Researcher Degrees of Freedom in Methodological Simulation Studies: An Illustration from Medical Statistics

Master Thesis

Author:
Sarah Musiol

Supervisor:
Prof. Dr. Anne-Laure Boulesteix

M.Sc. Statistik
Department of Statistics
Ludwig-Maximilians University Munich

June 5, 2023

Abstract

The design and analysis of a study involve many choices that are often arbitrary. These choices can affect the results and conclusions, referred to as researcher degrees of freedom. Recently, Pawel, Kook, and Reeve (2022) have shown that even simulation studies are affected by these questionable research practices.

Specifically, in medical research, a randomised controlled trial (RCT) might draw false-positive conclusions due to questionable research practices, which can be damaging. Many statistical methods that would also be used to analyse an RCT are primarily based on comparative simulation studies. However, treatment would not be approved after only completing the preclinical stage, which can be compared to a simulation study (Boulesteix, Wilson, & Hapfelmeier, 2017). Therefore, this thesis investigates researcher degrees of freedom in a simulation study based on an RCT for an ordinal outcome variable. The study consists of 600 combinations containing choices in the sample size, the number of categories of the ordinal outcome and the probability distribution in the allocation variable. Furthermore, the probability distributions are selected theoretically, but more realistic probabilities were included from previous studies.

In conclusion, the investigated researcher degrees of freedom have been shown to impact the conclusions drawn in specific scenarios. Specifically, the sample size and the number of categories have proven essential choices in the study design. However, these researcher degrees of freedom are themselves chosen arbitrarily. Therefore, further research on realistic probability distributions might be beneficial.

Contents

1 Introduction and Related Work

Degrees of freedom are the maximum number of logically independent values, which have the freedom to vary, in the data sample, as explained by Student (1908). Any study undergoes a decision-making process, often involving arbitrary choices, that likewise have the freedom to vary.

For instance, a researcher must decide on the evaluation criteria of their study (Pawel et al., 2022). Nevertheless, these choices could affect the outcome of a study and hence affect the conclusions drawn from the research. These choices are referred to as researcher degrees of freedom (Simmons, Nelson, & Simonsohn, 2011).

Opportunistic use of researcher degrees of freedom increases the chances of false-positive research findings. For instance, choosing evaluation criteria based on a better representation of the results would be an opportunistic use of a researcher degree of freedom. According to Wicherts et al. (2016), researcher degrees of freedom affect formulating hypotheses, designing, running, analysing, and reporting studies.

In medical research, scientists often conduct randomised controlled trials, which investigate the efficacy and superiority of a treatment. However, the efficacy and superiority of a statistical method are primarily based on comparative simulation studies. (Boulesteix, Hoffmann, Charlton, & Seibold, 2020; Morris, White, & Crowther, 2019)

Boulesteix et al. (2017) have compared a simulation study to the preclinical phase of medical research. The authorities would not approve a new treatment that has only completed the preclinical stage. Furthermore, a patient would not feel safe to take a drug only tested in the laboratory. Statistical methods, however, do not necessarily undergo a rigorous study procedure before their usage.

The development of a new method or algorithm often involves a trial-and-error learning task, as Boulesteix (2010) stated. Thus, a researcher would naturally fish for improvement. According to Jelizarow, Guillemot, Tenenhaus, Strimmer, and Boulesteix (2010), the trial-and-error phase in developing new algorithms represents an essential component of data analysis research, which results in an unpredictable search process. Consequently, the new algorithm is optimised for the data sets considered during development. Yousefi, Hua, Sima, and Dougherty (2010) have conducted an extensive study on this data set bias, which has also been addressed by Jelizarow et al. (2010).

Another source of this over-optimism, as described by Jelizarow et al. (2010), has also been addressed in the study conducted by Ullmann, Beer, Hünemörder, Seidl, and Boulesteix (2023), who have investigated the choice of competing methods amongst others. The primary method might seem superior to the selected comparisons by not mentioning other relevant algorithms. However, its performance might be similar to other existing methods not included in the study.

Furthermore, Ullmann et al. (2023) have illustrated that the variation of specific data parameters can impact the performance of a new algorithm. For instance, the sample size variation might provide information about the performance of a new approach, precisely, how robust the method is.

In addition, Nießl, Herrmann, Wiedemann, Casalicchio, and Boulesteix (2022) have performed a benchmark study discovering that the choice of performance

measure affects other decisions made in a study. The data set selection contributes another significant part of the variability in study results, which coincides with the results obtained by Jelizarow et al. (2010) and Ullmann et al. (2023).

With many optimisation biases, selecting an appropriate method for a given research question might seem challenging. Thus, Nießl et al. (2022) have also stated the importance of neutral benchmark studies to counteract those biases.

However, prior to a benchmark study on real data sets, a new method should be investigated by a comparative simulation study and show superiority in selected settings. Besides, some applications might only allow for simulation studies. Such comparative simulation studies are often the only guidance for researchers to choose a method for a given setting (Pawel et al., 2022). Hence the results of flawed and biased studies can be very damaging.

Recently, Pawel et al. (2022) illustrated that questionable research practices undermine the validity of results from comparative simulation studies. Specifically, the altering of the data-generating process, the removal of competitor methods and the selective reporting of simulation results have shown an impact on the research outcome. Therefore, simulation studies are not excluded from the effects of questionable research practices and over-optimism.

Thus far, the superiority of a method is influenced by several choices made throughout a study illustrated in different areas of data analysis. Moreover, Wicherts et al. (2016) have published a more extensive list of questionable research practices categorised according to different phases in a study. However, the questionable research practices in a comparative simulation study might differ from this list. Therefore, Pawel et al. (2022) have adapted those questionable research practices to comparative simulation studies.

Pawel et al. (2022) have illustrated questionable research practices in altering the data-generating process, removing competitor methods and selective reporting in comparative simulation studies. Hence, a comparative simulation study can investigate the performance of the methods in question and provide information on which method to choose for a research question. However, if the results are corrupted by the researcher degrees of freedom, one method might be favoured, even if another could perform better. Therefore, in medical research, a randomised controlled trial might draw false-positive conclusions favouring a treatment that might not be as beneficial.

The following simulation study is based on a randomised controlled trial illustrating researcher degrees of freedom and their impact by comparing various methods for an ordinal outcome variable. This study does not focus on one particular method and its comparison to competitor methods as researched in Pawel et al. (2022). Still, it investigates the overall performance and researcher degrees of freedom in the data-generating process and their impact. The study design is explained in section 2, along with questionable research practices and the corresponding methods. The study results will be presented in section 3, followed by a discussion and conclusion in section 4 and section 5, respectively.

2 Study Design and Methods

In medical research, a randomised controlled trial (RCT) is conducted to investigate the efficacy and superiority of a new treatment. A possible primary outcome measured in an RCT is an ordinal variable. For instance, the modified Rankin scale is a commonly used ordinal outcome variable in stroke trials, which measures the degree of disability among individuals who have suffered a stroke (Selman, Lee, & Mahar, 2022).

A comparative simulation study investigates the plethora of methods available and informs about appropriate methods. However, the efficacy of a treatment is influenced by several choices in the study design besides the choice of method. Thus, the conclusions of an RCT might be impacted by decisions made throughout a study. Many research practices can often be unjustified because they consist of arbitrary choices.

This simulation study illustrates researcher degrees of freedom in the data-generating process based on an RCT for an ordinal outcome. This section addresses questionable research practices, along with the study design and the corresponding methods.

2.1 Questionable Research Practices

A researcher has to make many arbitrary choices during their study. Nevertheless, these decisions might impact the conclusions drawn from their research. Often, researchers need to be made aware of any guidelines or impacts on those choices, which can yield false-positive results. Wicherts et al. (2016) have presented a list of so-called researcher degrees of freedom impacting study results by making various choices in the study design.

Wicherts et al. (2016) detected the problem of p-hacking in psychological research and, being a psychologist himself, developed this list of researcher degrees of freedom categorised by the stages of a psychological study. Studies in other research fields have a similar structure to those presented in Wicherts et al. (2016).

In simulation studies, however, researchers do not face many decisions about measuring and collecting data and such. Thus, it might seem reasonable to assume that simulation studies do not face similar challenges. However, Pawel et al. (2022) have adapted questionable research practices to comparative simulation studies. The stages in the research process consist of the design, the execution, and the reporting of the study. Questionable research practices concerning the data-generating process can be found in the design and execution of a study. The data-generating process is either not precisely defined in the design stage or adapted during execution.

The following study focuses on researcher degrees of freedom in the data-generating process investigating many possible scenarios.

2.2 The Study Design according to the ADEMP Structure

A comparative simulation study might lead to over-optimistic or misleading conclusions in many different fields of methodological research. As stated in Boulesteix et al. (2020), one of the so-called seven sins of methodological research is the lack of reporting guidelines. There are many decisions to be made in comparative

simulation studies, and transparent reporting of those decisions is necessary to justify modified conditions (Pawel et al., 2022).

Poorly reported results and a careless study design can be caused by statisticians either lacking the necessary understanding to execute a simulation study confidently or being overconfident (Morris et al., 2019). However, Morris et al. (2019) developed a guideline for planning simulation studies to prevent reporting bias and restore confidence. This structured approach to planning a simulation study covers the aim of the study, the data-generating mechanism, the estimands, the methods which are compared and the performance measures used (ADEMP).

This simulation study is based on a randomised controlled trial investigating the effect of a fictional treatment where the outcome variable is an ordinal variable with at least three categories. Table 1 shows the contingency table for J categories on the binary treatment allocation variable.

	1	2	...	J	Total
Treatment	n_{11}	n_{12}	...	n_{1J}	n_{1+}
Control	n_{21}	n_{22}	...	n_{2J}	n_{2+}
Total	n_{+1}	n_{+2}	...	n_{+J}	n

Table 1: Notation for the cell counts in a $2 \times J$ contingency table

The study aims to illustrate researcher degrees of freedom in the data-generating process. The decisions during the data-generating mechanism involve the sample size as well as the number of categories in the case of ordinal data and the corresponding probabilities with which each category occurs in the data set. The probabilities are drawn from a parametric distribution. Generally, these distributions are decided theoretically. This is the first part of this study. However, if information about the probability distributions is available, a simulation on previous parameters might be more realistic and provide better guidance. Therefore, on top of the first part of the study, previous probabilities were researched, and researcher degrees of freedom were also illustrated in those cases.

Henceforth, the study design of this simulation study is explained according to the aforementioned ADEMP structure by Morris et al. (2019).

2.2.1 Aim

This simulation study aims to identify possible researcher degrees of freedom and their impact on the performance of specific methods. Regarding medical research, randomized controlled trials are essential to investigating treatment effects. Therefore, putting a lot of effort into making appropriate decisions in these studies should be paramount. Hence, identifying possible researcher degrees of freedom in simulation studies might prevent misleading conclusions.

2.2.2 Data-Generating Mechanism

The number of data sets included in a benchmark study is usually based on practical criteria such as availability or computational cost rather than statistical considerations (Nießl et al., 2022). However, a simulation is not bound to the availability of data sets. Therefore, the primary concern in the data-generating mechanisms in a simulation study is the computational power because the amount of integrated iterations and combinations depends on the run time for each scenario. This simulation study computed $n_{sim} = 10000$ iterations for each combination.

Further, the data can be generated by a parametric draw from a known distribution. The binary treatment allocation variable represents the patient allocation to the treatment or control group. The allocation should be random and contain roughly the same amount of patients in each arm. Thus, the allocation sample is drawn from a binomial distribution with probability $\pi = 0.5$ for a given number of observations.

Besides, a parametric draw from a multinomial distribution for each allocation observation generates the corresponding ordinal outcome variable. There are many different decisions in this step of the data-generating mechanism. Figure 1 summarises the choices considered in this simulation study for generating the multinomial data.

Different research questions use other primary outcome variables. Hence, the number of categories in ordinal outcome is subject to change. The number of categories included in this simulation are $categories = \{3, 5, 7, 9\}$ with cardinality $|categories| = 4$.

The choice of sample size is another researcher degree of freedom. For instance, Fisher's exact test for count data was developed for small sample sizes and should yield comparatively more robust results with fewer observations. The sample sizes considered here are $sample\ size = \{30, 50, 100, 200, 300, 500\}$ with cardinality $|sample\ size| = 6$.

One more choice is the probability of specific categories occurring in the data. Many possibilities exist in choosing a probability distribution. For instance, the same probability in the allocation groups illustrates the type-I error. However, a purely theoretical parametric simulation might be unrealistic. Hence, realistic probabilities from previous studies have been added to the considered probability distributions. The probabilities considered in this study are uniform, linear increasing, an increase in the intermediate category and geometric distribution, which are explained shortly. Further, the probabilities from previous work are based on the modified Rankin scale in patients who suffered a stroke. For the treatment and control groups, a multinomial outcome variable is sampled according to the five probability distributions, resulting in $|probabilities| = 5^2 = 25$ combinations.

The described researcher degrees of freedom are independent of each other. Hence, the total amount of data-generating mechanisms results in $|categories| \cdot |sample\ size| \cdot |probabilities| = 6 \cdot 4 \cdot 25 = 600$ combinations.

Figure 1 shows parts of 600 possible combinations between the number of categories, sample size and probability distribution. For instance, stroke trials commonly use the modified Rankin scale, which typically consists of seven categories. The path in black shows a combination of an analysis of 200 observations

and an ordinal variable with seven categories. The probability distribution corresponds to a previous study on the modified Rankin scale.

Uniform probability distribution A uniform probability distribution refers to the event of all categories having an equal probability of occurring in the data. The sequence of uniform probabilities can be written as $(prob_k^{uniform})_{k \in K} = \frac{1}{K}$. This probability is denoted in this thesis as *uniform*.

Linear increasing probability distribution The categories' frequencies might linearly increase, which describes the event of a linear increasing probability distribution. The sequence can be described as $(prob_k^{linear})_{k \in K} = k\frac{2}{K(K+1)}$. The linear increasing probability is abbreviated in this study as *linear*.

Probability distribution with an increase of the intermediate category If the frequency of only one category increases, specifically the intermediate category, the remaining categories stay equal according to the uniform probabilities. For instance, the sequence for three categories is $(prob_3^{intermediate}) = \{\frac{1}{6}, \frac{4}{6}, \frac{1}{6}\}$. This probability is abbreviated as *intermediate* from here onwards.

Geometric probability distribution Utilising the geometric sequence, the probability distribution for exponentially increasing category frequencies can be described as $(prob_k^{geometric})_{k \in K} = a2^{k-1}$ with $a = \frac{1}{\sum_{k=1}^{K} 2^{k-1}}$. The geometric probability is denoted as *geometric* in this study.

Probability distribution for patients who suffered a stroke The probabilities from previous work are taken from various studies conducted on stroke patients. Goyal et al. (2015) provided a probability distribution for the treatment and control groups with seven categories on the modified Rankin scale. The baseline patient characteristics in Table 6 in the study by Langhorne, Wu, Rodgers, Ashburn, and Bernhardt (2017) reported a premorbid modified Rankin scale with three categories for the treatment and control group. Moreover, the TARDIS study conducted by Bath et al. (2018) listed the outcome based on a modified Rankin scale for the treatment and control groups, which could be aggregated for five and nine categories specified in the post hoc analyses table. The probabilities from these trials are denoted as *stroke* in the following study and are specified in the appendix in Table 3.

2.2.3 Target

Most methods developed for ordinal outcome variables are hypothesis tests. Thus the target is the null hypothesis. Regarding regression models, the target remains the null hypothesis, rather than an estimand or prediction, to ensure comparability. However, this simulation study aims to illustrate researcher degrees of freedom through various scenarios. Therefore, this study's overall target is the design that aims to investigate researcher degrees of freedom.

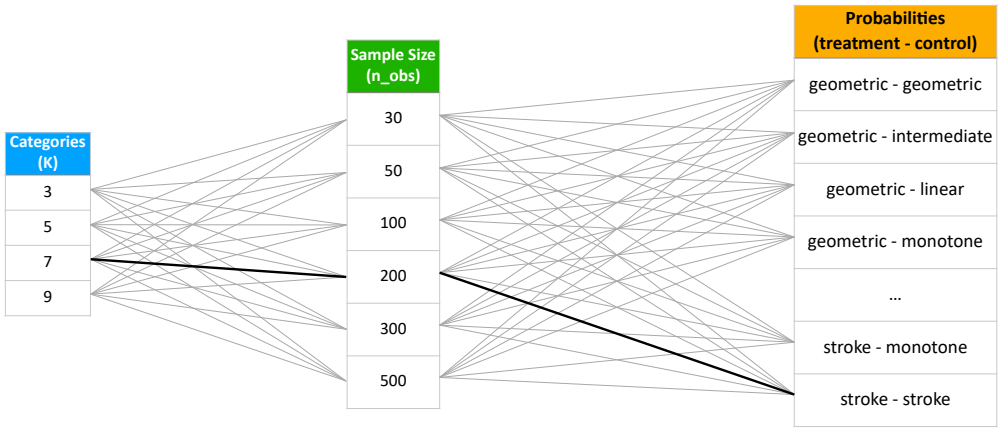

Figure 1: The 600 possible combinations in this simulation study consist of the number of categories, the sample size and the probability combinations for the treatment and control group. For instance, the path in black shows the combination of an ordinal variable with seven categories analysed for 200 observations and a probability distribution according to a real-life study conducted for stroke patients.

2.2.4 Methods

When comparing several methods, the selection of serious contenders is crucial. Serious contenders can be identified by researching previous work in the field of interest. The chosen method should also be implemented in software. Otherwise, making a convincing argument for its inclusion in the simulation study is difficult. The most common methods for an ordinal outcome are the Mann-Whitney test, the chi-squared test of independence, the Cochran-Armitage test, Fisher's exact test for count data, Stuart's τ_C, logistic regression for a dichotomised outcome variable and the proportional odds model. These methods are described in more detail in subsection 2.3.

2.2.5 Performance Measure

A performance measure describes a numerical quantity used to assess the performance of a method. A null hypothesis target's most common performance measures are type-I error rate and power.

Much information can be extracted from a hypothesis test, however, the primary objective of this study is the rejection rate. Most commonly, the decision about rejecting the null hypothesis is made by comparing the p-value to the significance level. Thus, the rejection rate can be calculated by the number of

rejections per method for n_{sim} iterations for each combination in the study design

$$\text{rejection rate} = \frac{\#\text{rejections}}{n_{sim}}. \tag{1}$$

All methods described in subsection 2.3 investigate the null hypothesis of independence. In comparative simulation study, independence is defined as the ordinal outcome variable being independent of the allocation variable. Thus, independence is achieved when the ordinal outcome of the treatment and control group have an equal probability distribution.

The type-I error occurs when the null hypothesis is rejected, although the null hypothesis is true. Hence, the rejection rate in case of equal probabilities for treatment and control group represents the type-I error.

Conversely, power is defined as rejecting the null hypothesis when the alternative hypothesis is true. The alternative hypothesis generally specifies no independence, corresponding to different probability distributions in the treatment and control groups.

2.3 Methods

The randomised controlled trial (RCT) analysed in this thesis consists of an allocation variable which indicates whether an observation is allocated to the treatment or control group. The allocation variable is a binomially distributed random variable denoted as X.

Further, the outcome variable of the RCT is ordinal. For instance, the disability of patients who suffered a stroke is categorised on the modified Rankin scale consisting of seven categories. Therefore, the outcome variable is a discrete multinomial random variable denoted as Y, which contains J categories. The primary objective of an RCT is to investigate the presence of a treatment effect. In other words, this RCT investigates the efficacy of treating stroke patients.

The methods utilised for this simulation study are explained in this section. The $2 \times J$ contingency table in Table 1 depicts the number of observations for allocation variable X and the J categories of Y. If not specified otherwise, the following methods are based on the explanations in the book by Agresti (2002).

2.3.1 Mann-Whitney U Test

A method for investigating a treatment effect has been proposed by Wilcoxon (1945). It utilises ranking methods, which substitute the numerical data with scores. Moreover, a treatment effect is defined by a shift in location. Therefore, the data sample is divided into a ranked treatment and control sample.

However, Wilcoxon (1945) considered only the case of equal sample sizes in the treatment and control groups ($n_{1+} = n_{2+}$). The allocation variable X is binomially distributed with $\pi = 0.5$. Nevertheless, the allocation samples are not necessarily the same size. Hence, Mann and Whitney (1947) have proposed a statistic U which considers unequal sample sizes.

Hypothesis The allocation samples are denoted as X_1 and X_2 for the treatment and control group, respectively, with cumulative distribution functions F_1 and

F_2. If a shift in the location of a cumulative distribution function is detected, a treatment effect exists. Thus, the null hypothesis is

$$H_0 : F_1(a) = F_2(a); \quad \forall a, a \in \mathbb{R}. \tag{2}$$

A location shift can occur in only one direction. Consequently, the alternative hypothesis typically specifies that X_1 is stochastically smaller (or larger) than X_2, which corresponds to $F_1(a) > F_2(a)$ (or $F_1(a) < F_2(a)$) for every a.

The U statistic In order to calculate the U statistic, let the combined sample of n_{1+} X_1-values and n_{2+} X_2-values be arranged in order. Given the continuity assumption, this arrangement is unique with probability $\mathbb{P}(x_{1i} = x_{2i}) = 0$. The statistic T proposed by Wilcoxon (1945) specifies the sum of the ranks of variable X_2 in the ordered sequence of x_1's and x_2's. However, the statistic U counts the number of times x_2 precedes x_1 and can be computed as

$$U = n_{1+}n_{2+} + \frac{n_{2+}(n_{2+} + 1)}{2} - T. \tag{3}$$

Mann and Whitney (1947) have tabulated the probabilities of U for small sample sizes. It was shown that for large sample sizes (greater than $n_{1+} = n_{2+} = 8$), the distribution of the test statistic U is asymptotically normal.

Assumptions and Applicability The Mann-Whitney U test relies on some assumptions. For instance, it assumes continuous cumulative distribution functions, as stated by Mann and Whitney (1947). The continuous cumulative distribution functions imply that the probability of obtaining the same value in populations X_1 and X_2 is 0 ($\mathbb{P}(x_{1i} = x_{2i}) = 0$). However, the ordinal data structure in the described setting of the RCT does not satisfy this continuity assumption. Indeed, the probability of observing the same value in the treatment and control samples is not zero. To satisfy this assumption, the ordinal data structure must become continuous. The test statistic U analyses ranks and the rank distribution. Therefore the ranking consists of only a few unique ranks and many ties. In the case of ties, mid-ranks are used.

In addition to the continuity assumption, Hollander, Wolfe, and Chicken (2013) have mentioned the assumption of independent and identically distributed samples X_1 and X_2, respectively, as well as X_1 and X_2 being mutually independent. This assumption is satisfied, considering that the nature of an RCT ensures independence and identical distributions.

Software This simulation study is executed in RStudio. The inbuilt function called `wilcox.test()` is used to analyse the Mann-Whitney U test. Hothorn, Hornik, van de Wiel, and Zeileis (2008) have implemented a class of tests organised in the `coin` package. The Mann-Whitney U test is included in the `coin` package. Both functions are further reviewed in section 3.

2.3.2 Chi-Squared Test of Independence

The chi-squared test of independence was first proposed by Pearson (1900), introducing the Pearson chi-squared statistic, which measures the goodness of fit.

This hypothesis test investigates the fit of a theoretical frequency distribution to an observed one. Table 1 in subsection 2.2 depicts the observed frequencies for the allocation groups and ordinal outcome variable. The corresponding cell probabilities are $\{\pi_{ij}\}$.

Hypothesis The null hypothesis of the chi-squared test investigates the statistical independence of two categorical variables

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j} \quad \text{for all } i \text{ and } j. \tag{4}$$

Generally, the marginal probabilities $\{\pi_{i+}\}$ and $\{\pi_{+j}\}$ are unknown and have to be estimated. The maximum likelihood estimates are the sample's marginal proportions and the estimated expected frequency $\hat{\mu}_{ij}$ results in

$$\hat{\mu}_{ij} = n\hat{\pi}_{i+}\hat{\pi}_{+j} = n\left(\frac{n_{i+}}{n}\right)\left(\frac{n_{+j}}{n}\right) = \frac{n_{i+}n_{+j}}{n}. \tag{5}$$

If the null hypothesis is true, the observed cell frequency n_{ij} should be close to the estimated expected frequency μ_{ij}. The alternative hypothesis states that there is at least one cell whose frequency is unequal to the product of its marginal frequencies.

Chi-squared statistic Pearson (1900) has described the measure of goodness of fit as the sum of the residuals between the observed and the estimated expected frequencies. Hence, the Pearson chi-squared statistic for testing H_0 is

$$\chi^2 = \sum_{ij} \epsilon_{ij}^2 = \sum_{ij} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}, \tag{6}$$

where ϵ_{ij} are the Pearson residuals. The χ^2 statistic is asymptotically chi-squared distributed with $df = (I-1)(J-1)$.

Assumptions and Applicability As mentioned, the chi-squared test of independence analyses the independence of two nominal variables. Hence, the ordinal structure of the data is not considered, and the test statistic will not change when changing the order of the categories.

Moreover, the chi-squared approximation improves as $\{\mu_{ij}\}$ increases where $\{\mu_{ij} \geq 5\}$ is sufficient for a decent approximation. Also, the sampling distribution of χ^2 gets closer to the chi-squared distribution as the sample size n increases relative to the number of cells.

Software The inbuilt function for the chi-squared test of independence is called `chisq.test()`. A continuity correction for 2×2 contingency tables can be applied. Usually, more than two categories are sampled for the outcome variable. However, low sample sizes and a low number of categories might cause samples with only two categories. In section 3, the difference in the rare cases with continuity correction is further explored.

2.3.3 Fisher's Exact Test for Count Data

The methods described so far can be applied in large sample settings. Usually, the corresponding distribution can be approximated. However, when the number of observations n is small, it might be recommended to use an exact distribution. Fisher's exact test initially considers the case for 2×2 tables. An extension for $I \times J$ tables is required with an ordinal outcome variable of more than two categories.

Hypothesis Within the $I = 2$ treatment groups, the ordinal outcome J is assumed to be independent and sampled from a multinomial distribution. The null hypothesis for testing the independence between the allocation variable and the ordinal outcome variable is

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j} \quad \text{for all } i \text{ and } j. \tag{7}$$

The corresponding alternative hypothesis considers either a positive or negative association.

Multiple hypergeometric distribution Rather than relying on distributions, Fisher's exact test calculates the null probabilities exactly. The null probability can be calculated through the following equation, which is the probability function of $\{n_{ij}\}$ conditional on $\{n_{+j}\}$ and $\{n_{i+}\}$,

$$\frac{(\prod_i n_{i+}!)(\prod_j n_{+j}!)}{n! \prod_i \prod_j n_{ij}!}. \tag{8}$$

It can be shown that Equation 8 corresponds to the multivariate hypergeometric distribution.

Applicability Generally, Fisher's exact test investigates a one-sided alternative hypothesis. However, a two-sided alternative might yield different p-values. An approach for calculating the p-value of a two-sided alternative is identical to $\mathbb{P}(\chi^2 \geq \chi_o^2)$, where χ_o^2 is the observed Pearson statistic.

Software The inbuilt function `fisher.test()` calculates their p-value exactly. The computing time increases immensely with increasing n and J in this study.

2.3.4 Stuart's τ_c

Kendall's rank correlation coefficients (τ_a and τ_b) are measures of strength of association in the case of two ordinal variables cross-tabulated into $I \times J$ contingency tables. These coefficients, however, assume data with no ties and an equal number of rows and columns. A modification to Kendall's rank correlation coefficients was proposed by Stuart (1953), which is independent of a scoring system and depends only on ordinal properties, called Stuart's τ_c rank correlation coefficient. In addition, it allows for an unequal number of rows and columns.

Hypothesis Stuart's τ_c's range is $[-1, +1]$, with $+1$ and -1 representing complete association and complete dissociation, respectively. Independence implies that $\tau_c = 0$. Therefore, the null hypothesis is

$$H_0 : \tau_c = 0. \tag{9}$$

The alternative hypothesis is $H_1 : \tau_c \neq 0$.

Stuart's τ_c Stuart's τ_c is a measure of ordinal association for ordinal contingency tables with $I \neq J$ and possible ties,

$$\tau_c = \frac{2m(C - D)}{n^2(m - 1)}. \tag{10}$$

The strength of association is denoted as $(C - D)$, where C is the number of concordant pairs, whereas D is the number of discordant pairs. A pair of observations is concordant if the observation ranking higher on X_1 also ranks higher on X_2. If the observation ranking higher on X_1 ranks lower on X_2, the pair is discordant. The total amount of observations in Equation 10 is n, and m is defined as $m = \min(I, J)$, which standardises the measure of the strength of association. Therefore, the test statistic ranges $-1 \leq \tau_c \leq +1$. Further, the null probability is calculated precisely.

Applicability The treatment variable in the RCT has two categories which are not necessarily ordered, and the outcome variable generally has more than two categories. Hence, the resulting contingency table consists of unequal row and column numbers.

Furthermore, the number of observations n equals at least 30. Thus, ties are present in the data.

Software Stuart's τ_c is not naturally implemented in RStudio. However, the package `DescTools` by Signorell (2023) provides the function `StuartTauC`, which computes Stuart's τ_c. The supplied confidence interval calculates the corresponding p-value.

2.3.5 Cochran-Armitage Test

Primarily, the order of the categories is not focused in most tests. However, utilising the ordering in categorical variables is essential. Rather than testing for a significant difference in the proportions, Armitage (1955) and Cochran (1954) have proposed a trend statistic to show a trend considering the ordering of the categories. The Cochran-Armitage test partitions the Pearson statistic to test the null hypothesis of independence. The $I \times 2$ tables consist of I ordered rows which in turn hold binomially distributed variates $\{y_i\}$ with probability π_i, which correspond to $\{x_i\}$ in this RCT.

Hypothesis Trends are often shown through linear regression models. Similarly, the Cochran-Armitage test uses a linear probability model,

$$\pi_i = \alpha + \beta x_i, \tag{11}$$

which is fitted by ordinary least squares. Hence, the probability π_i is independent of x_i (or y_i here) if $\beta = 0$, resulting in the following null hypothesis

$$H_0 : \beta = 0. \tag{12}$$

The alternative hypothesis can look in either direction, whether it is a positive or negative trend.

Cochran-Armitage Statistic For testing $H_0 : \beta = 0$, the Cochran-Armitage statistic in Equation 13 equals the score statistic of the linear logit model,

$$z^2 = \frac{\left(\sum_i (x_i - \bar{x}) y_i\right)^2}{p(1-p)\sum_i n_i (x_i - \bar{x})^2}, \tag{13}$$

where $p = (\sum_i y_i)/n$ (or $p = (\sum_i x_i)/n$ in this case). The statistic z^2 is chi-squared distributed with $df = 1$.

Applicability / Comments An RCT's objective is to investigate a significant difference in the outcome between the treatment groups. Hence, the outcome variable y would be the ordinal variable. The Cochran-Armitage test, however, defines the outcome variable as the binary variable x. Moreover, the similarity to the linear logistic regression model suggests a binary outcome with categorical explanatory variables. Therefore, the Cochran-Armitage test does not seem applicable in an ordinal outcome. The score statistic measures how far from zero the score function is when evaluated at the null hypothesis. However, the null hypothesis tests the independence of the ordinal and treatment variables. Hence, the test should be applicable as long as the null hypothesis is not rejected. The interpretation of the test in case the null hypothesis is rejected should be handled with care.

Software In the case of an ordinal outcome, the Cochran-Armitage test can be applied in RStudio by specifying the ordinal variable as the outcome variable. The inbuilt function `prop.trend.test()` computes the Cochran-Armitage test. In addition, the `coin` package by Hothorn et al. (2008) provides a suitable function. Both functions are discussed in section 3.

2.3.6 Dichotomised Logistic Regression

A regression model for binary outcome variables is the logistic regression model. A somewhat naïve approach to deal with more than two categories for the outcome variable is splitting J categories into $1, \ldots, j$ and $j+1, \ldots, J$ categories resulting in a dichotomous variable.

The logit link function of the logistic regression model has the linear relationship

$$\text{logit}\left(\pi(x)\right) = \log \frac{\pi(x)}{1-\pi(x)} = \beta_0 + \beta_1 x. \tag{14}$$

For all possible thresholds for an ordinal outcome, $J-1$ logits are calculated, hence choosing the best out of $J-1$ separate models.

Hypothesis For the logistic model with a single predictor specified in Equation 14, the null hypothesis of independence,

$$H_0 : \beta_1 = 0, \tag{15}$$

could be tested through the Wald, likelihood ratio, or score tests. The `glm()`-function in R utilises the Wald test by Wald (1943) to test the given hypothesis.

Wald statistic The parameter β_1 is estimated through the maximum likelihood method. The corresponding test statistic,

$$z^2 = \frac{\hat{\beta}_1^{\,2}}{SE^2}, \tag{16}$$

uses the log-likelihood at $\hat{\beta}_1$ and its standard error SE. Under the null hypothesis, z^2 is asymptotically χ_1^2 distributed.

Applicability The ordinal outcome variable in the RCT contains more than two categories. Since the outcome variable is not binary, the categories must be split into two categories. With a simulation, the best threshold is not necessarily known. Therefore, the result will be the best fit according to the lowest p-value after fitting models for each possible threshold. A Bonferroni correction is applied before comparison.

The threshold's choice considers the ordering of the categories. However, the model itself does not consider the order of a binary variable. An alternative to this naïve approach is a proportional odds model, which considers the ordering.

Software The inbuilt function `glm(., family = binomial)` is utilised for the logistic regression model. The p values are Bonferroni corrected with the function `p.adjust()`.

2.3.7 Proportional Odds Model

The logistic regression model explained in subsubsection 2.3.6 describes a model for a binary outcome. Applying this model in an ordinal case results in $J-1$ logits choosing the lowest p-value. The more appropriate method is to use all $J-1$ logits simultaneously, specifically all cumulative logits,

$$\begin{aligned} logit(\mathbb{P}(Y \leq j|x)) &= \log \frac{\mathbb{P}(Y \leq j|x)}{1 - \mathbb{P}(Y \leq j|x)} && (17)\\ &= \log \frac{\pi_1(x) + \ldots + \pi_j(x)}{\pi_{j+1}(x) + \ldots + \pi_J(x)}, \qquad j = 1, \ldots, J-1. && (18) \end{aligned}$$

The corresponding proportional odds model uses all cumulative logits simultaneously, $logit(\mathbb{P}(Y \leq j|x)) = \beta_{0j} + \beta_1 x$, with $j = 1, \ldots, J-1$. McCullagh (1980) refers to this model as the proportional odds model since the log cumulative odds ratio is proportional to the distance between x_1 and x_2.

Hypothesis and Statistic The null hypothesis of independence in the proportional odds model is

$$H_0 : \beta_1 = 0, \tag{19}$$

with the alternative of $H_1 : \beta_1 \neq 0$.

Typically, the statistic of the proportional odds model for this null hypothesis uses the score statistic

$$\frac{(\partial L(\beta)/\partial \beta^0)^2}{-\mathbb{E}(\partial^2 L(\beta)/\partial (\beta^0)^2)}, \tag{20}$$

with β^0 being the null value of the hypothesis.

McCullagh (1980) noted that the score test typically used in the proportional odds model for the $2 \times J$ table is equivalent to the Mann-Whitney test described in subsubsection 2.3.1.

Applicability The proportional odds model is prevalent for modelling an ordinal outcome. In the case of an ordinal outcome with a binary treatment allocation variable as the explanatory variable, the proportional odds model might be a natural choice. However, it should be noted that the proportional odds model requires the proportional odds assumption. It is impossible to check this assumption for the amount of simulation iterations calculated. Hence, the proportional odds assumption might only be met for some scenarios.

Software Various packages are implemented in R for the proportional odds model. This study utilises the `lrm()` function from the `rms` package by Harrell Jr (2023).

3 Results

This simulation study is based on randomised controlled trials for an ordinal outcome. As explained in subsection 2.2, the data-generating process consists of 600 combinations, including the number of categories, the sample size and the probability distribution. This study aims to illustrate possible researcher degrees of freedom in the data-generating process. The results are illustrated for theoretical probability distributions in subsection 3.1, and based on that, probability distributions from previous studies are investigated in subsection 3.2.

The analysis and the results presented in this section are available in the eponymous GitHub repository by Musiol (2023) to ensure reproducibility.

3.1 Theoretical Probability Distributions

The theoretical probability distributions in this study include a uniform, a linear increasing and a geometric distribution, as well as probabilities with an increase in the intermediate category. More details on these probabilities are found in subsection 2.2.

3.1.1 Overall Type-I Error

The type-I error is defined for equal probability distributions in the allocation variable, represented in the columns of Figure 2. The rows portray the number of categories, whereas the sample size is depicted on the x-axis. To get a smooth representation, a local polynomial regression model, the locally estimated scatterplot smoothing (loess), was applied over the rejection rate of the methods described in subsection 2.3.

Generally, the smoothed rejection rate seems to be accumulated around 0.05, represented by the black horizontal line, corresponding to the assumed significance level $\alpha = 0.05$. Across all probability distributions, the width of the confidence band increases with an increase in the number of categories. Furthermore, the smoothed rejection rate suggests that the scenarios with three categories might not accept the treatment for smaller sample sizes. Additionally, the geometric and the increase in the intermediate category probability distributions show a smoothed rejection rate below 0.05 across all sample sizes with an increase in the number of categories.

The underlying methods show a high discrepancy in the rejection rates for small sample sizes. The rejection rates seem to become more robust with an increase in the sample size. Whether the authorities would accept a treatment relies on the choice of method for small sample sizes. However, with sample sizes as large as 500, all rejection rates are below 0.05, and the treatment would be accepted no matter the probabilities, methods or the number of categories.

The logistic regression model displays a lower rejection rate overall, suggesting a very well fit across all combinations. However, the logistic regression model only considers one logit at a time instead of all categories simultaneously. Therefore, it chooses the best model for each combination, which might lead to overfitting.

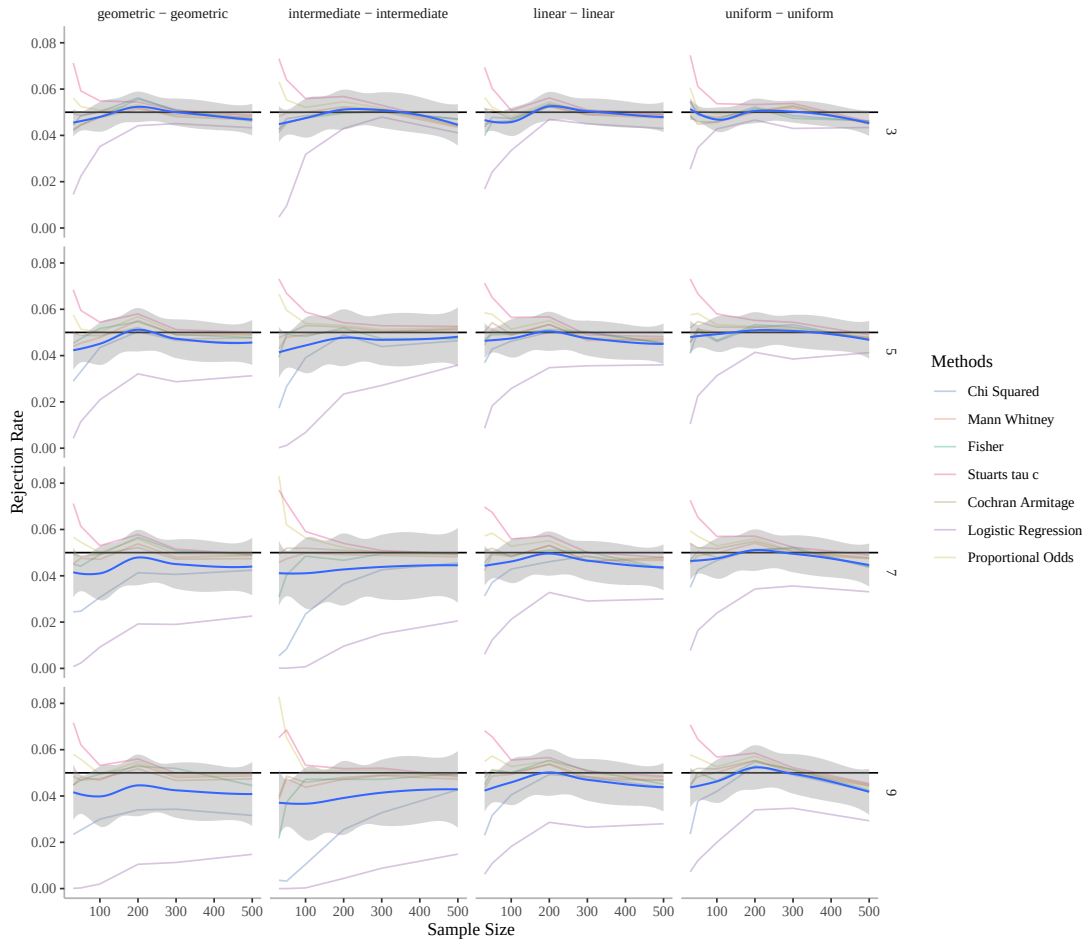

Figure 2: The smoothed rejection rate in blue was calculated through local polynomial regression along with a confidence band. The scenarios are represented for the type-I error across sample sizes and numbers of categories.

3.1.2 Type-I Error for the Chi-Squared Test with and without Correction

The chi-squared test approximates the binomial cell frequencies by the chi-squared distribution, and to satisfy the approximation, no cell can have an expectancy of less than five. This assumption might not be met in the case of 2×2 contingency tables. Therefore, Yates (1934) has proposed a continuity correction for 2×2 contingency tables to prevent overestimation by subtracting 0.5 from the difference between the observed and expected value.

The correction can be specified in the inbuilt function `chisq.test()` in RStudio, which applies only to 2×2 contingency tables. A sample with only two categories might occur with certain probability distributions, low sample sizes, and fewer categories.

Table 2 shows the cases in which a continuity correction was applied, and therefore, the results differ from the chi-squared test without continuity correction. The combinations with theoretical probability distributions report a slightly lower rejection rate for the chi-squared test with the continuity correction. However, a 2×2 sample was only detected with a sample size of 30 observations, primarily considering three categories. The chances of randomly sampling only

two out of three categories in 30 observations are significant, especially since the probabilities for specific categories are relatively low. For instance, the geometric sequence for three categories is $\{1/7, 2/7, 4/7\}$, with the third category taking up more than 50% of the occurrences.

The type-I error in Table 2 includes the first eight rows, which display similar probabilities in the treatment and control groups. The rejection rate of those combinations is about 0.05, which is an expected type-I error. The chi-squared test with correction displays minor type-I errors regarding Table 2 and should be preferred to the test without continuity correction. Nevertheless, the remaining rows show the corresponding power, which is slightly lower with continuity correction. In this case, the chi-squared test without continuity correction would be preferred.

Nevertheless, the continuity correction is applied in the case of two categories which might occur under certain circumstances. However, these situations might benefit from methods considering a binary outcome variable.

Sample Size (n_obs)	Categories (K)	Probability (treatment - control)	Rejection Rate (**without** correction)	Rejection Rate (**with** correction)	Difference in Rejection Rate
30	3	geometric - geometric	0.0444	0.0437	0.0007
30	3	intermediate - intermediate	0.0425	0.0421	0.0004
30	3	linear - linear	0.0437	0.0434	0.0003
30	5	geometric - geometric	0.0290	0.0288	0.0002
30	5	intermediate - intermediate	0.0173	0.0172	0.0001
30	7	intermediate - intermediate	0.0054	0.0053	0.0001
30	9	intermediate - intermediate	0.0036	0.0034	0.0002
50	9	intermediate - intermediate	0.0032	0.0031	0.0001
30	3	geometric - intermediate	0.5871	0.5856	0.0015
30	3	geometric - linear	0.0558	0.0556	0.0002
30	3	intermediate - geometric	0.5920	0.5915	0.0005
30	3	intermediate - linear	0.4419	0.4411	0.0008
30	3	linear - geometric	0.0533	0.0529	0.0004
30	3	linear - intermediate	0.4442	0.4436	0.0006
30	3	uniform - intermediate	0.3518	0.3516	0.0002

Table 2: The chi-squared test with and without correction show differences in the rejection rate for theoretical probabilities. The 15 combinations in the data-generating process exhibit differences in the type-I error (upper part) and power (bottom part).

3.1.3 Type-I Error for the Chi-Squared Test and Fisher's Exact Test for Count Data

Fisher's exact test for count data was developed specifically for small sample sizes and calculates exact null probabilities. For higher sample sizes, the computing time increases, but the results should be similar to the chi-squared test of independence.

Figure 3 depicts the type-I error for both methods according to the number of categories and sample size. The lower left quadrant of Figure 3 shows a higher discrepancy between both methods, whereas the rejection rate of the remaining quadrants is more similar, especially towards higher sample sizes.

Considering ordinal data with three categories shown in the first row of Figure 3, the rejection rate of both methods indicates no distinct preference. The rejection rate at smaller sample sizes shows slightly higher differences. Nonetheless, both methods are around the expected type-I error.

The difference in the rejection rate across the probability distributions seems to diverge with an increased number of categories. The higher the number of categories analysed, the higher the difference between both methods seems to be. The geometric sequence and the increase of the intermediate category seem to have the most considerable discrepancies in the nine categories. On the other hand, the uniform probability displays the lowest differences in the probability distributions.

In those discrepancies, the chi-squared test shows lower rejection rates which would be desirable. However, Fisher's exact test was developed to forgo the necessity for approximations which might not produce accurate results. Therefore, the chi-squared test might yield slightly inaccurate results, at least for small sample sizes.

The treatment would not be accepted according to Fisher's exact test for count data for more scenarios compared to the chi-squared test. These differences become more distinct with the increase in the number of categories.

The results from Figure 3 show that the sample size, the number of categories and the probability distribution cause the rejection rates to differ and yield different results and conclusions. In addition, Fisher's exact test seems to yield steadier rejection rates, whereas the chi-squared test is subject to change.

3.1.4 Type-I Error for the Regression Models

As mentioned, the logistic regression model shows the lowest rejection rate regarding the type-I error, as seen in Figure 2. The rejection rate seems over-optimistic since the logistic regression model chooses the lowest p-value from all $J-1$ models.

Overall, a smaller sample size seems to yield more optimistic results. However, considering the samples with three categories, the logistic regression model reaches 0.05 as the remaining models. Moreover, the geometric and the intermediate probabilities show higher deviations than the linear and uniform probabilities.

Generally, the proportional odds model shows more robust rejection rates than the logistic regression model. Figure 9 in the appendix represents the type-I error of both models separate from the remaining methods.

Furthermore, McCullagh (1980) have claimed that the hypothesis test results of the proportional odds model are equivalent to the Mann-Whitney U test for this specific case of $2 \times J$ contingency tables. A representation of the type-I error for the proportional odds model and the Mann-Whitney U test can be found in Figure 10 in the appendix. The rejection rates are similar to each other, however, some discrepancies can be spotted throughout the various scenarios.

3.1.5 Overall Power

If the allocation probabilities differ for the treatment and control groups, the results represent the power. Usually, an RCT is supposed to have a power of at least 80%, which is depicted with a black horizontal line in Figure 4 and Figure 5. The rejection rates across all methods are summarized by loess as

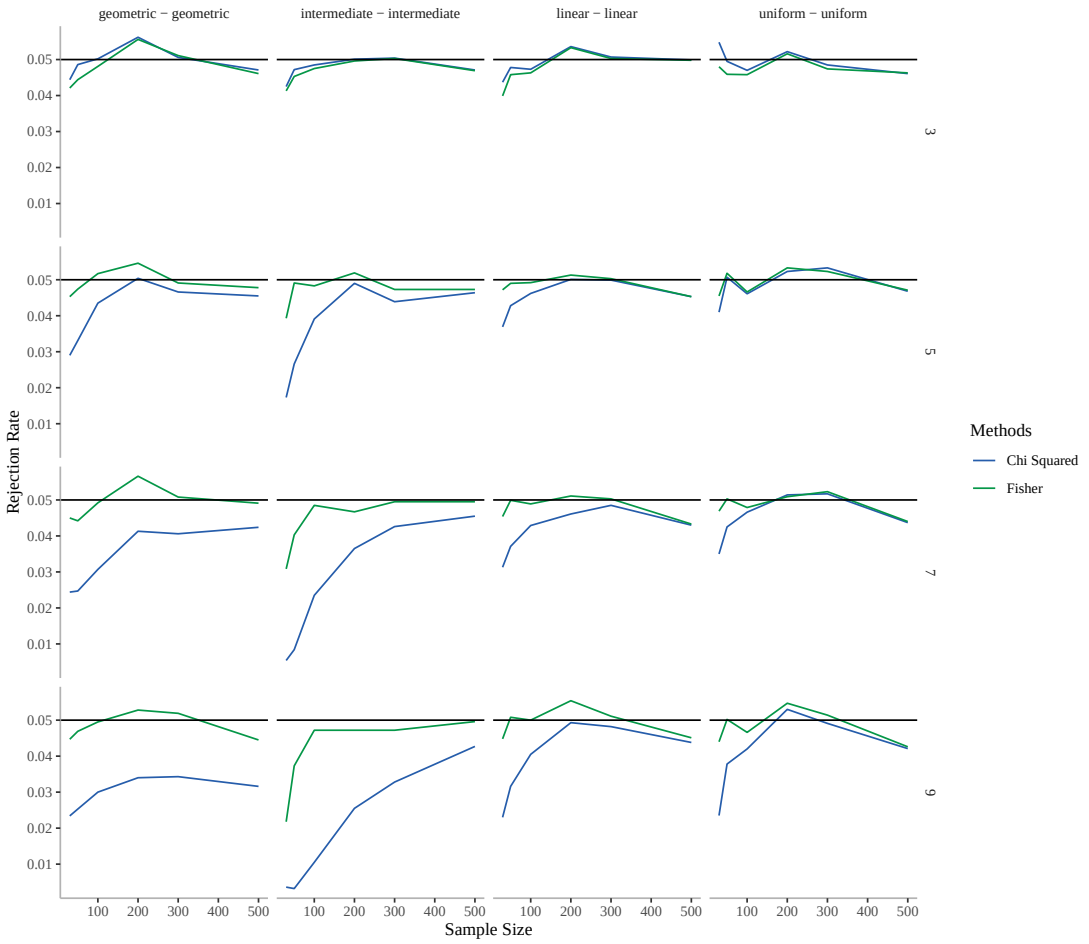

Figure 3: The rejection rates for the selected method are displayed for the chi-squared and Fisher's exact tests. The scenarios are represented for the type-I error across sample sizes and numbers of categories.

explained for the type-I error. The rejection rates of the allocation groups do not show any differences when the probabilities are reversed. Therefore, the results are illustrated for one probability combination, and the reversed ones are depicted in the appendix.

Figure 4 shows rejection rates that converge quickly to 100% with an increasing sample size. Many scenarios show a rejection rate of less than 80% for small sample sizes, which would not be sufficient for treatment approval. However, a sample size of about 200 is improving the performance considerably.

Moreover, the scenarios with three categories show the lowest rejection rates for small sample sizes. With more categories, the rejection rates keep mostly above 80%, where the geometric-intermediate probabilities perform best. Further, the width of the confidence band of the smoothed rejection rate increases with the number of categories. However, the methods' rejection rate does not show a high variability overall.

Alas, some probability distributions show an unexpected rejection rate, as depicted in Figure 5. The linear - uniform probabilities are similar to Figure 4. The convergence toward the power of 100% increases with the sample size. However, the rejection rates start below 50% and only reach 80% at sample sizes of at least

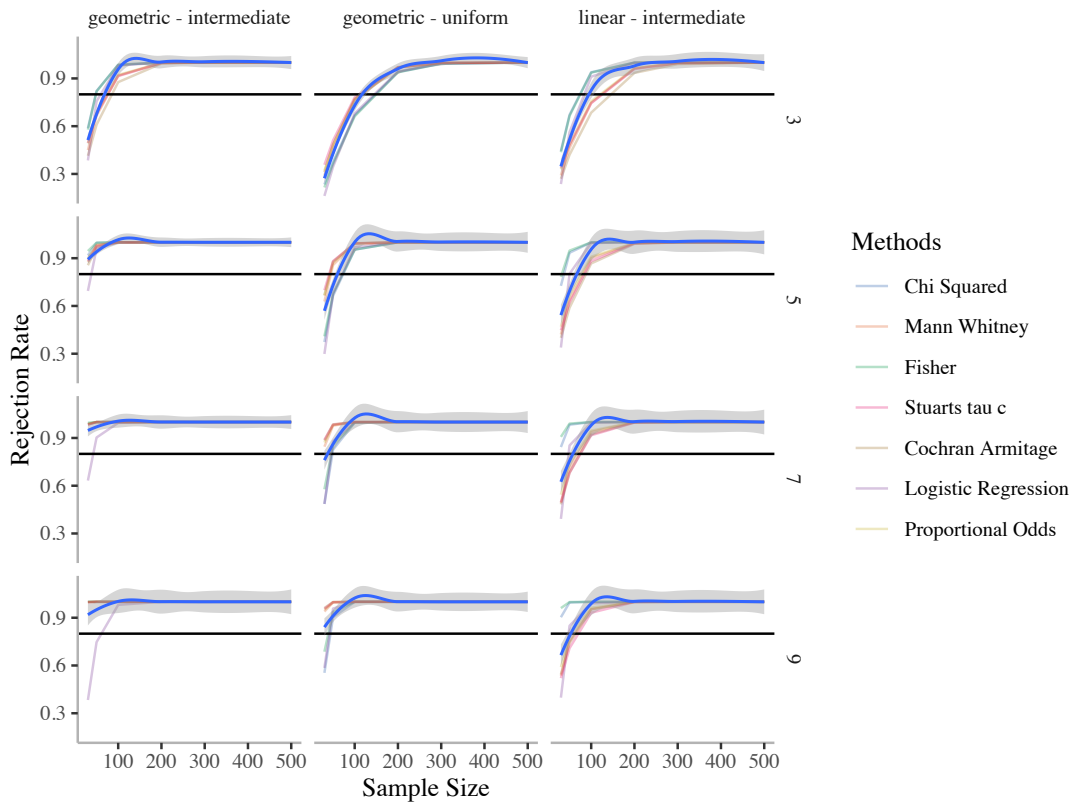

Figure 4: The smoothed rejection rate in blue was calculated through local polynomial regression along with a confidence band. The power is displayed for three probability distributions. The reversed allocation groups show no distinct differences. Those probability distributions are represented across sample sizes and numbers of categories. The rejection rates suggest a fairly good performance.

200.

The geometric-linear probabilities show a meagre power of less than 50% across all sample sizes for the three-categories scenarios. However, the probabilities for the geometric sequence and the linear increase are very similar with $\{1/7, 2/7, 4/7\} = \{6/42, 12/42, 24/42\}$ and $\{1/6, 2/6, 3/6\} = \{7/42, 14/42, 21/42\}$, respectively. Therefore, it might be hard to spot that the treatment and control groups are not independent, especially for small sample sizes. Regardless, the rejection rate improves considerably with an increase in the number of categories. Evidently, the number of categories considerably impacts the performance in the case of geometric-linear probabilities.

The intermediate-uniform distribution exemplifies the most considerable discrepancy. Some methods detect independence even though the intermediate and uniform probabilities are different. Once more, the probabilities are similar with $\{1/6, 4/6, 1/6\}$ and $\{1/3, 1/3, 1/3\}$, respectively. More details on those methods are found in the following sub-subsection.

3.1.6 Power of intermediate and uniform probability distributions

The rejection rate of the intermediate and uniform probability distributions show different results for different methods, as discussed in subsubsection 3.1.5. The

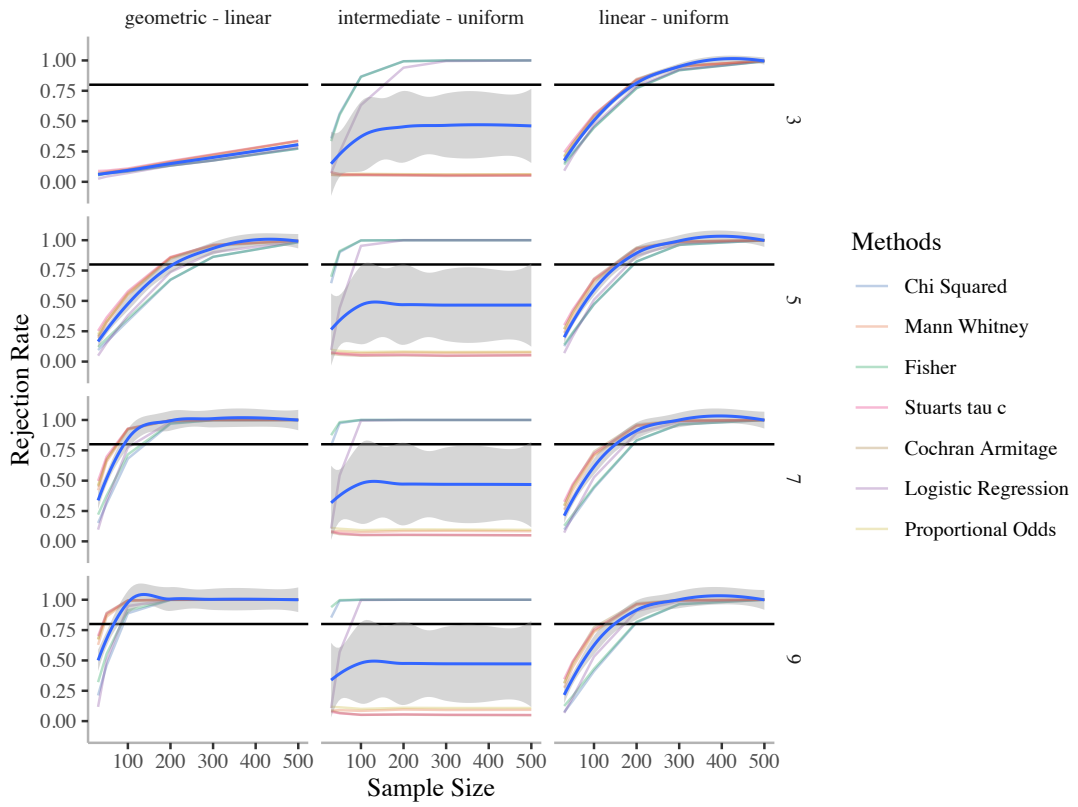

Figure 5: The rejection rates are displayed for three probability distributions across sample sizes and numbers of categories indicating discrepancies regarding the researcher degrees of freedom. The reversed allocation groups show no distinct differences. The smoothed rejection rate in blue was calculated through local polynomial regression along with a confidence band.

logistic regression model, the chi-squared test and Fisher's exact test of independence converge to a power of 100%, as seen with other probability distributions. However, the remaining methods display a low rejection rate as seen in Figure 6.

The rejection rate of the methods in Figure 6 is less than 0.12. Therefore, the rejection of the null hypothesis occurs in less than 12% of the time. The desired power of at least 80% for different probability distributions is not reached.

The discrepancy between the methods increases with a higher number of categories. Samples with three categories show narrow rejection rates of the methods considered in Figure 6 than samples with nine categories. Moreover, whether the treatment or control groups are uniformly distributed does not seem to affect the rejection rate strongly. The variation of the rejection rate in the methods seems similar to both probability combinations.

The proportional odds model shows the highest rejection rate of the four methods in Figure 6. The proportional odds model considers all cumulative logits simultaneously, which might be the reason for not detecting the increase in the intermediate category or simply not putting much weight on it. Out of the two regression models, the naïve logistic approach has detected the difference between the probability distributions. However, Figure 5 shows higher power with larger sample sizes. Considering that the logistic regression model is calculated $J - 1$

times, the increase in the intermediate category is represented in at least one of the dichotomous groups. Therefore, the logistic regression model seems to detect this difference very well.

Furthermore, the Cochran-Armitage test and the Stuarts τ_c seem to perform similarly badly, with rejection rates close to 0.05. Their performance does seem to be most robust among the displayed methods in Figure 6, with slightly higher variations towards a higher number of categories. Regarding Stuart's τ_c statistic that investigates concordant and discordant pairs, most pairs of individuals are probably yielding many ties with only a few concordant or discordant pairs, thus assuming independence. On the other hand, the Cochran-Armitage test probably does not detect the difference well because it uses the score statistic as the proportional odds model does. Therefore, the results might not put much weight on the increased category.

Lastly, the performance of the Mann-Whitney U test is also not satisfactory. The rejection rate increases with a higher number of categories. However, a power of below 12% is not desirable. The Mann-Whitney U test investigates a shift in location by counting the number of times that the control group's value precedes the treatment group's value. The shift in location can only occur in one direction. Hence, it is not considering the location shifting back. The control group preceding the treatment group might not happen often. Therefore it might be assumed that there is no shift in location.

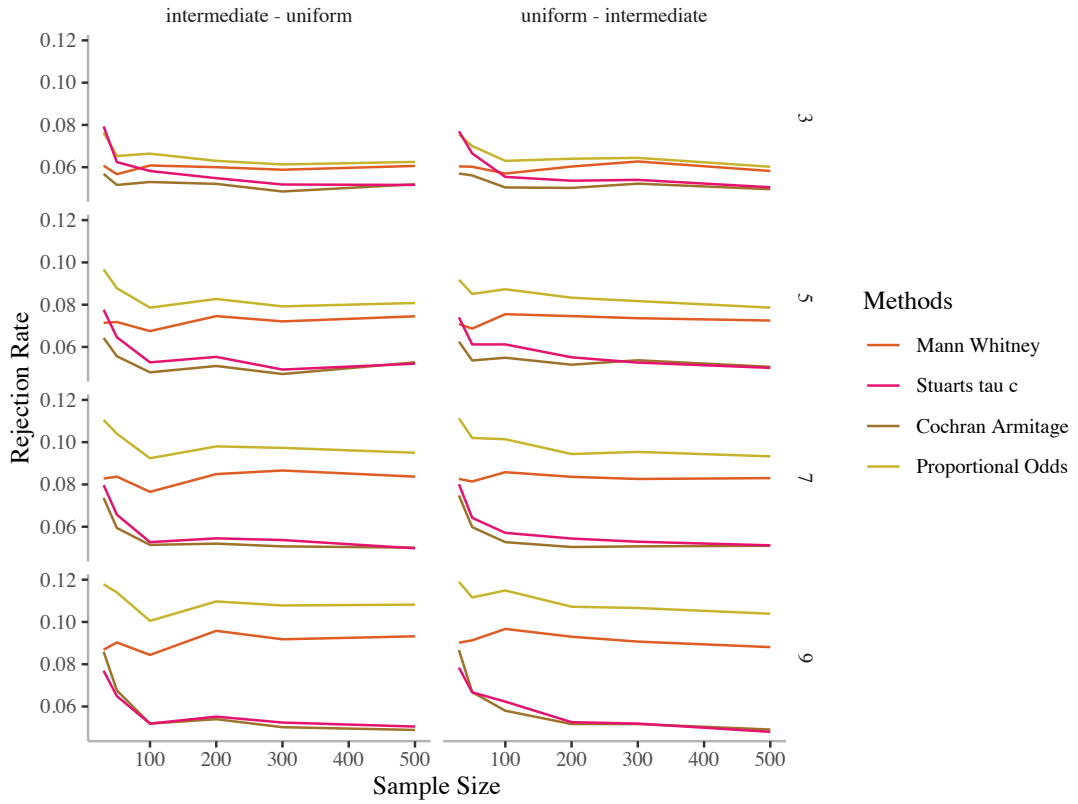

Figure 6: The combination of intermediate and uniform probability distributions display a low rejection rate for certain methods. The scenarios are represented across sample sizes and numbers of categories, where rejection rates are desired to be above 80%.

3.1.7 Additional results

The `coin` package by Hothorn et al. (2008) provided a function that enables an alternative calculation for several independence tests. However, the application could be more straightforward. The Mann-Whitney U test and the Cochran-Armitage could also be computed using the `independence_test()` function.

The Mann-Whitney U test shows no significant discrepancies in both functions. Figure 12 to Figure 14 in the appendix show rejection rates for both functions by type-I error, power and stroke probabilities. The mean difference in the rejection rate of both functions over all 600 combinations is 0.0015. Hence, either function is an appropriate choice for the combinations considered in this study.

Similarly, the Cochran-Armitage test does not show high discrepancies between the inbuilt function and the `coin` package implementation. Figure 15 to Figure 17 in the appendix illustrate the rejection rate for type-I error, power and the stroke example. The mean difference in the rejection rate for the Cochran-Armitage functions over all 600 combinations is 0.0025, which makes either function an appropriate choice.

3.2 Probabilities from previous studies

Generating data samples from parametric draws might be unrealistic. An RCT might yield different conclusions in practice. Hence, another way to generate data samples is by considering previous studies. The researched studies are based on the modified Rankin scale. The probabilities corresponding to each number of categories have been extracted from previous stroke trials. The probabilities are summarized in the appendix in Table 3.

Figure 7 illustrates the rejection rates for stroke probabilities in the treatment and control groups, where a local polynomial regression model calculated the smoothed rejection rate in blue. The smoothed rejection rates for the scenarios with five, seven and nine categories seem to increase with higher sample sizes. However, the rejection rate of the sample with three categories converges towards 0.05. The samples with three, five and nine categories display a rejection rate below 0.07 with a wide confidence band. Larger sample sizes suggest that the treatment would not be accepted, considering that the rejection rate would be too high for an acceptable type-I error and too low for a desired power. The rejection rate of the scenarios with seven categories represent a high rejection rate converging to 100%. A sample size of about 200 is sufficient to display an appropriate power.

According to the selected probabilities, the rejection rates increase with the sample size. This increase, however, is only beneficial for the sample with seven categories, which converges to 100%. The remaining categories depict rejection rates below 0.05 for small sample sizes, which would indicate an appropriate type-I error. In contrast, an increase in the sample size results in inconclusive rejection rates.

Another data-generating mechanism consideration is the combination of stroke probabilities and theoretical ones. The probabilities in the allocation groups of samples with three, five and nine categories are illustrated in Figure 18 in the appendix. The rejection rate for all those scenarios is high and converges towards

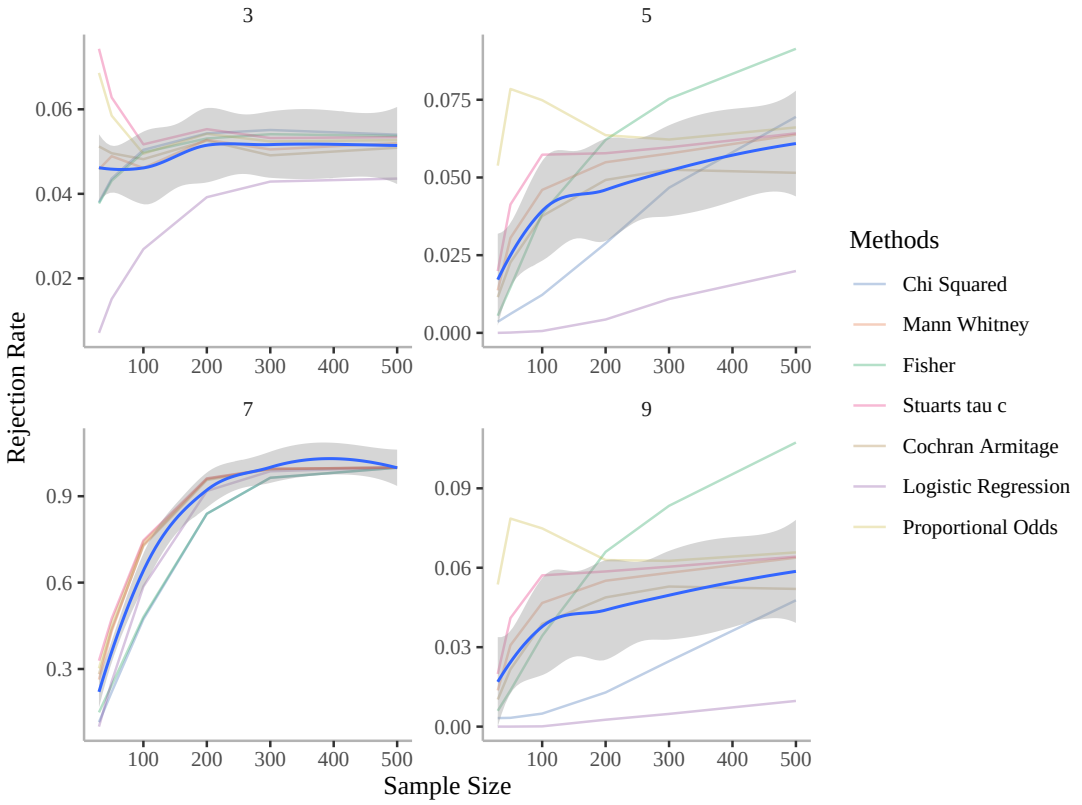

Figure 7: The rejection rates are depicted across sample sizes and numbers of categories for probabilities from previous stroke trials. The smoothed rejection rate in blue was calculated through local polynomial regression along with a confidence band.

100% with increasing sample size. The rejection rates do not fall below 80%, except for a few scenarios with low sample sizes with three categories, specifically for the logistic regression model. Therefore, the choice of small sample sizes in three categories might yield different conclusions, as well as choosing the logistic regression model for small sample sizes.

The samples considering seven categories illustrate a different picture in Figure 8. The probability combinations containing stroke and uniform and the linear-stroke probabilities depict a linear increase of the rejection rate with an increasing sample size finishing above 80%. However, a large sample size of 500 is required. Furthermore, the intermediate combinations depict a slower ascent starting below 80%, but reaching 100% with a sample size of 500. The most promising results seem to be the geometric and the stroke-linear combinations with a steep ascent and all methods reaching 100% at around 200 observations. Moreover, the stroke-geometric combination shows a rejection rate above 80% for all sample sizes.

In conclusion, using probabilities from previous studies illustrate more realistic results. Significantly, the samples with seven categories show that different probability distributions yield different results with higher sample sizes improving the rejection rate.

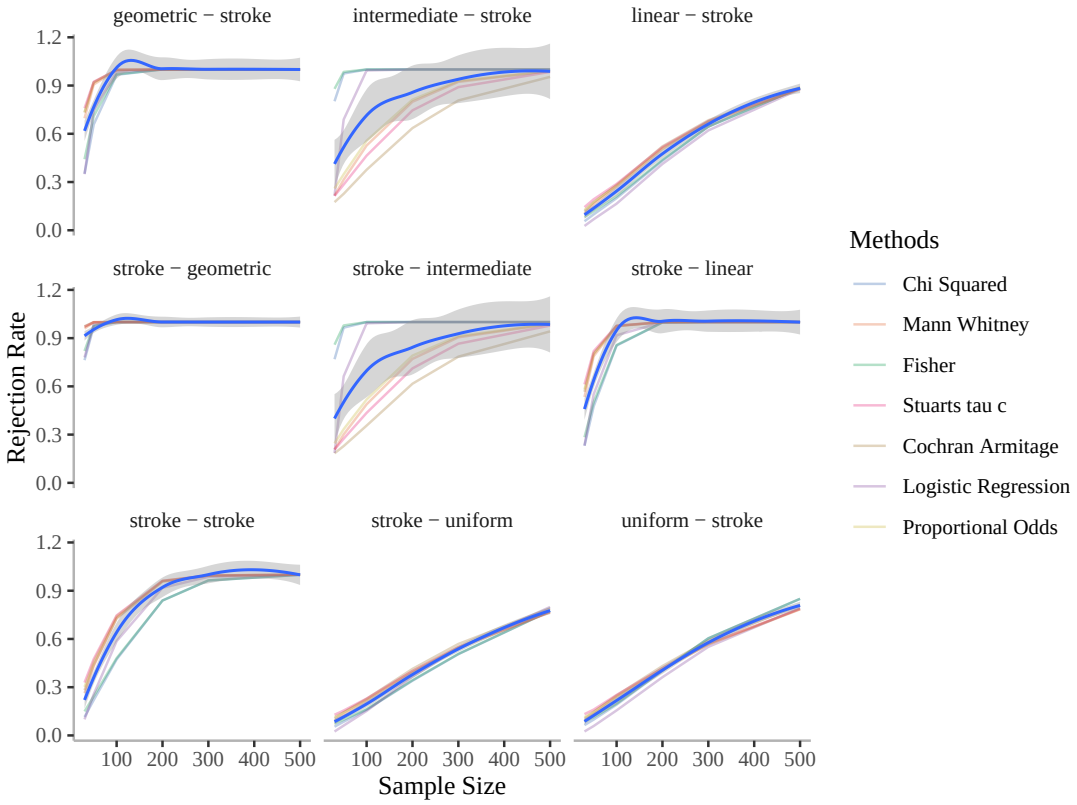

Figure 8: The probabilities from previous stroke trials and theoretical ones are depicted for the samples with seven categories across sample sizes. The smoothed rejection rate in blue was calculated through local polynomial regression along with a confidence band.

4 Discussion

Researchers might find themselves making many choices, often arbitrary, that involve the design and analysis of a study. Those choices are referred to as researcher degrees of freedom, and the results and conclusions can be affected by choices made throughout a study. Assuming that statistical methods are safe after only one comparative simulation study can be damaging, and the focus has yet to be on methodological research. However, Pawel et al. (2022) have shown that comparative simulation studies are included in questionable research practices. This simulation study investigated the effects of researcher degrees of freedom based on a randomised controlled trial with an ordinal outcome including several methods. The results of this simulation study indicate that choices in the study design can affect the conclusions drawn from research. The study aims to investigate researcher degrees of freedom based on a randomised controlled trial with an ordinal outcome variable. The simulations were conducted for 600 combinations, including the sample size, the number of categories and probability distributions of the categories' occurrences. Theoretical probabilities were chosen, and the more realistic probabilities from previous studies.

The rejection rate regarding the theoretical probabilities of the considered methods showed that the intermediate and uniform combination in the probability distribution illustrates high discrepancies in the methods. While the logistic regression model, the chi-squared test of independence and Fisher's exact test for count data detected the difference in the ordinal data distribution, the remaining methods struggled to perform accurately. These results suggest that the probability distributions in the allocation groups influence the conclusion drawn from similar probabilities especially involving a change in one category. Moreover, the choice of method in this situation is crucial in the interpretation of the outcome of a study. Additionally, concerning the power investigation in Figure 4 and Figure 5, the increase in sample size provides more robust rejection rates. With a higher number of categories, the rejection rates start higher and, in some scenarios, do not fall below 80%. Therefore, the sample size and the number of categories affect the rejection rate of all methods.

Furthermore, the rejection rate of the methods regarding the type-I error indicates decent performances hovering around 0.05. However, slight changes can be detected with an increasing number of categories. Figure 2 depicts a higher variation in the rejection rates with an increase in the number of categories.

Additionally, the function from the `coin` package by Hothorn et al. (2008) is indifferent to the inbuilt function for the considered methods. Further, the continuity correction of the chi-squared test does not seem applicable in this study. However, further research might be helpful when investigating small sample sizes.

As for the probabilities taken from previous stroke trials, no high discrepancies are discovered for samples with three, five and nine categories. The samples with seven categories depicted various rejection rates across the probability distributions. Generally, the rejection rates increase with higher sample sizes. However, some scenarios needed a sample size of 500 to reach at least 80%, while others reached 100% with 200 observations. These probabilities depicted the variability in real data samples quite well.

In conclusion, the sample size, the number of categories, the probability dis-

tribution and the choice of methods can affect the performance and whether the authorities would accept a treatment.

The researcher degrees of freedom illustrated in this thesis add to the state of the art in simulation studies. The benchmark studies conducted by Nießl et al. (2022), Ullmann et al. (2023), and Jelizarow et al. (2010) have discovered that the choice of data set accounts for a large part of the variability in those studies. This simulation study is not based on a specific data set. Nevertheless, the effects of data generated in different settings are apparent even in simulation studies. Furthermore, the sample size variation has been shown to provide more robust results for high sample sizes, which coincides with the findings of Ullmann et al. (2023). In addition to the results by Pawel et al. (2022), this thesis illustrated researcher degrees of freedom focusing on altering the data-generating process specifically for ordinal data analysis, expanding on questionable research practices within the generation of data samples.

Besides, simulation studies are not bound to the availability of data sets. Hence the limitation regarding data samples is subject to the computing power. This study was conducted with 10000 simulations per combination, resulting in 6000000 data samples. A laptop computer with eight CPU cores would have taken around a month to compute these many data samples. Such high computing times might be due to Fisher's exact test for count data that increase drastically with higher sample sizes and number of categories. For high-scale simulation studies, higher computing power might be required. An increase in the number of iterations in the simulation study yields more robust results and provides better insight into the variation of the rejection rates.

Furthermore, a small sample size and higher numbers of categories might lead to data samples with only one category. For instance, in the stroke trial, one of nine categories occurs in more than 90% of the observations. An analysis of one category is impossible, so the p-value for that data sample is not available. The rejection rate is then calculated according to the remaining data samples. However, another sample could be drawn until 10000 data samples are available, which requires even more computing time.

As the results have shown, similar probabilities in treatment and control groups might yield conflicting rejection rates in the methods considered in this study. To further explore the discrepancies with changes in one category, investigating those probabilities might provide more information. Furthermore, these particular researcher degrees of freedom were somewhat arbitrarily chosen. Therefore, investigating different probabilities, higher numbers of categories and different sample sizes might be beneficial.

Moreover, probabilities from previous studies provide more realistic results and contribute information for further studies. However, in the search for probabilities in ordinal data, many journals failed to report the distribution of their ordinal outcome. The investigation of available probabilities of previous studies might be worthwhile.

5 Conclusion

This simulation study investigated potential researcher degrees of freedom in methodological research based on a randomised controlled trial with an ordinal outcome variable. The performance of the methods considered in this thesis has been shown to be affected by the sample size, the number of categories in the ordinal outcome and the probability distribution in the allocation variable. Similar probability distributions depict high discrepancies in the methods' rejection rate. Moreover, probabilities from previous studies provide more realistic information for further studies.

List of Figures

List of Tables

References

Agresti, A. (2002). *Categorical data analysis.* New York: Wiley.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, *11*, 375–386.

Bath, P. M., Woodhouse, L. J., Appleton, J. P., Beridze, M., Christensen, H., Dineen, R. A., ... Sprigg, N. (2018). Triple versus guideline antiplatelet therapy to prevent recurrence after acute ischaemic stroke or transient ischaemic attack: The tardis rct. *Health Technology Assessment*, *22*.

Boulesteix, A.-L. (2010). Over-optimism in bioinformatics research. *Bioinformatics*, *26*, 437–439.

Boulesteix, A.-L., Hoffmann, S., Charlton, A., & Seibold, H. (2020). A replication crisis in methodological research? *Significance*, *17*, 18–21.

Boulesteix, A.-L., Wilson, R., & Hapfelmeier, A. (2017). Towards evidence-based computational statistics: Lessons from clinical research on the role and design of real-data benchmark studies. *BMC Medical Research Methodology*, *17*, 1–12.

Cochran, W. G. (1954). Some methods for strengthening the common χ 2 tests. *Biometrics*, *10*, 417–451.

Goyal, M., Demchuk, A. M., Menon, B. K., Eesa, M., Rempel, J. L., Thornton, J., ... Hill, M. D. (2015). Randomized assessment of rapid endovascular treatment of ischemic stroke. *New England Journal of Medicine*, *372*, 1019–1030.

Harrell Jr, F. E. (2023). rms: Regression modeling strategies [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=rms (R package version 6.6-0)

Hollander, M., Wolfe, D. A., & Chicken, E. (2013). *Nonparametric statistical methods.* New Jersey: John Wiley & Sons.

Hothorn, T., Hornik, K., van de Wiel, M. A., & Zeileis, A. (2008). Implementing a class of permutation tests: The coin package. *Journal of Statistical Software*, *28*, 1–23. (R package version 1.4.2) doi: 10.18637/jss.v028.i08

Jelizarow, M., Guillemot, V., Tenenhaus, A., Strimmer, K., & Boulesteix, A.-L. (2010). Over-optimism in bioinformatics: An illustration. *Bioinformatics*, *26*, 1990–1998.

Langhorne, P., Wu, O., Rodgers, H., Ashburn, A., & Bernhardt, J. (2017). A very early rehabilitation trial after stroke (avert): A phase iii, multicentre, randomised controlled trial. *Health Technology Assessment*, *21*, 1–119.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, *18*, 50–60.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *42*, 109–127.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*, 2074–2102.

Musiol, S. (2023). *Researcher Degrees of Freedom in Methodological Simulation Studies: An Illustration from Medical Statistics.* Retrieved from https://github.com/SarahM95/RDOF doi: 10.5281/zenodo.1234

Nießl, C., Herrmann, M., Wiedemann, C., Casalicchio, G., & Boulesteix, A. L. (2022). Over-optimism in benchmark studies and the multiplicity of design and analysis options when interpreting their results. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *12*. doi: 10.1002/widm.1441

Pawel, S., Kook, L., & Reeve, K. (2022). Pitfalls and potentials in simulation studies. *arXiv preprint arXiv:2203.13076*.

Pearson, K. (1900). X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *50*, 157–175.

Selman, C. J., Lee, K. J., & Mahar, R. K. (2022). Statistical analyses of ordinal outcomes in randomised controlled trials: protocol for a scoping review. *arXiv preprint arXiv:2208.06154*.

Signorell, A. (2023). Desctools: Tools for descriptive statistics [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=DescTools (R package version 0.99.48)

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

Stuart, A. (1953). The estimation and comparison of strengths of association in contingency tables. *Biometrika*, *40*, 105–110.

Student. (1908). The probable error of a mean. *Biometrika*, *6*, 1–25.

Ullmann, T., Beer, A., Hünemörder, M., Seidl, T., & Boulesteix, A.-L. (2023). Over-optimistic evaluation and reporting of novel cluster algorithms: an illustrative study. *Advances in Data Analysis and Classification*, *17*, 211–238.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society*, *54*, 426–482.

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*, 1832.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*, 80–83.

Yates, F. (1934). Contingency tables involving small numbers and the χ 2 test. *Supplement to the Journal of the Royal Statistical Society*, *1*, 217–235.

Yousefi, M. R., Hua, J., Sima, C., & Dougherty, E. R. (2010). Reporting bias when using real data sets to analyze classification performance. *Bioinformatics*, *26*, 68–76.

A Study Design and Methods

This simulation study utilises probability distributions from previous work. The probabilities haven been sourced from stroke trials based on the modified Rankin scale for categories three, five, seven and nine shown in Table 3.

	Bath et al. (2018)		Goyal et al. (2015)		Bath et al. (2018)		Langhorne et al. (2017)	
Categories	Treatment	Control	Treatment	Control	Treatment	Control	Treatment	Control
1	0.94	0.93	0.15	0.07	0.94	0.93	0.76	0.75
2	0.02	0.03	0.21	0.01	0.02	0.03	0.14	0.15
3	0.003	0.003	0.18	0.12	0.01	0.01	0.1	0.1
4	0.006	0.009	0.16	0.15	0.02	0.02	-	-
5	0.008	0.007	0.13	0.24	0.01	0.005	-	-
6	0.006	0.008	0.07	0.12	-	-	-	-
7	0.004	0.005	0.1	0.19	-	-	-	-
8	0.003	0.001	-	-	-	-	-	-
9	0.008	0.005	-	-	-	-	-	-

Table 3: Probability distributions have been sourced from various stroke trials for three, five, seven and nine categories.

B Results

Besides the results discussed in section 3, further figures have been added to illustrate more details in the studies results.

B.1 Type-I Error for the Regression Models

The results corresponding to the type-I error have been discussed for the logistic regression model and the proportional odds model. The differences between both models are shown in Figure 9.

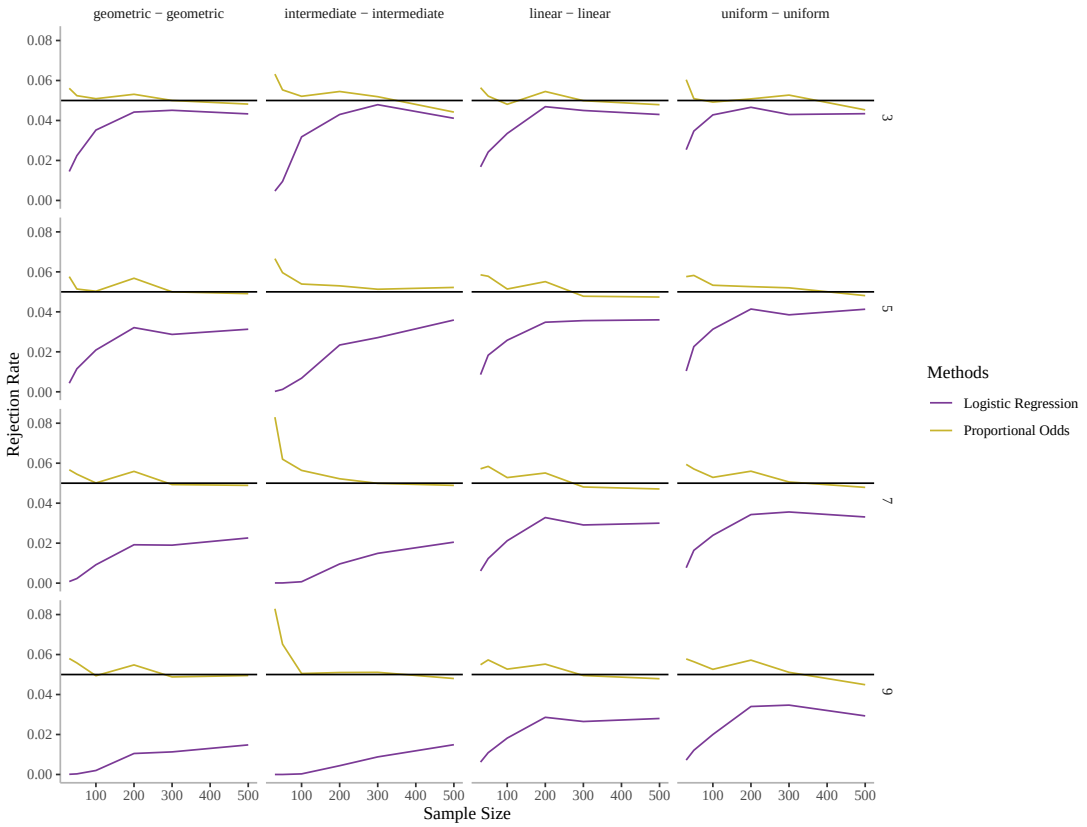

Figure 9: The rejection rate is depicted for the logistic regression model and the proportional odds model. The type-I error is represented across the sample sizes and numbers of categories.

Further, McCullagh (1980) has claimed that for $2 \times J$ contingency tables the hypothesis test for the proportional odds model is equivalent to the Mann-Whitney test. These results are depicted in Figure 10.

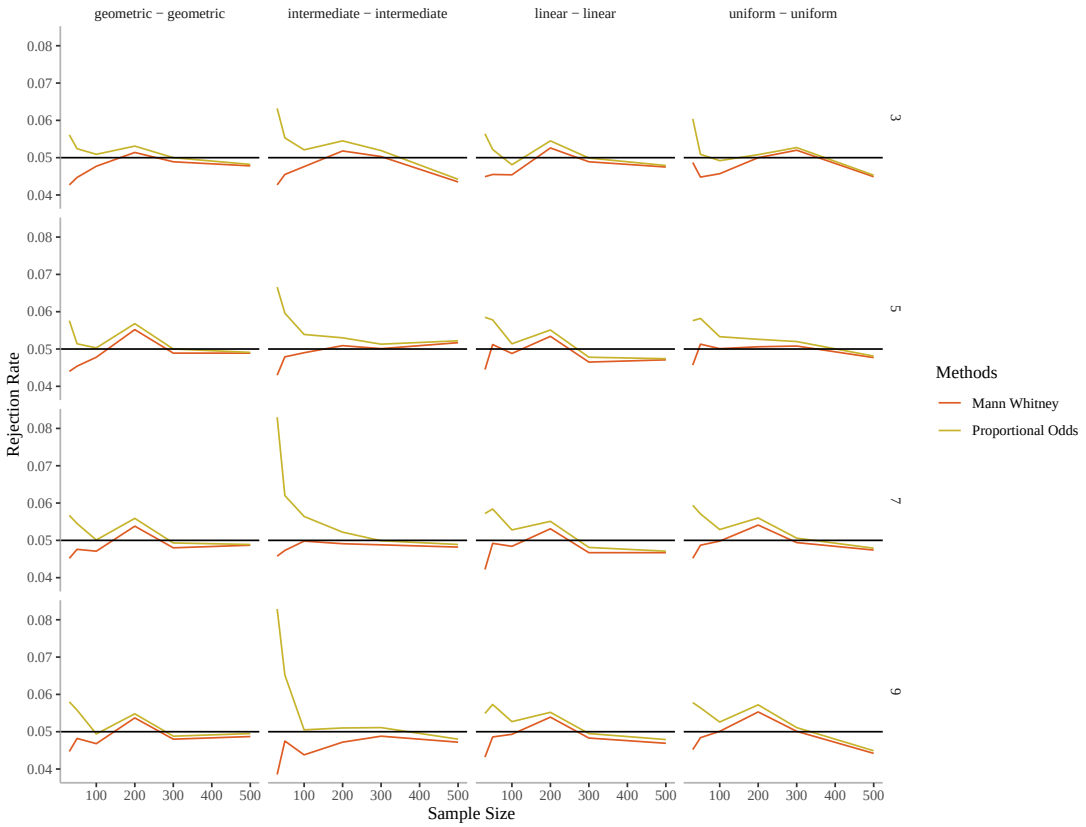

Figure 10: The rejection rates for the proportional odds model and the Mann-Whitney test depict minor differences in the type-I error. Small sample sizes show higher discrepancies.

B.2 Overall Power

Many combinations in the probability distributions exist, therefore the illustration has been restricted to one set of probabilities. No distinct differences in the probabilities in the allocation variables were detected, therefore the remaining combinations are depicted in Figure 11.

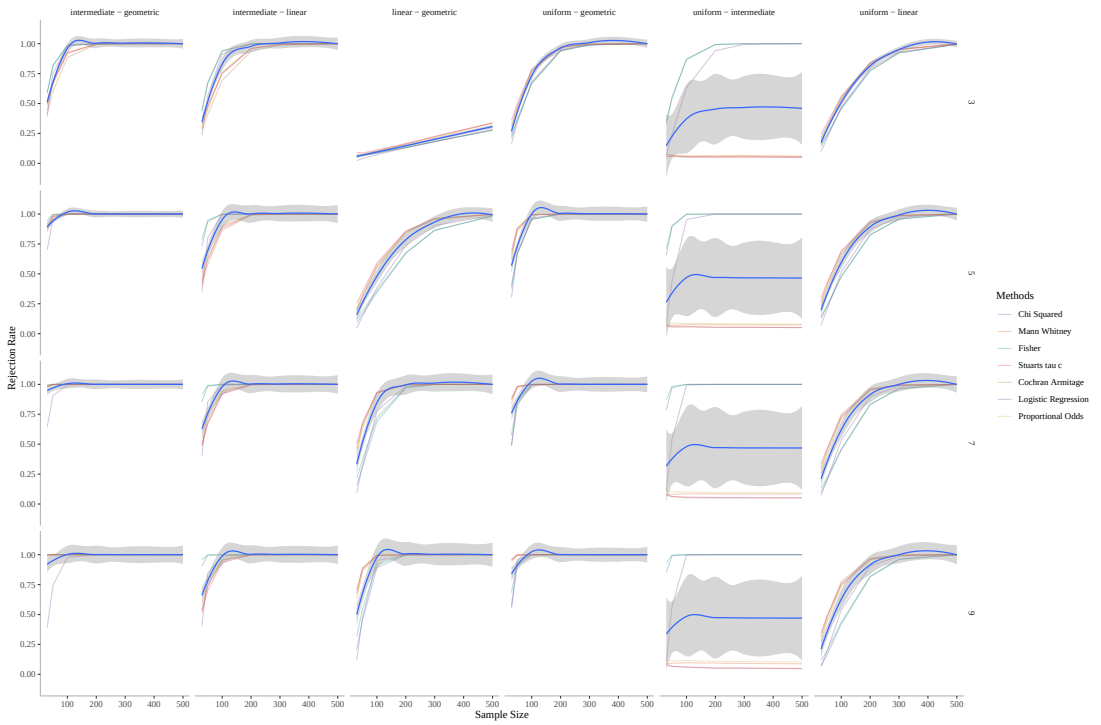

Figure 11: The rejection rates are depicted for theoretical probabilities regarding the power across the sample sizes and numbers of categories. The smoothed rejection rate in blue was calculated through local polynomial regression along with a confidence band.

B.3 Mann-Whitney U test

The Mann-Whitney U test is computed using the inbuilt function and the function from the `coin` package. The differences are shown for theoretical distributions in Figure 12 and Figure 13, whereas the rejection rates for the probabilities from previous studies are depicted in Figure 14.

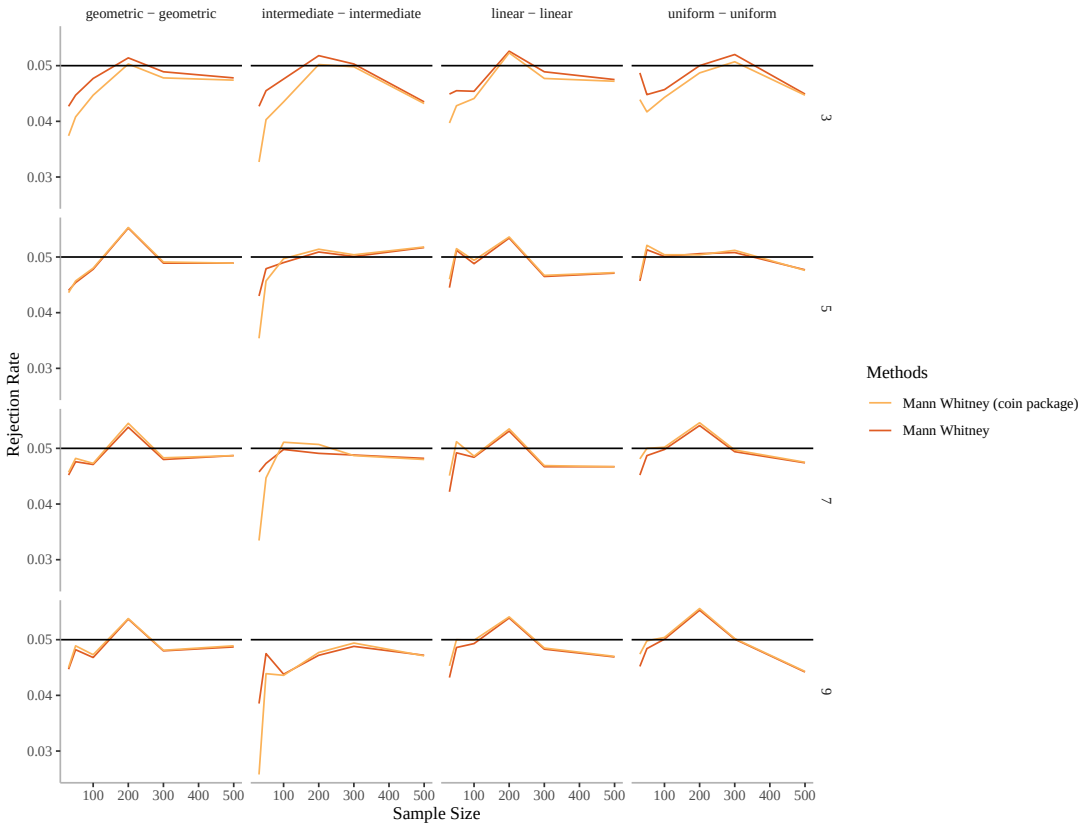

Figure 12: The rejection rates for the Mann-Whitney U test are computed regarding the type-I error across sample sizes and numbers of categories.

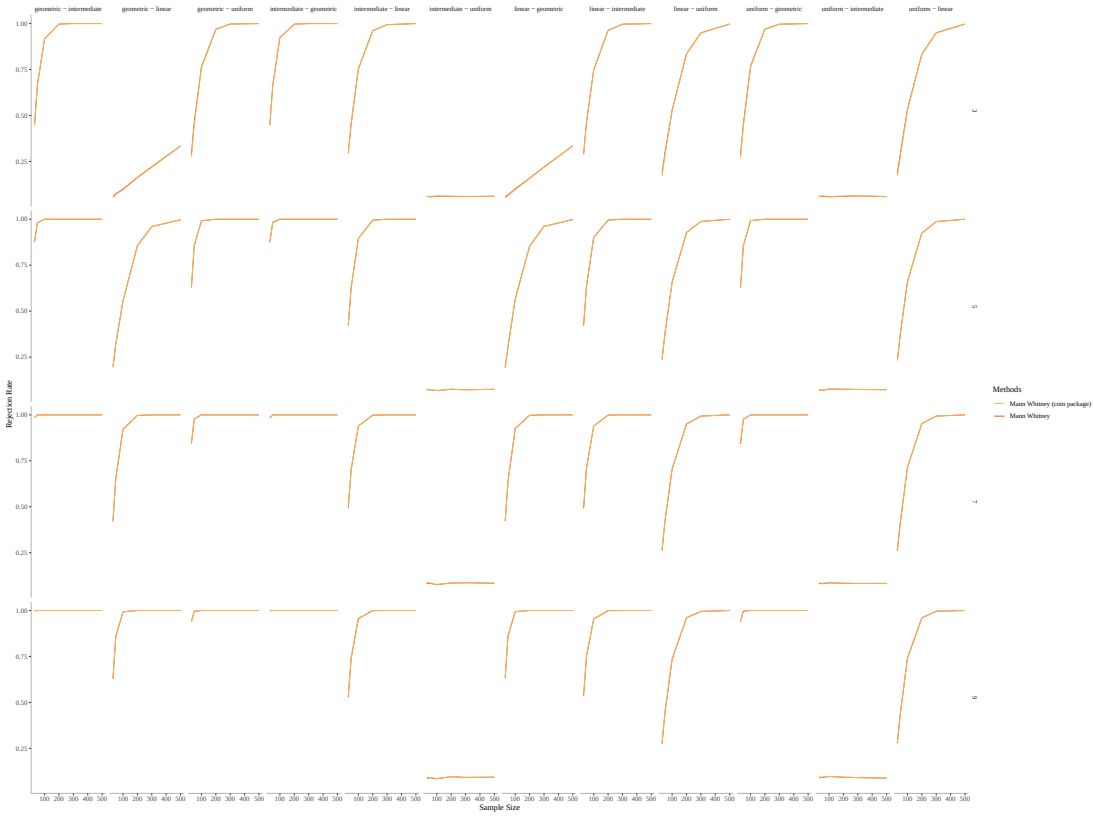

Figure 13: The rejection rates for the Mann-Whitney U test are computed regarding the power across sample sizes and numbers of categories.

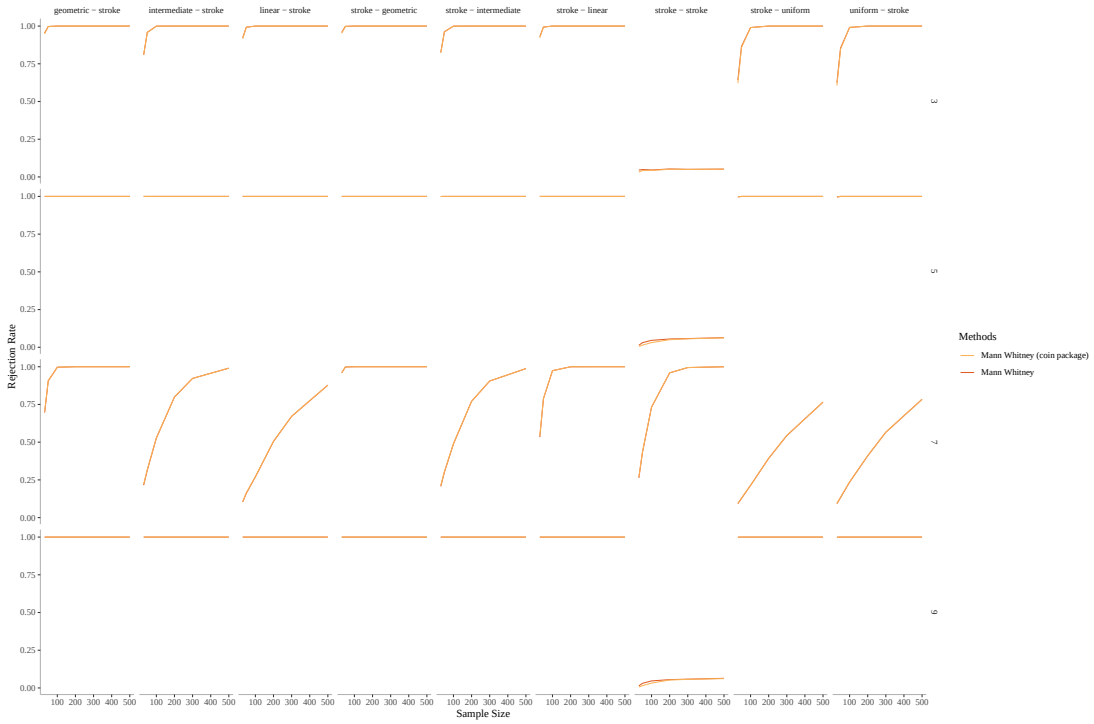

Figure 14: The rejection rates for the Mann-Whitney U test are computed across sample sizes and numbers of categories for probabilities from previous stroke trials.

B.4 Cochran-Armitage test

The `coin` package provided a function for the Cochran-Armitage test, which is compared to the inbuilt function for theoretical distributions in Figure 15 and Figure 16. The differences in both functions for the probabilities from previous studies are shown in Figure 17.

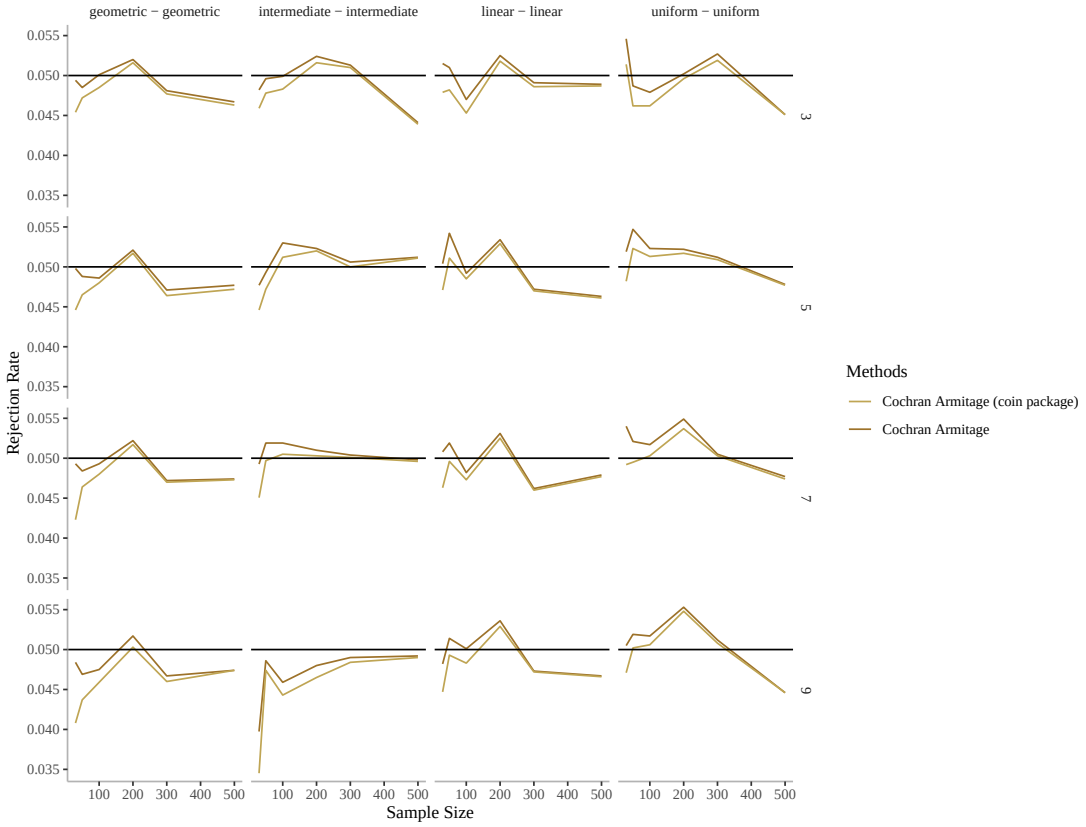

Figure 15: The rejection rates for the Cochran-Armitage test are computed regarding the type-I error across sample sizes and numbers of categories.

Figure 16: The rejection rates for the Cochran-Armitage test are computed regarding the power across sample sizes and numbers of categories.

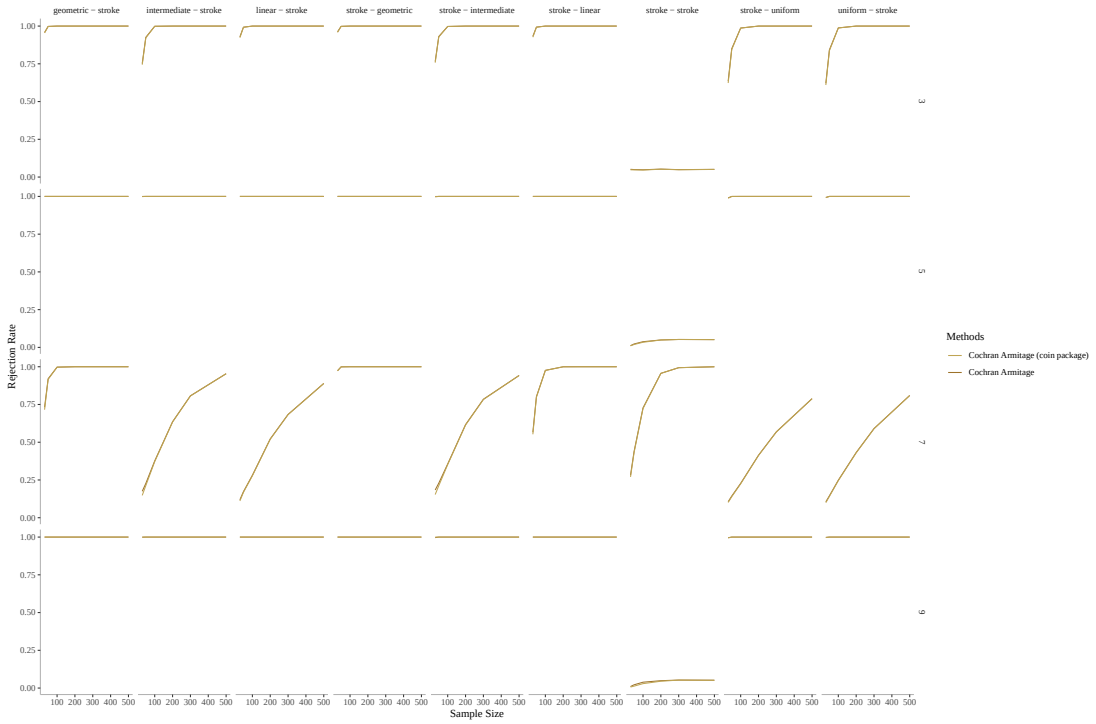

Figure 17: The rejection rates for the Cochran-Armitage test are computed across sample sizes and numbers of categories for probabilities from previous stroke trials.

B.5 Probabilities from previous studies

The rejection rates for probabilities from previous stroke trials were discussed in section 3 for samples with seven categories. The remaining categories display very similar rejection rates across the sample sizes and probability distributions shown in Figure 18.

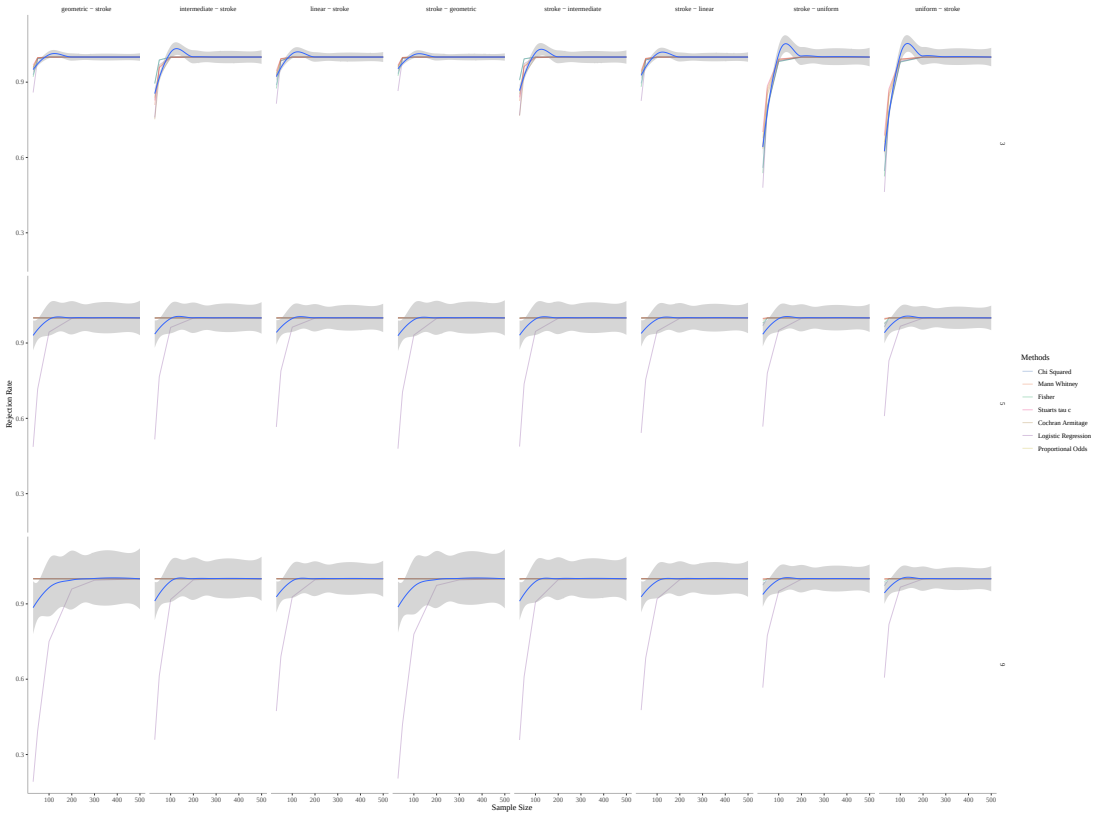

Figure 18: The smoothed rejection rate in blue was calculated through local polynomial regression along with a confidence band. The scenarios are represented for the probabilities from previous stroke trials across sample sizes and three, five, and nine categories.

Erklärung zur Masterarbeit

Hiermit versichere ich, dass die vorliegende Arbeit von mir selbstständig verfasst wurde und dass keine anderen als die angegebenen Quellen und Hilfsmittel benutzt wurden. Diese Erklärung erstreckt sich auch auf in der Arbeit enthaltene Graphiken, Zeichnungen, Kartenskizzen und bildliche Darstellungen.

Master's thesis statement of originality

I hereby confirm that I have written the accompanying thesis by myself, without contributions from any sources other than those cited in the text and acknowledgements. This applies also to all graphics, drawings, maps and images included in the thesis.

.....	.....
Ort und Datum	Unterschrift
Place and date	Signature