter, chi-square tests will be applied to the simulated data as well as to the HapMap data, with the goal of assessing the expected bias through permutation checks. For each simulated dataset a chi-square test of independence is performed between the outcome Y and each of the four explanatory variables X_i. The description of the chi-square test in this section is based on Grabmeier and Lambe (2007) and Ugoni and Walker (1995). For the testing purpose, contingency tables of Y and each of the X_i are created, four in total per dataset. Each of the contingency tables contains the observed frequencies O_{ij}, with $i = 1, 2$ being the row index for the two groups of the outcome variable and $j = 1, 2$ the column index for the groups of the explanatory variables. In the case that the explanatory variable consists of three groups, $j = 1, 2, 3$. Next, the expected frequencies are calculated for each cell of the contingency table:

$$E_{ij} = \frac{R_i * C_j}{N}. \tag{10}$$

R_i is the total count of observations in the ith row, $C_j$is the total count of observations in the jth column and N is the total count of observations in the entire contingency table. The expected frequency for a particular cell is thus the product of the row total count and the column total count divided by the grand total of observations. The test hypotheses for each distinct test are given by $H_0 : O_{ij} = E_{ij}$ vs $H_1 : O_{ij} \neq E_{ij}$ and Pearson's test statistic is calculated as follows (Grabmeier and Lambe 2007):

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}. \tag{11}$$

Under the null hypothesis, the χ^2 test statistic is χ^2-distributed with $(R-1)*(C-1)$ degrees of freedom, R being the number of rows and C the number of columns of the contingency table the test is performed on. The analysis is can be done with Yate's correction for continuity. This correction was introduced by the statistician Frank Yates in 1934 as a way to correct the chi-square test statistic for small sample sizes, especially when dealing with 2x2 contingency tables (Hitchcock 2009). Yate's correction involves adjusting the chi-square test statistic by subtracting 0.5 from the absolute difference between each observed frequency and its corresponding expected frequency before squaring and summing them up to calculate the test statistic (Hitchcock 2009):

$$\chi^2 = \sum_i \sum_j \frac{(\mid O_{ij} - E_{ij} \mid -0.5)^2}{E_{ij}}. \tag{12}$$

While Yates' correction is generally recommended for 2x2 contingency tables with small sample sizes, typically when any expected frequency is less than 5, it's important to note that it can also be applied to larger contingency tables, although its effectiveness diminishes as sample sizes increase. In fact, when performing a chi-square test in RStudio, Yate's correction is applied per default when dealing with 2x2 contingency tables.

The output of interest of the test procedure is the p-value, with the significance threshold being set to $\alpha = 0.05$. The chi-square test of each of the four explanatory variables against the outcome variable is done on 5 random permutations of each of the simulated datasets, before performing it on the original data. That way, any potential bias created by the test can already be detected before even conducting the final analysis and thus be taken into account for the interpretation of the results.

Low p-values of the chi-square test are obtained for high test statistics. Hence in the context of the paper by Boulesteix et al. (2012), a low p-value is comparable to a high Gini VIM, as a high Gini VIM implies a high test statistic. In the paper, the bias in favour of evenly distributed explanatory variables shows in a higher Gini VIM for those variables, compared to unevenly distributed variables. For the present thesis, the bias in favour of evenly distributed explanatory variables thus translates into lower p-values.

Analogous to the procedure on the simulated data, a potential bias of the test on the HapMap data is also investigated through five permutation checks. A chi-square test is then applied to each of the 9670 polymorphisms to test its independence with the outcome variable, the phenotype.

3.4 Simulation results

Results on two-categorical explanatory variables with Yate's correction

To start with, the expected bias is inspected with the help of permutation checks on simulated data where the explanatory variables are two-categorical. As explained in the previous section, datasets with four explanatory binary variables and one binary outcome variable are simulated. The probabilities $P(X_i = 1)$ for the explanatory variables are set to $\boldsymbol{X.proba} = (0.05, 0.15, 0.3, 0.5)$. That way, the group counts of the variable X_1 are very unevenly distributed as $P(X_1 = 1) = 5\%$, while the group counts of variable X_4 are evenly distributed, with both groups being equaly likely. The coefficients vector is set to $\boldsymbol{\beta} = (0.5, 0.5, -0.5, -0.5)$, such that the X_i all have an effect of the same magnitude on the outcome Y, as it is the case for every simulation. Each dataset contains 1000 observations of those variables. Next, the outcome is permuted five times for every simulated dataset and on each permuted dataset the chi-square test of independence between each of the explanatory variables and the outcome is performed with Yate's correction for continuity. The simulation, permutation and testing procedure is performed 10000 times and the results are shown in terms of p-values.

Figure 6 shows the p-values of the chi-square tests with Yate's correction, that are obtained after the permutation checks. In fact, the dependencies between the explanatory and the outcome Y that were introduced by the $\boldsymbol{\beta}$-vector are not showing in the boxplots, indicating that the permutation checks break those relationships as intended. The differences in the p-values between the variables that remain, are purely a result of the testing procedure. Although there are no big differences visible through the boxplots, it appears that the p-values of the chi-square tests between X_1 and Y are overall higher than those of the other variables, especially with the 75th percentile and the median demonstrating higher values. The same can be observed for X_2, but to a smaller extent. This difference in the p-value distribution could already be an indicator of the bias of the chi-square test in favour of variables with evenly distributed groups.

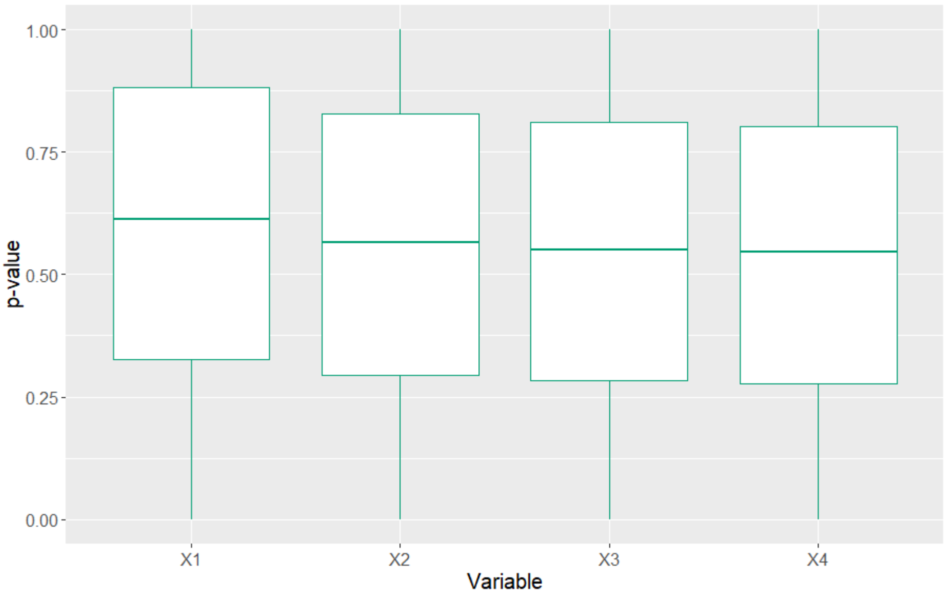

Figure 6: Boxplots of p-value distributions from chi-square tests of independence with Yate's correction between the two-categorical explanatory variables X_1, ..., X_4 and the outcome variable. 10000 datasets are simulated, tests done on five permutations of each dataset.

To confirm this, Table 2 displays two kinds of information: the second column indicates the percentages of how many times each of the explanatory variables had the lowest p-value among the four variables for a dataset. This means that for every simulated and permuted dataset it is checked which chi-square test between an explanatory variable and the outcome has the lowest p-value, the final percentage being denoted in the table. The third column shows the percentage of significant p-values for each of the explanatory variables, the threshold still being set to $\alpha = 0.05$.

Table 2: Percentages of occurence of lowest and significant p-values for the chi-square tests of the two-categorical explanatory variables with Yate's correction. 10000 datasets are simulated, tests done on five permutations of each dataset.

	Lowest p-value	Significant p-values
X_1	22.2%	3.3%
X_2	25.2%	3.9%
X_3	26.0%	4.3%
X_4	26.7%	4.2%

As each of the simulated datasets is permuted five times, there are 50000 datasets in total on which the chi-square tests are performed. Out of these 50000 datasets, X_1 has the lowest p-value in 22.2% of the cases. This percentage increases for each variable, with X_4 having the lowest p-value in 26.7% of the datasets. Since through the random permutations of the outcome variable there are no dependencies with the explanatory variables, those differences are a result of the expected bias of the chi-square test. In fact, X_4, which has evenly distributed groups, has more frequently the lowest p-value that the variables with less evenly distributed groups. The same pattern shows when considering the percentage of significant p-values for each explanatory variable: while 3.3% of the chi-square tests between X_1 and Y are significant, this percentage increases to 4.2% for X_4. It is also noticeable that the significance level of 5%, which would be reached with 2500 significant p-values, is not fully exploited, indicating that the chi-square test with Yate's correction for continuity is underpowered.

The permutation checks reveal a bias in favour of evenly distributed explanatory variables, that should be kept in mind when looking at the p-value distributions of the chi-square tests of the original data in Figure 7.

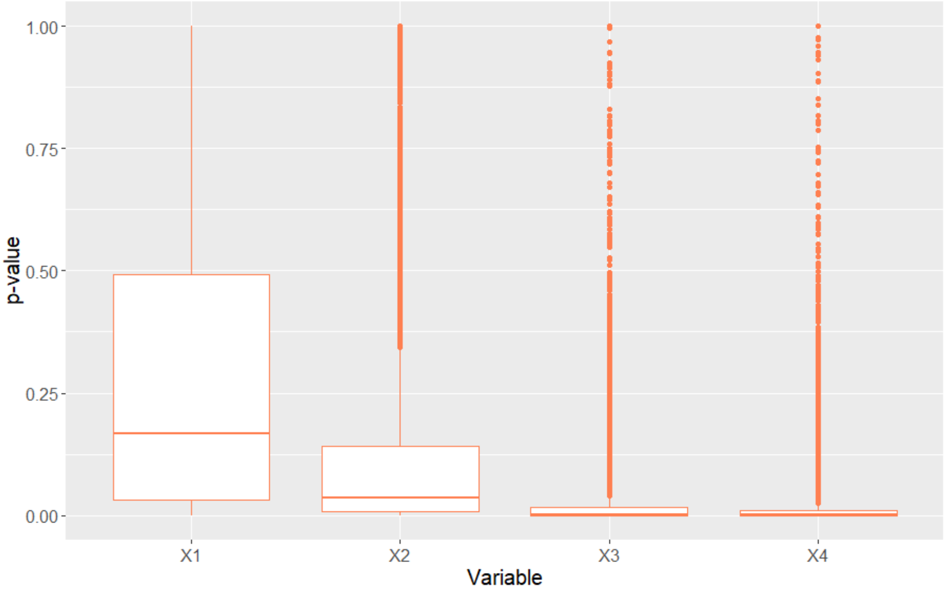

Figure 7: Boxplots of the p-value distribution of chi-square tests of independence with Yate's correction between the two-categorical explanatory variables X_1, ..., X_4 and the outcome variable. 10000 datasets are simulated, tests on original data.

Even though through the simulation process all four explanatory variables have the same influence on the outcome Y, the p-value distributions of the chi-square tests look very differently across the variables. Without knowing the ***beta***-coefficient setting, it would seem like the dependencies between Y and the explanatory variables increase from X_1 to X_4, with X_3 and X_4 having almost exclusively significant p-values for the chi-square test. The same picture shows when looking at Table 3, containing the same informations as Table 2:

Table 3: Percentages of occurence of lowest and significant p-values for the chi-square tests of the two-categorical explanatory variables with Yate's correction. 10000 datasets are simulated, tests done on original data.

	Lowest p-value	Significant p-values
X_1	3.8%	30.5%
X_2	7.5%	55.8%
X_3	39.5%	86.4%
X_4	49.2%	90.7%

The table confirms that out of the 10000 chi-square tests of independence between Y and X_4, 90.7% have significant p-values. As the previously performed permutation checks already predicted, there is a bias in favour of evenly distributed variables, showing through the decrease in the p-values the more evenly the variables are distributed. Nevertheless, the differences seen in Figure 7 cannot fully be explained by that bias of the chi-square test, as they are excessively large. There seems to be another bias that occurs only in the presence of an effect and is thus not caught by the permutations.

Results on two-categorical explanatory variables without Yate's correction

The identical workflow as before is now repeated, keeping the same ***X.proba*** and $\boldsymbol{\beta}$-vectors as well as the same number of simulations and permutations per simulation. The only difference is that now Yate's correction for continuity is not applied anymore.

Figure 8 displays the p-values of the chi-square tests on permuted data. Looking at the boxplots, the p-value distributions seem identical across all four explanatory variables, showing no apparent bias. Additionally to being identical, the p-values also appear uniformly distributed. Those results differ from the results observed in Figure 6, indicating that Yate's correction, which is not applied here anymore, is the source of the previously observed bias.

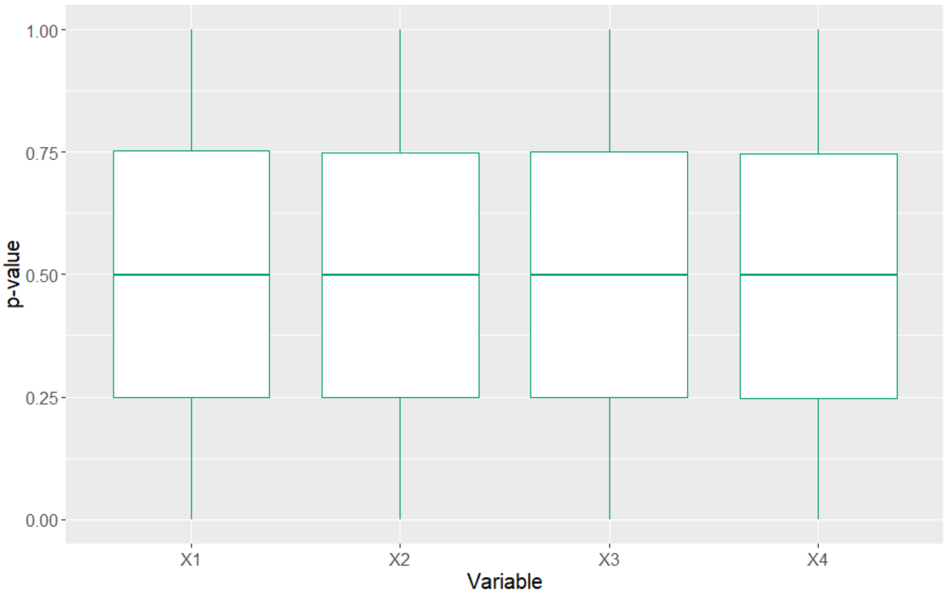

Figure 8: Boxplots of p-value distributions from chi-square tests of independence without Yate's correction between the two-categorical explanatory variables X_1, ..., X_4 and the outcome variable. 10000 datasets are simulated, tests done on five permutations of each dataset.

The significant p-value counts of Table 4 confirm this finding: contrary to before, there is no systematic increase in the counts of significant p-values from X_1 to X_4 when testing on the permuted data. In fact every explanatory variable has between 4.9% to 5.2% significant test results among the 50000 performed tests, confirming the absence of a bias in favour of evenly distributed variables in this scenario.

Even though according to the permutation checks no bias should be present anymore, Figure 9 still shows some clear differences between the p-value distributions of the explanatory variables on the original data, as seen in the previous simulation scenario including Yate's correction. In comparison, Table 4 indicates that the differences between the explanatory variables are slightly less pronounced than in Table 3, with the counts of significant and lowest p-values before the permutation checks having increased a little for the very unbalanced variables. Those differences are a result of the bias created by Yate's correction. The remaining differences among the explanatory variables are not captured by the permutation checks and remain unexplained.

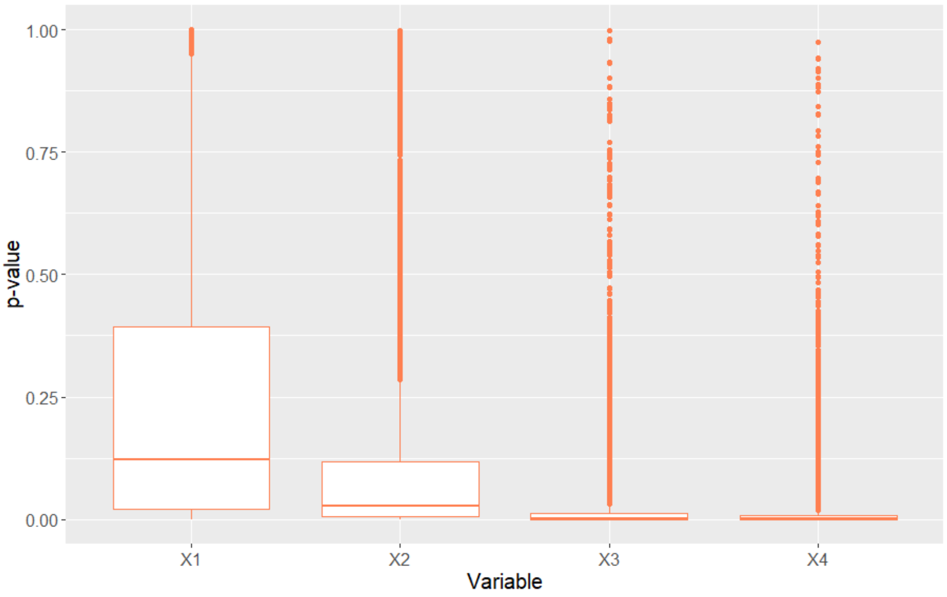

Figure 9: Boxplots of the p-value distribution of chi-square tests of independence without Yate's correction between the two-categorical explanatory variables X_1, ..., X_4 and the outcome variable. 10000 datasets are simulated, tests on original data.

Table 4: Percentages of occurence of lowest and significant p-values for the chi-square tests without Yate's correction of the two-categorical explanatory variables. 10000 datasets are simulated, tests done on original data and on five permutations of each dataset.

	Before Permutation		After Permutation	
	Lowest p-value	Significant p-values	Lowest p-value	Significant p-values
X_1	4.4%	35.9%	25.0%	5.1%
X_2	7.7%	60.3%	25.1%	5.0%
X_3	39.4%	88.1%	24.8%	5.2%
X_4	48.6%	91.8%	25.1%	4.9%

Through those two simulation scenarios, the permutations checks reveal a bias of Yate's correction for continuity with the chi-square test. When applied, the correction induces a bias favouring evenly distributed variables. As there are less than 5% significant results after the permutation checks, the bias does not seem to create an increased type I error and is thus not source of falsely identified dependencies between two variables. Problems are more likely to arise in cases where the p-values or test statistics are compared to each other and decisions or interpretations are made based on the differences, independently of the significance of the test.

Results on three-categorical explanatory variables

The inspection of the bias is now extended to three-categorical explanatory variables. The $\boldsymbol{\beta}$-vector remains identical, the group probabilities are set to $\boldsymbol{X.proba} = ((0.05, 0.1), (0.1, 0.2), (0.15, 0.3), (0.33, 0.33))$, ensuring that the explanatory variables always have three categories while keeping the pattern of increasingly more evenly distributed variables from X_1 to X_4. In the presence of only three-categorical explanatory variables, Yate's correction is never applied. This simulation is done to gain more certainty in the occurrence and source of the bias of the chi-square test before moving on to a scenario where two- and three-categorical explanatory variables are considered simultaneously.

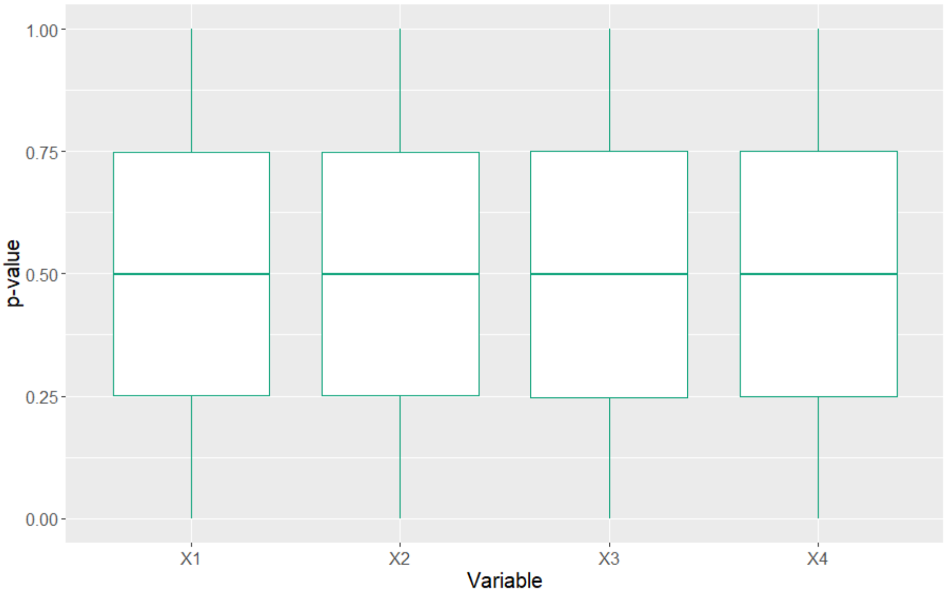

Figure 10: Boxplots of p-value distributions from chi-square tests of independence between the three-categorical explanatory variables X_1, ..., X_4 and the outcome variable. 10000 datasets are simulated, tests done on five permutations of each dataset.

The results of the permutation checks are considered first. The boxplots of the p-value distributions in Figure 10 as well as the percentages of significant p-values in Table 5 do not indicate any signs of biases. Since the two previous simulations revealed a bias coming from Yate's correction, this is an expected result as the correction is not used in the three-categorical case. Nevertheless the results on the original data in Figure 11 show again different p-value distributions for the explanatory variables, even though all of them have the same effect. This bias, also in favour of evenly distributed variables, could also be observed for the two-categorical explanatory variables and it is again not detected by the previously performed permutation checks.

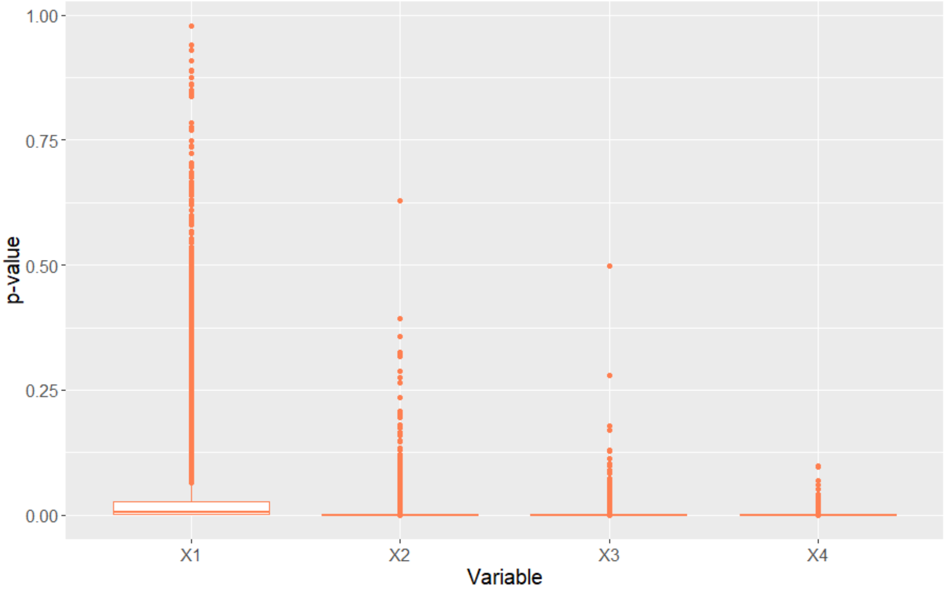

Figure 11: Boxplots of the p-value distribution of chi-square tests of independence between the three-categorical explanatory variables X_1, ..., X_4 and the outcome variable. 10000 datasets are simulated, tests on original data.

Table 5: Percentages of occurence of lowest and significant p-values for the chi-square tests of the three-categorical explanatory variables. 10000 datasets are simulated, tests done on original data and on five permutations of each dataset.

	Before Permutation		After Permutation	
	Lowest p-value	Significant p-values	Lowest p-value	Significant p-values
X_1	0.5%	83.0%	24.8%	5.0%
X_2	20.6%	99.2%	25.0%	5.1%
X_3	21.0%	99.6%	25.3%	5.0%
X_4	57.9%	100.0%	24.9%	5.1%

Results on two/three-categorical explanatory variables with Yate's correction

For the next simulations, the data is generated using $\boldsymbol{X.proba} = ((0.003, 0.1), (0.1, 0.2), (0.15, 0.3), (0.33, 0.33))$ as category probabilities for the explanatory variables. By setting the first pair of probabilities to $(0.003, 0.1)$, the probability for the category "2" $P(X_1 = 2)$ is so low that in approximately half of the simulated datasets, X_1 has only two instead of three categories while the other explanatory variables are three-categorical. This setting mimics the genetics scenario from Boulesteix et al. 2012, where some variants are very rare, leading to variables with different numbers of groups. Except for the generation of the explanatory variables, the remaining simulation process remains the same. For now, the chi-square tests are performed using Yate's correction for continuity, which is applied when X_1 is two-categorical. Again the permutation checks are performed before conducting the analysis on the original data, in order to identify potential errors and biases.

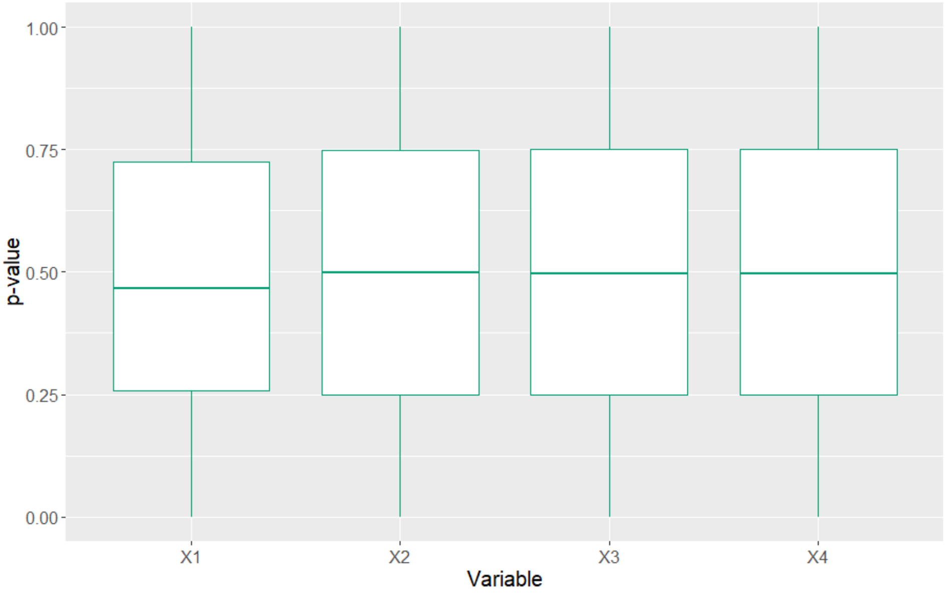

Figure 12: Boxplots of p-value distributions from chi-square tests of independence with Yate's correction between the two/three-categorical explanatory variables X_1, ..., X_4 and the outcome variable. 10000 datasets are simulated, tests done on five permutations of each datasets.

After the permutation checks the p-values are identically and uniformly distributed for X_2, X_3 and X_4 (Figure 12). Since Yate's correction only applies to two-categorical variables, those three explanatory variables are not affected by it and there is no bias among them. The p-values of x_1 on the other hand are partially affected by the correction, which is applied in the cases where it has only two categories. As a result, the distribution of the p-values of X_1 differs slighlty from the other variables. While in Figure 12 the median and the 75th percentile of the p-value distribution of X_1 are smaller than for the other explanatory variables, Table 6 reveals a bias in favour of X_2, X_3 and X_4: they all have similar amounts of significant p-values around the 5%-threshold while X_1 only has 3.9% of significant p-values. Whether this bias originates from Yate's correction or from the different number of categories, has to be inspected in the next step where testing is done without Yate's correction for continuity. Either way, the permutation checks indicate a bias in favour of X_2, X_3 and X_4 which needs to be kept in mind when conducting the analysis on the original data, seen in Figure 13.

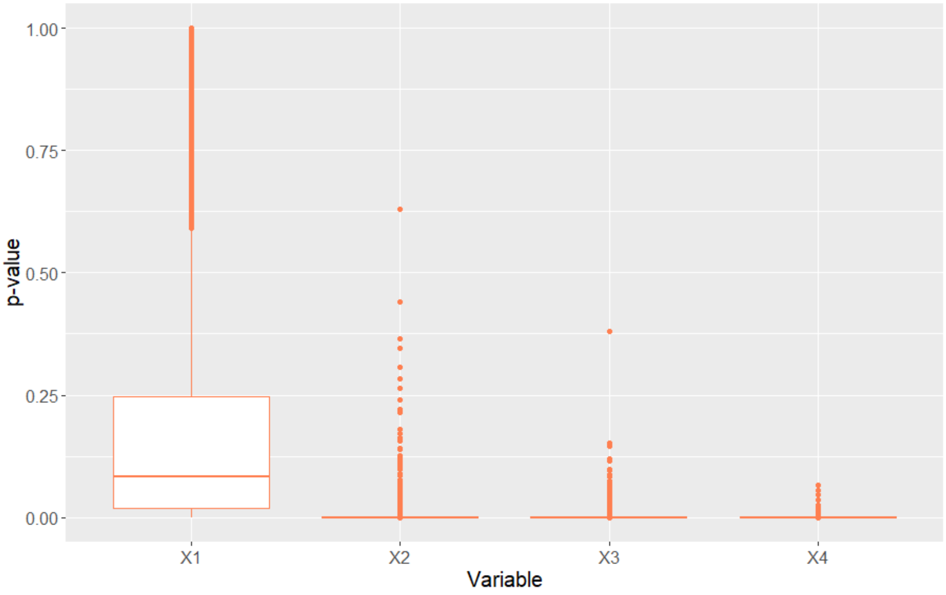

Figure 13: Boxplots of the p-value distribution of chi-square tests of independence with Yate's correction between the two/three-categorical explanatory variables X_1, ..., X_4 and the outcome variable. 10000 datasets are simulated, tests on original data.

Table 6: Percentages of occurence of lowest and significant p-values for the chi-square tests with Yate's correction of the two/three-categorical explanatory variables. 10000 datasets are simulated, tests done on original data and on five permutations of each dataset.

	Before Permutation		After Permutation	
	Lowest p-value	Significant p-values	Lowest p-value	Significant p-values
X_1	0.02%	39.6%	24.8%	3.9%
X_2	20.1%	99.4%	25.1%	5.1%
X_3	21.2%	99.7%	25.2%	5.0%
X_4	58.7%	100%	24.9%	5.1%

Figure 13 displays the p-value distributions of the chi-square tests on the original datasets. The dependencies between Y and each of the explanatory variables are equally strong, nevertheless the distributions look differently. As expected according to the permutation checks, there is a clear difference between the p-values of X_1 and the other explanatory variables, resulting from a bias. Even though the permutation checks predicted this sort of bias, its strength is still surprising as it is not caught by the permutations.

Results on two/three-categorical explanatory variables without Yate's correction

The simulation that is performed now is exactly the same as the previous one with the only difference, that Yate's correction for continuity is not applied anymore, not even in the cases where X_1 has only two instead of three categories. This will help indentify the source of the bias that has been previously revealed by the permutation checks.

Considering the results of the chi-square tests on the permuted data as well as the original data in Figure 14 and 15 respectively, the distributions of the p-values look almost identical to those that were obtained with Yate's correction. Small differences to the testing with Yate's correction can be found in Table 7. Especially after the permutations there is a small increase in the significant p-values of X_1 from 3.9% with Yate's correction to 4.0% significant p-values without Yate's correction, translating to 30 more tests being significant. The bias that is caught here by the permutation checks is the bias originating from Yate's correction, which increases the p-values of very unevenly distributed variables like X_1. However, the break that could already be observed previously with Yate's correction between X_1 and the other explanatory variables remains and is also detected by the permutations. This is the bias caused by the differing number of groups, in favour of variables with more groups.

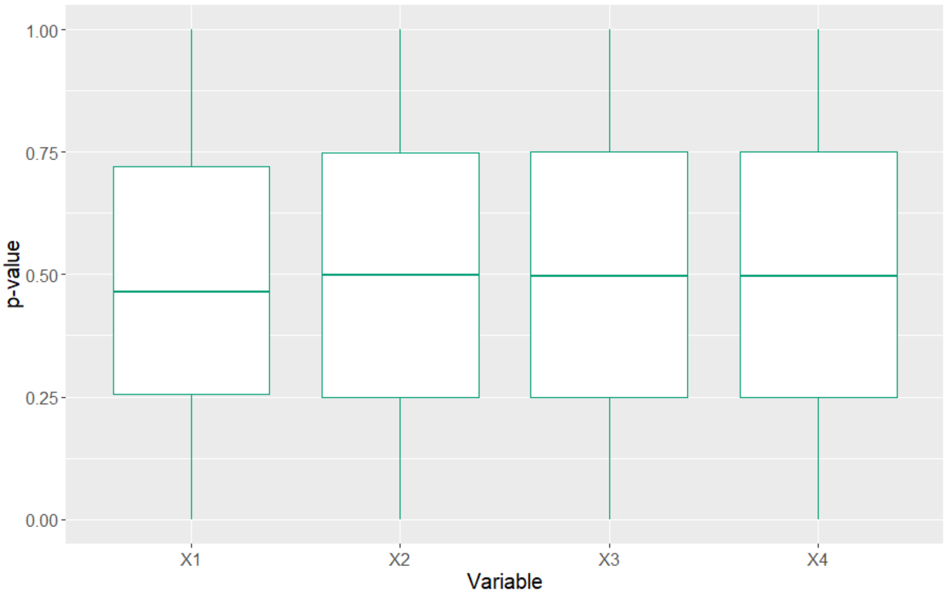

Figure 14: Boxplots of p-value distributions from chi-square tests of independence without Yate's correction between the two/three-categorical explanatory variables X_1, ..., X_4 and the outcome variable. 10000 datasets are simulated, tests done on five permutations of each dataset.

Table 7: Percentages of occurence of lowest and significant p-values for the chi-square tests without Yate's correction of the two/three-categorical explanatory variables. 10000 datasets are simulated, tests done on original data and on five permutations of each dataset.

	Before Permutation		After Permutation	
	Lowest p-value	Significant p-values	Lowest p-value	Significant p-values
X_1	0.02%	39.9%	24.9%	4.0%
X_2	20.1%	99.4%	25.1%	5.1%
X_3	21.2%	99.7%	25.1%	5.0%
X_4	58.7%	100.0%	24.9%	5.1%

Keeping in mind those biases revealed by the permutation checks, the results on the original data seen in Figure 15 should be interpreted with caution. When the underlying dependencies between the outcome and the explanatory variables are unknown, the results of the permutation checks can help identify whether the differences in p-values among the explanatory variables are due to a true effect or merely a result of bias.

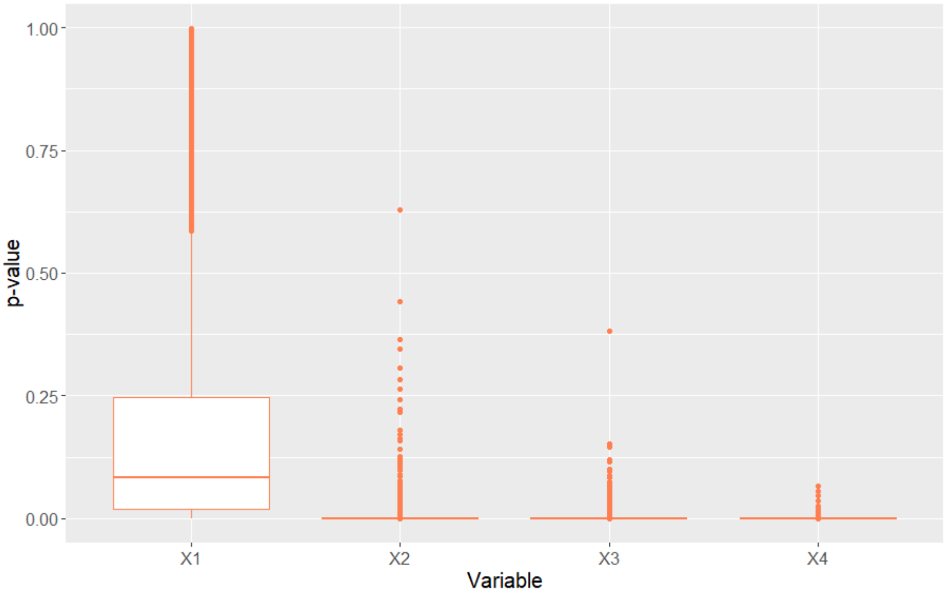

Figure 15: Boxplots of the p-value distribution of chi-square tests of independence without Yate's correction between the two/three-categorical explanatory variables X_1, ..., X_4 and the outcome variable. 10000 datasets are simulated, tests on original data

3.5 Real-world data application

In their paper, Boulesteix et al. 2012 apply the permutation checks to simulated data as well as to the HapMap dataset in order to assess the bias originating from the Gini VIM. Analogous to that, the permutation checks are used here to inspect whether Yate's correction for continuity creates biased results for chi-square tests on the HapMap dataset. For this purpose, the outcome variable, which is the phenotype, is randomly permuted five times and chi-square tests are done in order to obtain results that are free of any possible effect and thus reflect a potential bias. A chi-square test of independence is performed between each of the 9670 SNPs and the outcome for each permuted dataset. The testing is done once with and once without Yate's correction. Only after identifying potential biases thanks to the permutation checks, the chi-square tests are performed on the original data, thus following the recommended workflow. For visualization purposes and since there might be an expected difference in the results between two- and three-categorical SNPs, the obtained p-values are split according to the number of categories of the SNPs.

Results of the permutation checks

As mentioned, out of the 9670 SNPs of the HapMap dataset, 366 only have two instead of three categories. After randomly permuting the data five times and performing the chi-square tests, the results consist of 1830 p-values for two-categorical SNPs and 46520 p-values for three-categorical SNPs. Those results are obtained twice, once when testing is done with Yate's correction and once when testing is done without it. The p-values are plotted in Figures 16 and 24, with the relative frequencies of the lesser common category on the x-axis and the p-values on the y-axis. Figure 24 can be found in Appendix A, as it is almost identical to Figure 16. Looking at the p-value distribution of the two-categorical SNPs there does not seem to be a bias favouring evenly distributed SNPs as it was observed in Section 3.4, even with Yate's correction. This bias is absent here because of the small frequency range of the lesser frequent category on the two-categorical SNPs. Overall there are visually hardly any apparent differences in the p-value distribution between testing on the permuted data with and without Yate's correction.

However the percentage of significant p-values reveals some differences (Table 8): When Yate's correction is applied, only 3.7% of the chi-square tests on two-categorical SNPs are significant, as opposed to 5.6% significant p-values for three-categorical SNPs. When testing without Yate's correction the proportions of significant p-values become almost identical, with 5.5% for two-categorical and 5.6% for three-categorical SNPs, meaning there is also no bias favouring SNPs with more categories apparent. Because of Yate's correction there are 1.8% fewer tests identified as significant and the α-level is not fully exploited.

Table 8: Percentage of significant p-values of the chi-square test from five permutation checks of the HapMap data.

	With Yate's cor.	Without Yate's cor.
2-cat. SNPs	3.7%	5.5%
3-cat. SNPs	5.6%	5.6%

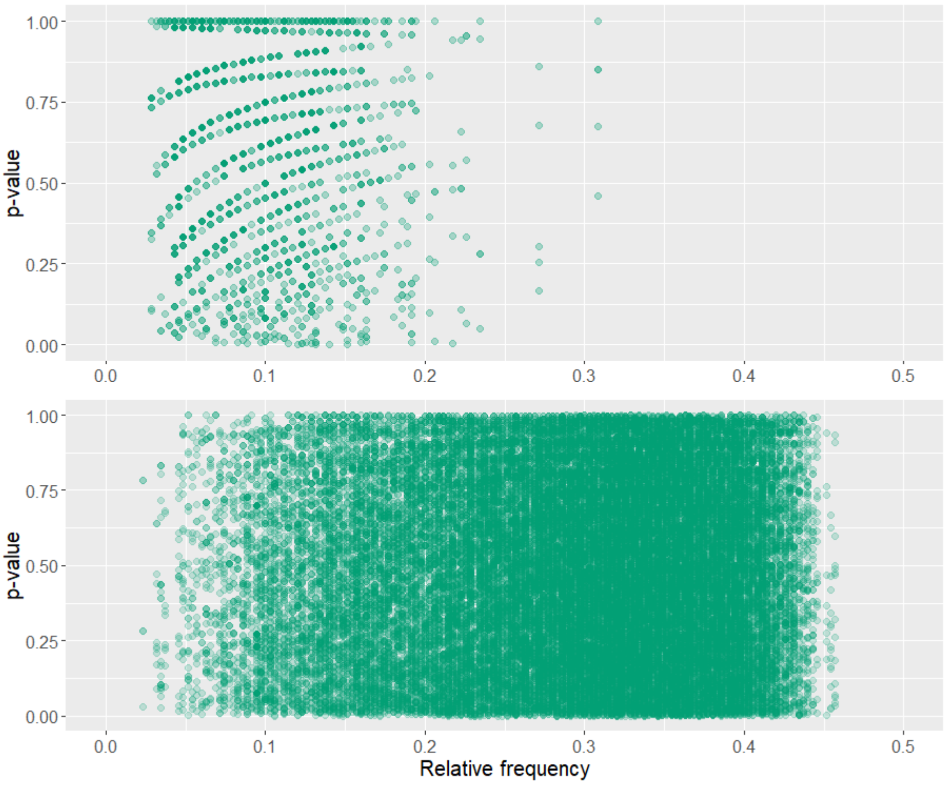

Figure 16: P-values of the chi-square tests on five permutations of the HapMap dataset with Yate's correction applied, split in SNPs with two categories (top) and SNPs with three categories (bottom)

Overall, the permutation checks revealed a bias introduced by Yate's correction, leading to almost 2% less significant test results for two-categorical SNPs. This finding can now have an impact on the analysis of the original data: the chi-square tests can either be performed without Yate's correction to completely avoid that bias or the correction can be used and the results are then interpreted with more caution. For comparison, both are done in the next part.

Results of the analysis on the original data

Now the originally planned analysis, namely chi-square tests of indenpendence between each SNP and the outcome, is performed on the HapMap data. The tests are be performed once with and once without Yate's correction, acknowledging the previously found bias.

Figure 17 and 18 display the distributions of the p-values over the relative frequency range, with and without Yate's correction respectively. By looking at the plots, it already shows that the majority of the p-values appear to be below the significance threshold of $\alpha = 5\%$. Especially in comparison with the results on the permuted data, the p-values are now much lower, showing once again that the permutations break existing effects. By the first look, not many differences between the plots with and without Yate's appear. Particularly the plots for the three-categorical SNPs are identical since Yate's correction is not applied at all there. Regarding the two-categorical SNPs, most of the p-values seem to decrease without Yate's correction (Figure 18). This trend also shows in the median of the p-values: with Yate's correction the median p-value of the two-categorical SNPs is 0.005, without Yate's correction it is 0.003. The percentages of significant p-values in Table 9 reflect this bias as well, with 71.9% against 74.6% significant p-values with and without the correction, respectively.

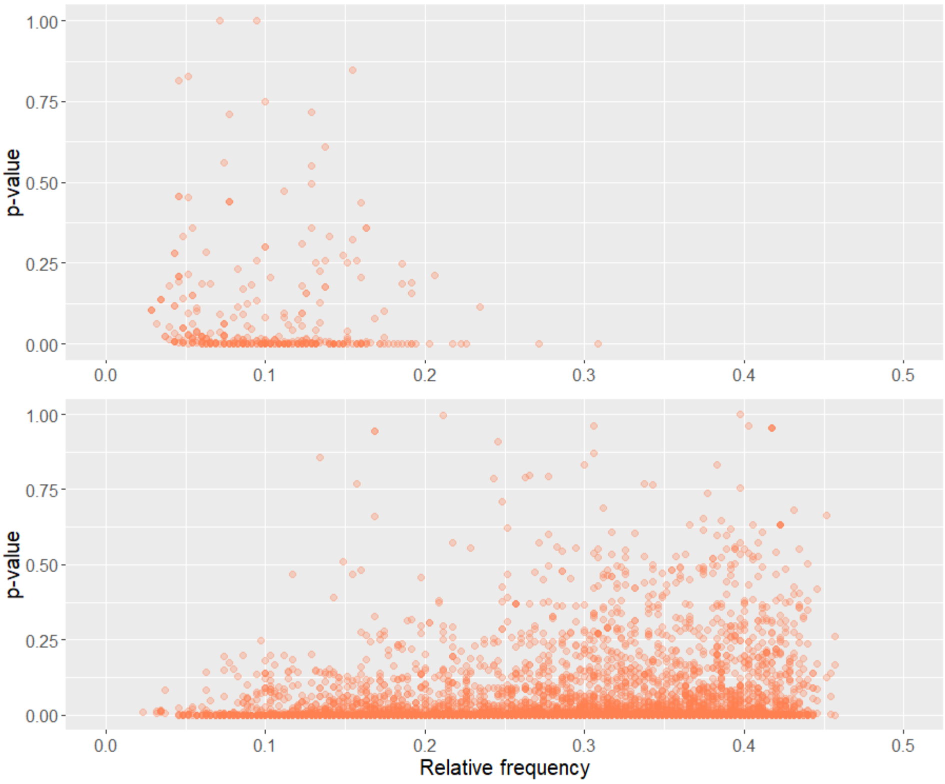

Figure 17: P-values of the chi-square tests of the original HapMap dataset with Yate's correction applied, split in SNPs with two categories (top) and SNPs with three categories (bottom)

According to the findings of the permutation checks, the results obtained without Yate's correction can be considered unbiased and the SNPs presenting a significant test result can be interpreted as not being independent with the phenotype. On the other hand, if the analysis is conducted using Yate's correction, the interpretation of the results should be done with caution, knowing that dependencies between some of the two-categorical SNPs and the outcome might not be recognized by the test.

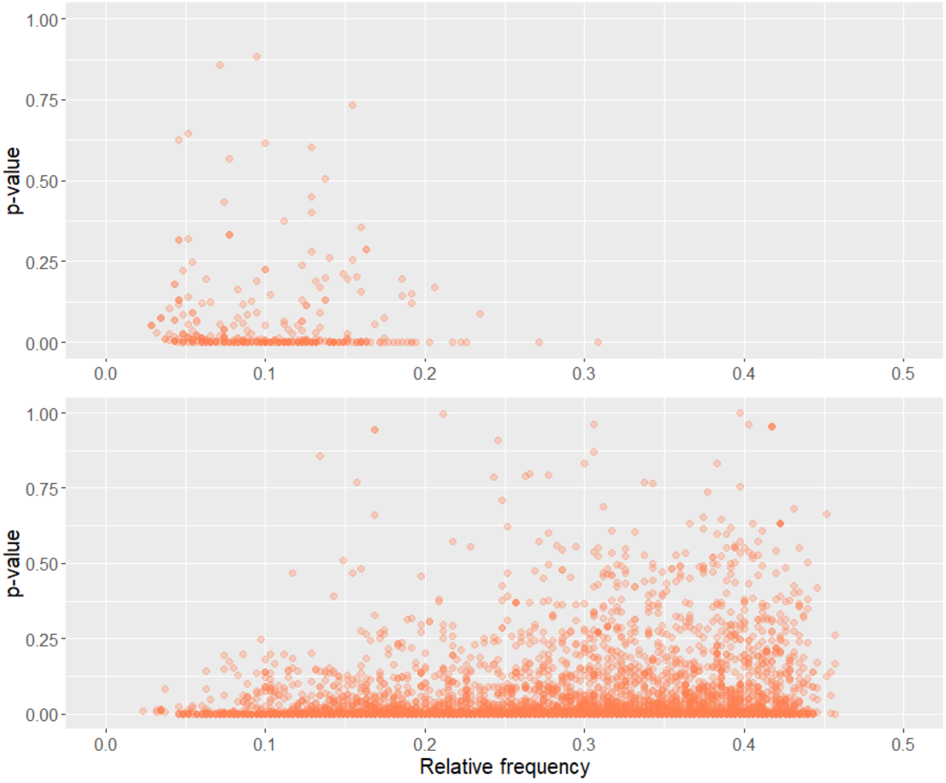

Figure 18: P-values of the chi-square tests of the original HapMap dataset without Yate's correction applied, split in SNPs with two categories (top) and SNPs with three categories (bottom)

Table 9: Percentage of significant p-values of the chi-square test on the original HapMap data.

	With Yate's cor.	Without Yate's cor.
2-cat. SNPs	71.9%	74.6%
3-cat. SNPs	86.4%	86.4%

4 Extension of the permutation checks to the Cochran-Mantel-Haenszel test

Now that the bias of the chi-square test with Yate's correction in favour of evenly distributed explanatory variables is shown through permutation checks, the next step is to inspect another statistical test to see if the bias extends to it: the Cochran-Mantel-Haenszel test (CMH test), which is closely related to the chi-square test. Developed by three statisticians in the mid-20th century, the test is used to assess the association between two categorical variables while controlling for a third confounding variable. The theory of the CMH test presented in Sections 4.1 and 4.2 is based on Agresti (2012).

4.1 Aim and simulation process

Again the bias is assessed through permutation checks using simulated data. The start of the simulation process is identical to the one presented in Section 3.2 where data for the chi-square test is simulated, as the required data format for the Mantel-Haenszel test is similar. Each simulated dataset consists of a binary outcome variable Y and four binary explanatory variables X_i, $i = 1, ..., 4$. Again for each of the X_i a different probability of success can be declared and the data-generating function offers the possibility to create dependencies between the X_i and the outcome Y. Next, indices are sampled randomly without replacement from the row numbers of the dataset. The resulting vector of indices is then cut in three parts and the dataset is split accordingly, in order to mimic a three-categorical confounder variable. The dataset is then converted into contingency tables between the X_i and the outcome Y, resulting in a total of 12 tables for one original simulated dataset: each of the four X_i has a separate contingency table for each level of the confouder, as seen in Table 10.

Table 10: Contingengy table for one X_i and one level $k = 1, 2, 3$ of the confounding variable

	$X_i = 0$	$X_i = 1$	Row total
$Y = 0$	A_k	B_k	N_{1k}
$Y = 1$	C_k	D_k	N_{2k}
Column total	M_{1k}	M_{2k}	T_k

Two different data and testing situations are simulated, both consisting of 10000 simulations. The datasets always contain 1000 observations. The probability vector is set to $\boldsymbol{X.proba} = (0.05, 0.15, 0.3, 05)$ and the $\boldsymbol{\beta}$-vector to $\boldsymbol{\beta} = (-0.5, 0.5, -0.5, 0.5)$. That way all the X_i all have an effect of the same magnitude on Y, identically to the simulations for the chi-square tests. The difference between the simulation is again Yate's correction for continuity, as the expected bias originates from it.

4.2 Methods and estimands

The method of interest in this section is the CMH test, commonly used in epidemiology and biostatistics to analyze the relationship between two variables while accounting for the influence of a third variable. The null hypothesis assumes the conditional independence of the two variables of interest. Given Table 10 and $k = 1, 2, 3$, the test statistic looks as follows (Agresti 2012):

$$CMH = \frac{[\sum_k (A_k - \mathbb{E}(A_k)]^2}{\sum_k var(A_k)} \tag{13}$$

with

$$\mathbb{E}(A_k) = \frac{N_{1k} * M_{1k}}{T_k} \text{ and } var(A_k) = \frac{N_{1k} * N_{2k} * M_{1k} * M_{2k}}{T_k^2(T_k - 1)}. \tag{14}$$

The significance threshold is set to $alpha = 5\%$, the p-value is the output of interest with which the potential bias in favour of evenly distributed explanatory variables is assessed. If a bias is present, it is expected to show in the same way as for the chi-square test, with the p-values being lower the more evenly the variables are distributed. Yate's correction for continuity can also be applied to the CMH test identically as for the chi-square test, the bias is inspected with and without it.

4.3 Simulation results

Results of the permutation checks

First, the permutation checks are done to predict biases that could have an influence on the results of the analysis of the original data. Since a bias coming from Yate's correction for continuity is expected when performing the CMH test, five permutation checks are done on every simulated data set with and without the correction.

Figures 19 and 20 display the p-value distributions resulting from this procedure, indicating again differing results between testing with and without Yate's correction. In fact, the p-values are larger with Yates' correction for all four explanatory variables, with the 25th and 75th percentiles and the median lying above the values expected from a uniform distribution.

Besides general differences to the p-value distribution of the tests without Yate's correction, there are also differences within the p-value distributions of the explanatory variables with Yate's correction. Especially the p-value distribution of X_1, which is the most unevenly distributed explanatory variable, differs from the other ones. Compared to X_2, X_3 and X_4, the p-value distribution of X_1 is shifted towards higher values with a median of 0.61. Opposed to that, the p-value distributions of CMH tests done without Yate's correction seem uniformly distributed and no bias is apparent (Figure 20).

Table 11 confirms those results: without Yate's correction applied, all four explanatory variables have the lowest p-value in approximately a quarter of the cases and 5% of their p-values are significant. On the other hand, when Yate's correction is used while testing, a bias in favour of evenly distributed variables becomes apparent and the test is again underpowered. Overall the findings are similar to those on the chi-square test, leading again to interpret the results on the original data with caution.

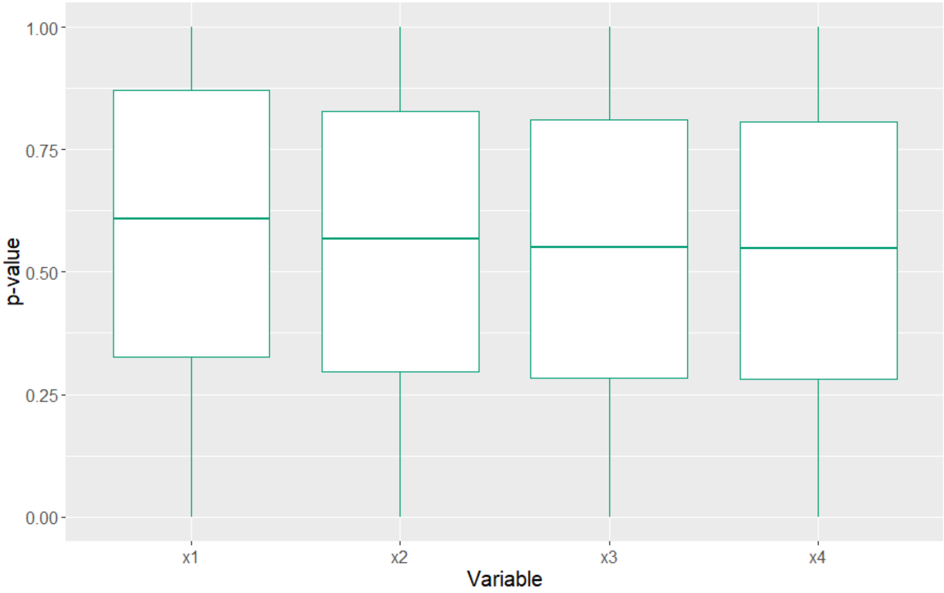

Figure 19: Boxplots of p-value distributions from CMH tests with Yate's correction between the two-categorical explanatory variables X_1, ..., X_4 and the outcome variable. 10000 datasets are simulated, tests done on five permutations of each datasets.

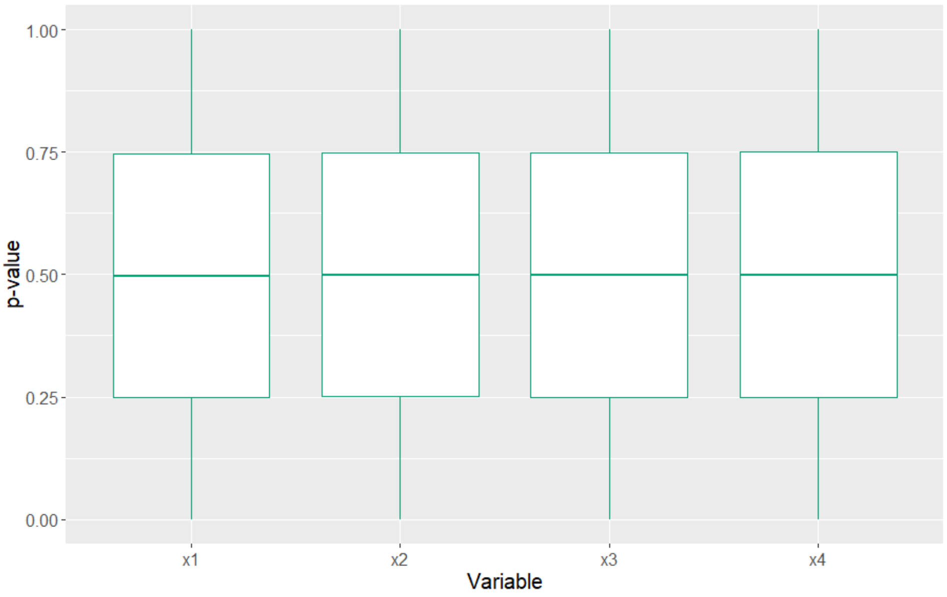

Figure 20: Boxplots of p-value distributions from CMH tests without Yate's correction between the two-categorical explanatory variables X_1, ..., X_4 and the outcome variable. 10000 datasets are simulated, tests done on five permutations of each datasets.

Table 11: Percentages of occurence of lowest and significant p-values for the CMH tests with and withourt Yate's correction of the two-categorical explanatory variables. 10000 datasets are simulated, tests done on five permutations of each dataset.

	With Yate's cor.		Without Yate's cor.	
	Lowest p-value	Significant p-values	Lowest p-value	Significant p-values
X_1	22.4%	3.2%	25.3%	4.9%
X_2	24.9%	3.9%	24.8%	5.0%
X_3	26.3%	4.1%	25.0%	5.0%
X_4	26.5%	4.1%	24.9%	5.0%

Results of the analysis on the original datasets

The permutation checks found a bias of the CMH test with Yate's correction in favour of evenly distributed variables, allowing a more correct assessment of the analysis results on the original data. The interpretation of the p-values obtained with Yate's correction, seen in Figure 21, needs to be done with caution. Thanks to the permutation checks it is known that the p-value distributions are biased, with especially the p-values of X_1 being higher than they should be. When working with real data where the underlying effects of the explanatory variables on the outcome are unknown, the permutation checks would allow to know that potential dependencies between some of the X_1 and the outcome Y might be overlooked and missed by the CMH test.

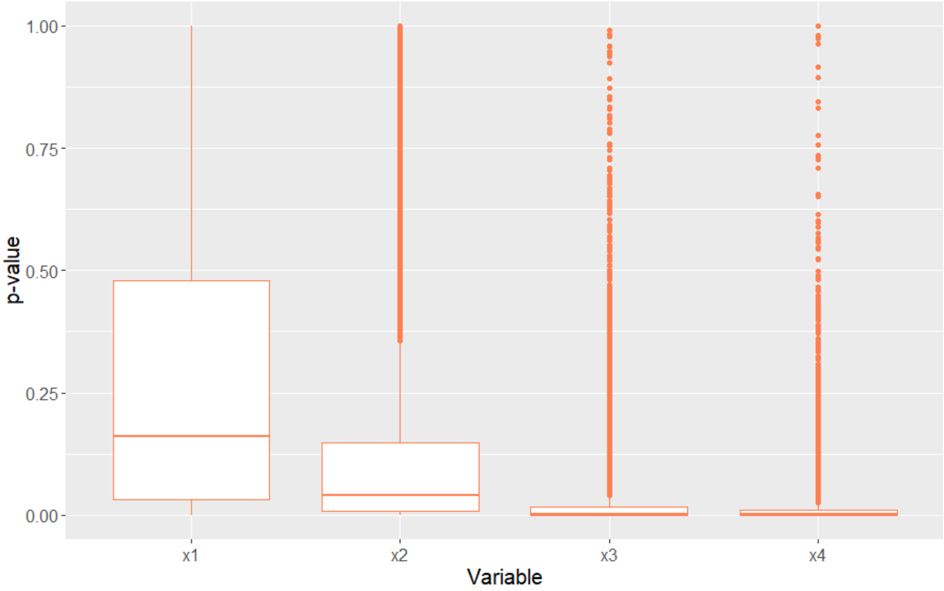

Figure 21: Boxplots of p-value distributions from CMH tests with Yate's correction between the two-categorical explanatory variables X_1, ..., X_4 and the outcome variable. 10000 datasets are simulated, tests on original data.

Concerning the CMH test results without Yate's correction on the other hand, the permutation checks didn't find any bias or anomaly. In the present simulation setting, it is known that all four explanatory variables have an effect of the same magnitude on the outcome. Despite this, Figure 22 displays different p-value distributions for each of the X_i. These biased distributions are not predicted by the permutation checks. If the true effects were unknown and the interpretation assumed no biases based on the permutation check results, the interpretation would be biased.

In Table 12 some of the differences between unevenly and evenly distributed explanatory variables can be explained by Yate's correction, nevertheless the main part of the differences remains unexplained. Taking for example the percentages of significant p-values , X_1 has almost 6% more significant p-values when Yate's correction is not used with the CMH test. However, X_2 has 22.5% more significant p-values than X_1, increasing up to 55.5%

more significant p-values for X_4, all of that in a simulation situation without Yate's correction where the permutation checks couldn't identify any biases.

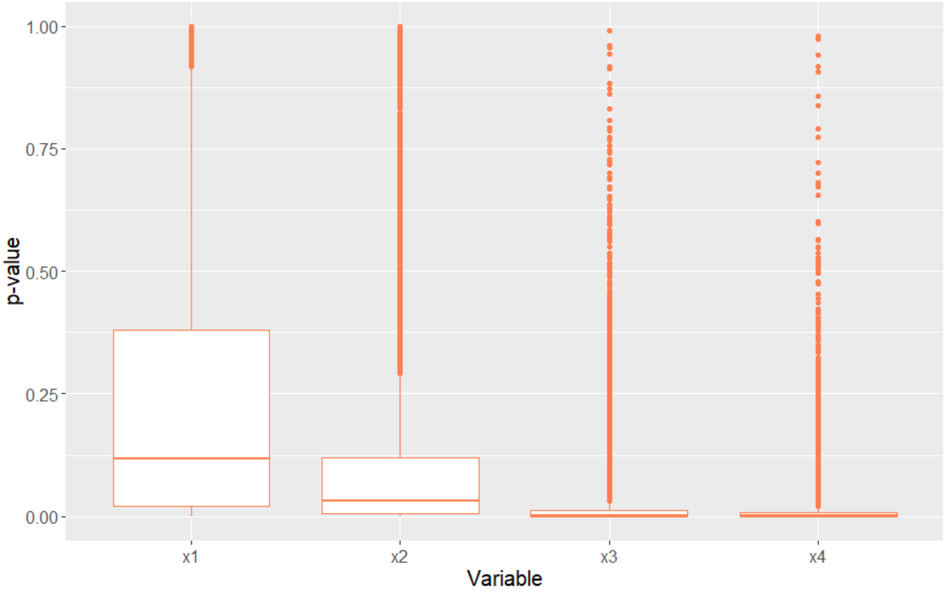

Figure 22: Boxplots of p-value distributions from CMH tests without Yate's correction between the two-categorical explanatory variables X_1, ..., X_4 and the outcome variable. 10000 datasets are simulated, tests done on original data.

Table 12: Percentages of occurence of lowest and significant p-values for the CMH tests with and without Yate's correction of the two-categorical explanatory variables. 10000 datasets are simulated, tests done on original data.

	With Yate's cor.		Without Yate's cor.	
	Lowest p-value	Significant p-values	Lowest p-value	Significant p-values
X_1	4.1%	30.6%	4.7%	36.1%
X_2	7.3%	54.3%	7.5%	58.6%
X_3	40.0%	86.8%	39.8%	88.4%
X_4	48.6%	90.3%	48.0%	91.6%

Analogous to the simulation results on the chi-square test, the permutation checks help finding a bias of the CMH test with Yate's correction for continuity in favour of evenly distributed variables. On the other hand, they fail to uncover a bias of the CMH test in favour of evenly distributed variables which is present independently of Yate's correction. This bias seems to occur only in the presence of an effect between the X_i and the outcome,explaining why it cannot be revealed by the permutation checks, since they break all possible dependencies.

5 Practical considerations for the permutation checks

5.1 Distribution of the outcome variable

In the simulation settings that have been considered until now, the outcome variables always have been fairly evenly distributed. In the HapMap data as well, both groups of the phenotype were almost equally large. Now let's consider the scenario of unevenly distributed outcome variables to inspect wether the permutation checks are still able to capture biases in those cases. This is will be tested using simulation and testing scenarios whom the bias to expect is known: the bias of the chi-square test with Yate's correction and two-categorical explanatory variables and the bias of the CMH test with Yate's correction in favour of evenly distributed variables. The simulation, permutation and testing workflow, which is repeated 10000 times, remains exactly the same with $\boldsymbol{X.proba} = (0.05, 0.15, 0.3, 0.5)$ and $\boldsymbol{\beta} = (-0.5, 0.5, -0.5, 0.5)$. The only change happens in the simulation of the outcome variable Y where the intercept is set to 3 instead of 1. This results in a probability of occurence for the lesser common category of the outcome variable between 4-7%, thus creating a unevenly distributed Y variable. Note that the explanatory variables are not affected by this change in the data-generating process.

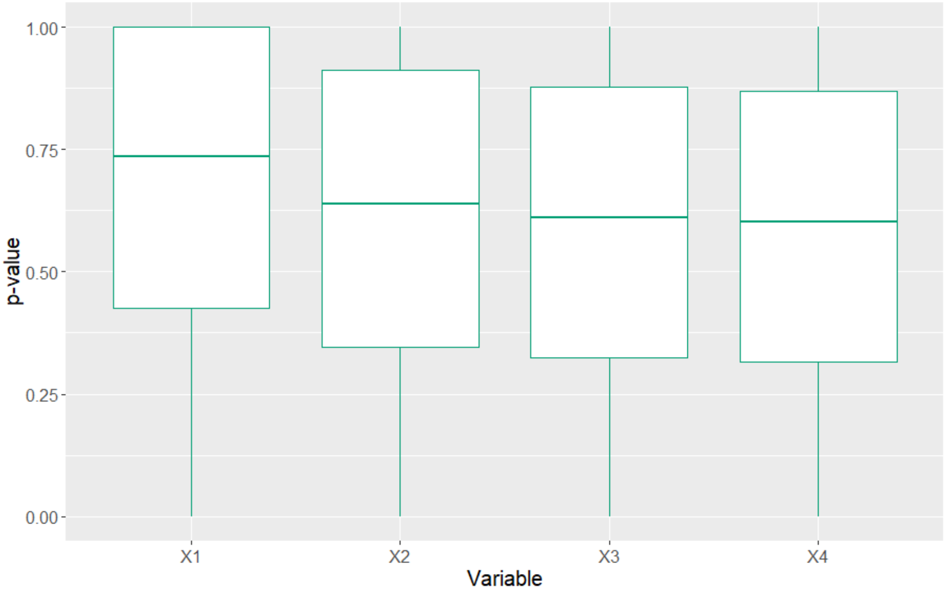

Figure 23: Boxplots of p-value distributions from chi-square tests of independence between the two-categorical explanatory variables X_1, ..., X_4 and the outcome variable, which is unevenly distributed. 10000 datasets are simulated, tests done on five permutations of each datasets.

Figure 23 displays the p-value distributions of the chi-square tests on permuted data with Yate's correction and an unevenly distributed outcome variable Y. The pattern of decreasing p-value distribution ressembles the one observed in Section 3.4, Figure 6, where the bias of the chi-square test with Yate's correction and an evenly distributed outcome could be identified. Again, the p-values increase the more unevenly the explanatory variables are distributed, confirmed by the results in Table 13. The percentage of significant p-values for the chi-square test of independence between each of the X_i and the outcome increases the more evenly the X_i are distributed.

Table 13: Percentages of occurence of lowest and significant p-values for the chi-square tests with Yate's correction of the two-categorical explanatory variables. 10000 datasets are simulated, tests done on five permutations of each dataset. The outcome Y is unevenly distributed.

	Lowest p-value	Significant p-values
X_1	18.8%	2.2%
X_2	25.3%	2.9%
X_3	27.7%	3.3%
X_4	28.2%	3.4%

Similar results can be observed when performing the CMH test (Figure 25 and Table 13 of Appendix A), where the permutations also reveal the bias introduced by Yate's correction, with Y being very unevenly distributed.

In the case of biases of Yate's correction with the chi-square and the CMH test, the permutation checks work regardless of how the binary outcome variable is distributed. Whether this finding also holds when applying permutation checks to other statistical analyses has not been investigated in this thesis; considerations should be made for each individual application.

5.2 Number of permutations

There is no clear instruction or guideline on the number of permutation checks needed in order to uncover potential errors or anomalies in the statistical analysis. While five random permutations of the outcome variable are made in Wünsch (2022), in the paper by Boulesteix et al. (2012) the outcome of interest is permuted ten times to assess the bias of the chi-square test on the HapMap dataset. Similarly, in this thesis ten permutations checks are done in the introductory examples on the t-tests and five are done for all other analyses. This choice is made based on the sample sizes of the individual simulations and datasets: on one hand the permutation checks should reach total sample sizes big enough in order to discover potential problems, on the other hand performing many permutation checks can quickly become computationally intensive and time consuming.

Let's consider the simulation scenario from Section 3.4 “Results on two-categorical explanatory variables without Yate's correction”. Here the permutation checks show no signs of a bias, as Yate's correction is not applied. If only one permutation check is done instead of five, the absence of a bias is not so clear, with a slightly increasing number of significant p-values from X_1 to X_4 (Table 14). This could be due to the implemented dependencies between the X_i and Y, a single permutation check does not seem to be enough to erase it out of the results. Moreover, with a single permutation check, 10000 chi-square tests are done while 50000 are done with five permutation checks, leading to more exact results. The corresponding plot for the p-value distributions can be found in Appendix A, Figure 26.

Table 14: Percentages of occurence of lowest and significant p-values for the chi-square tests without Yate's correction of the two-categorical explanatory variables. 10000 datasets are simulated, tests done on one permutation of the dataset.

	Lowest p-value	Significant p-values
X_1	24.7%	4.8%
X_2	24.9%	5.0%
X_3	24.6%	5.0%
X_4	25.9%	5.5%

The choice of the number of permutation checks can also depend on the statistical analysis to be inspected. For example in Wünsch (2022), the permutation checks helped finding genes that were consistently identified as differentially expressed across all random permutations. Several permutation checks are necessary to recognise such a pattern, as a systematic selection cannot be recognised in one or two random permutations.

6 Conclusion and outlook

In this thesis, the application possibilities and the role of permutations in detecting errors and anomalies in statistical analysis have been explored. Starting with simple scenarios, random permutations allowed for the detection of outliers, which impacted the results of t-tests. Based on the work of Boulesteix et al. (2012), the next goal was to use permutation checks to inspect a suspected bias in chi-square tests. Using simulated data sets, the permutations revealed a bias in the chi-square test along with Yates's correction for continuity, which is applied to two-categorical variables. The bias causes the p-values of evenly distributed variables to be lower and more often significant than the p-values of unevenly distributed variables. Furthermore, the chi-square test becomes underpowered when applying Yates's correction. The same simulations were conducted on three- and two/three-categorical data. A bias was found by the permutation checks only with two-categorical variables when Yates's correction was applied. Despite the permutations discovering this bias, it was apparent that another bias influenced the test results. This bias only appeared in the original, non-permuted data, where the p-values decreased the more evenly the variables were distributed, despite all variables having the same effect on the outcome. Overall, this thesis partly confirms the statement made by Boulesteix et al. (2012), showing a bias in the chi-square test with Yates's correction on two-categorical variables. A bias in another scenario could not be found by the permutations. In the application on the real-world HapMap data, the permutation checks performed before the analysis on the original data allowed for a more accurate assessment and evaluation of the test results by revealing again the bias of the chi-square test.

Subsequently, the permutation checks were applied to a method closely related to the chi-square test: the Cochran-Mantel-Haenszel (CMH) test. It was expected that the bias of Yates's correction in favor of evenly distributed variables would also be observed for the CMH test. Indeed, the permutation checks revealed this bias, which can lead to incorrect analysis results and interpretations if disregarded.

In the present simulation scenarios on the chi-square and CMH tests, the permutation checks were also able to identify the bias even when the outcome variable was very unevenly distributed. Furthermore, before performing the checks, consideration should be given to the number of conducted random permutations. Enough permutation checks should be performed to detect errors or anomalies while keeping in mind that it can become computationally intensive depending on the data and the chosen analysis.

Finally, permutations can be a simple and helpful tool to use before conducting statistical analysis on the original data. They can be applied when it is not known which error or bias to expect, allowing for a more precise interpretation of the results. Nevertheless, it should be noted that they fail to detect biases that occur only in the presence of a statistical effect or dependency, since random permutations break those. In this thesis, permutation checks were only applied to binary outcome variables. Future work could build on this by inspecting how they could be applied to outcome variables with more than two categories or even continuous outcomes, in order to expand the application field of permutations.

Bibliography

Agresti, Alan (2012). *Categorical data analysis.* Vol. 792. John Wiley & Sons.

Auguie, Baptiste, Anton Antonov, and Maintainer Baptiste Auguie (2017). "Package 'gridExtra'". In: *Miscellaneous functions for "grid" graphics.* URL: https://cran.r-project.org/web/packages/gridExtra/gridExtra.pdf.

Boulesteix, Anne-Laure et al. (2012). "Random forest Gini importance favours SNPs with large minor allele frequency: impact, sources and recommendations". In: *Briefings in Bioinformatics* 13.3, pp. 292–304.

Davison, Anthony Christopher and David Victor Hinkley (1997). *Bootstrap methods and their application.* 1. Cambridge university press.

Fodor, Anthony A, Timothy L Tickle, and Christine Richardson (2007). "Towards the uniform distribution of null P values on Affymetrix microarrays". In: *Genome biology* 8, pp. 1–16.

Gewers, Felipe L et al. (2021). "Principal component analysis: A natural approach to data exploration". In: *ACM Computing Surveys (CSUR)* 54.4, pp. 1–34.

Grabmeier, Johannes L and Larry A Lambe (2007). "Decision trees for binary classification variables grow equally with the Gini impurity measure and Pearson's chi-square test". In: *International journal of business intelligence and data mining* 2.2, pp. 213–226.

Hitchcock, David B (2009). "Yates and contingency tables: 75 years later". In: *Electronic Journal for History of Probability and Statistics* 5.2, pp. 1–14.

Ioannidis, John PA (2005). "Why most published research findings are false". In: *PLoS medicine* 2.8, e124.

Manfei, XU et al. (2017). “The differences and similarities between two-sample t-test and paired t-test”. In: *Shanghai archives of psychiatry* 29.3, p. 184.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.

Strobl, Carolin et al. (2007). “Bias in random forest variable importance measures: Illustrations, sources and a solution”. In: *BMC bioinformatics* 8, pp. 1–21.

Ugoni, Antony and Bruce F Walker (1995). “The Chi square test: an introduction”. In: *COMSIG review* 4.3, p. 61.

Wickham, Hadley (2020). “Package ‘reshape2”’. In: URL: `https://github.com/hadley/reshape`.

Wickham, Hadley, Winston Chang, and Maintainer Hadley Wickham (2016). “Package ‘ggplot2”’. In: *Create elegant data visualisations using the grammar of graphics. Version* 2.1, pp. 1–189. URL: `https://github.com/tidyverse/ggplot2`.

Wünsch, Milena (2022). “Over-Optimism in Gene Set Analyses: how does the choice of methods and tools influence the detection of enriched gene sets?” MA thesis. ”LMU Munich, department of Statistics”.

Appendices

A Appendix

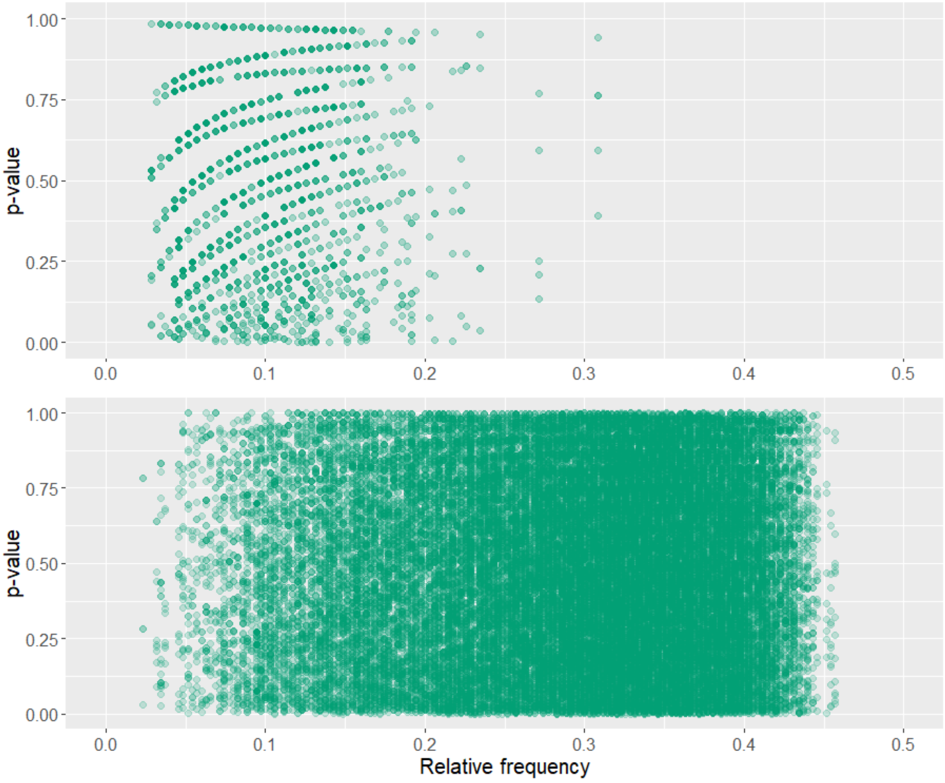

Figure 24: P-values of the chi-square tests on five permutations of the HapMap dataset without Yate's correction applied, split in SNPs with two categories (top) and SNPs with three categories (bottom)

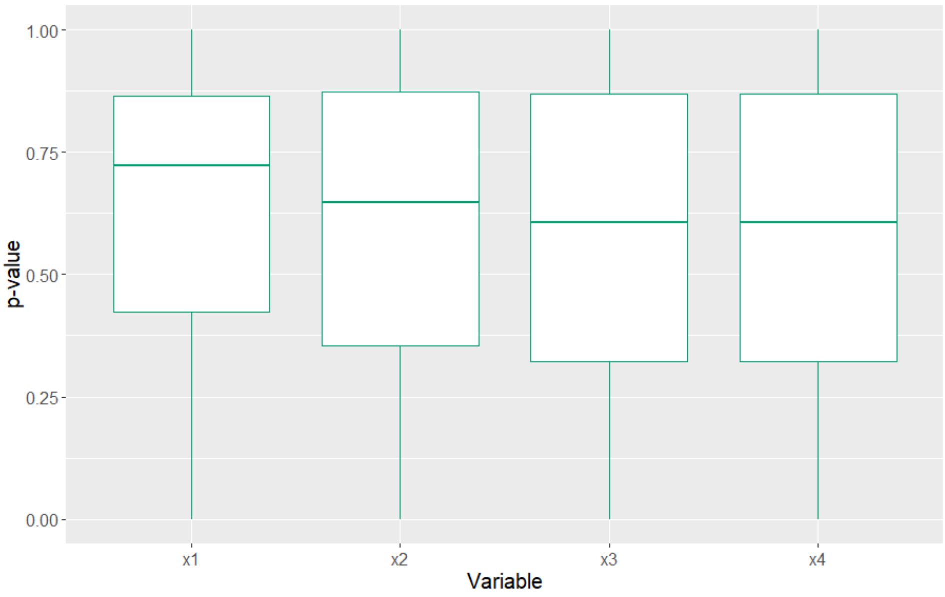

Figure 25: Boxplots of p-value distributions from CMH tests with Yate's correction between the two-categorical explanatory variables X_1, ..., X_4 and the outcome variable, which is unevenly distributed. 10000 datasets are simulated, tests done on five permutations of each dataset.

Table 15: Percentages of occurence of lowest and significant p-values for the CMH tests with Yate's correction of the two-categorical explanatory variables. 10000 datasets are simulated, tests done on five permutations of each dataset. The outcome Y is unevenly distributed.

	Lowest p-value	Significant p-values
X_1	19.3%	2.1%
X_2	24.9%	2.7%
X_3	27.6%	3.3%
X_4	28.1%	3.3%

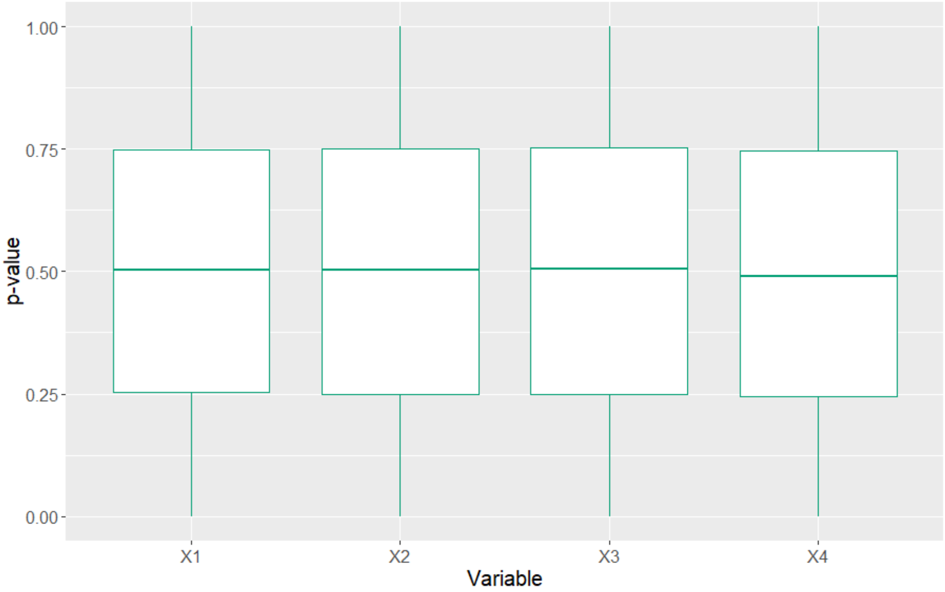

Figure 26: Boxplots of p-value distributions from chi-square tests of independence without Yate's correction between the two-categorical explanatory variables X_1, ..., X_4 and the outcome variable. 10000 datasets are simulated, tests done on one permutation of each dataset.

B Electronic appendix

The electronic appendix consists of an electronic copy of this thesis (`Thesis_Orzelek.pdf`), the HapMap dataset as prepared and processed by Boulesteix et al. (2012) and four `R` code files. The file `t_test_simulations` contains the code to reproduce the results of Chapter 2, the file `chi_square_simulations` contains the code to reproduce the results of Section 3.4, the file `chi_square_HapMap` contains the code to reproduce the results of Section 3.5 and the file `CMH_test_simulations` contains the code to reproduce the results of Section 4.3. Further explanations for the application of the code are given in the code files themsleves.

Declaration of authorship

I hereby confirm that I have authored this Master's thesis independently and without use of other resources other than those indicated. The ideas taken directly or indirectly from external sources are duly acknowledged in the text. The material, either full or in part, has not been previously submitted for grading at this or any other academic institution.

Munich, May 23, 2024

Anna Maria Orzelek