Pim Keer

Hypothesis Testing in Contingency Tables

A Discussion, and Exact Unconditional Tests for $r \times c$ Tables

Master's thesis to obtain the degree of MSc Applied Mathematics
Responsible Supervisor: Prof.dr. G. Jongbloed
TU Delft Supervisor: Dr. H. P. Lopuhaä
NTNU Supervisor: Dr. Ø. Bakke
Thesis Committee: Dr. H. P. Lopuhaä, Dr. C. Kraaikamp, Dr. Ø. Bakke
Delft, June 2023

Delft University of Technology
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft Institute of Applied Mathematics

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

ABSTRACT

Every time one counts the number of occurrences of a pair of values for two categorical variables, one obtains a contingency table. These tables are one of the simplest representations of data in order to statistically test for the presence of some association between the two variables under consideration. Although naturally occurring in so many scientific disciplines, there is still a lot of debate on the appropriate way to perform tests of significance on these contingency tables.

Especially when one wants to use exact methods, i.e., methods that are based on the exact probabilities of observing the table of interest, there is great disagreement on which marginal totals one should treat as fixed for inference. This has led to the development of the conditional tests, most famously Fisher's exact test, and unconditional tests, of which Barnard's CSM test was the first example. Mostly due to philosophical objections and computational challenges, the unconditional test has received far less attention over the years. This is especially true for contingency tables with more than 2 rows or columns. To our knowledge, there are no implementations available of exact unconditional tests for these larger tables.

The aim of this text is two-fold. First, we give a historical account on the rivalry between conditional and unconditional test, and argue that there is a case to be made to research exact unconditional methods in greater depth. Second, we will present implementations of exact unconditional tests that are applicable to general $r \times c$ contingency tables. Some of these implementations are generalisations of existing methods for the 2×2 table, such as Barnard's CSM test, with some additions in order to increase the computational efficiency. In addition, we also introduce a new approach that translates the classical Neyman–Pearson procedure of constructing a critical region for a given significance level α into a a mixed integer linear programming problem. The latter can be solved efficiently with one of many existing optimisation software packages.

This will eventually lead to a power study comparing 14 different tests, of which 12 unconditional ones, for different table dimensions and marginal totals. Although no test comes out as most powerful in every situation, the tests using a linear programming formulation have comparable, and often higher power than the classical unconditional approaches. This comes at a cost however, the critical regions produced via this optimisation approach are not guaranteed to be nested, i.e., they are not necessarily contained in each other for increasing values of α. This limits their use and interpretability. Further research should point out whether additional requirements can be formulated that would make the critical regions nested, while still keeping the advantages of the linear programming formulation.

PREFACE

You are about to start your reading of my Master thesis, written in order to obtain the degree of Master of Science in Applied Mathematics from the Delft University of Technology. However, this thesis has not been written in Delft. Instead, from January to June 2023, I have been fortunate enough to do my research at the Norwegian University of Science and Technology. It is in the beautiful Norwegian city of Trondheim that I have been working on the topic of hypothesis testing in contingency tables.

Of course, it was partially a certain spirit of adventure that has drawn me to Norway. The short winter days covered in snow back in January, and now the 4-hour twilight which is intended to serve as night, made for an ever-changing environment to work in. However, the research topic made it even more attractive to come down here. This is because the problem setting of testing significance in contingency tables can intuitively be explained in around a minute. Nevertheless, diving deeper into the different testing approaches, it becomes clear that behind this simple object hide both a very rich mathematical theory, and a long history of debate on the appropriateness of each testing method. Because of this two-sided character, which will clearly be visible in the text you are about to read, there was never a dull moment when writing this thesis. On one hand, there was always the possibility to work on interesting and challenging mathematical statistics. On the other hand, I could always switch and read up on the many philosophical considerations that can go unnoticed when one is only writing code or handling equations. This is something I would not have expected beforehand to enjoy so much.

This thesis made me realise even more that statistics is not just probability, but also causality, experimental design and philosophy. I also realised, with every article I read, that I am far from an expert on the matter. However, this thesis certainly helped me believe that I will graduate as a more well-rounded statistician, ready to learn more.

It is for these reasons that I would like to say a very sincere "*Tusen hjertelig takk*" to my supervisor at NTNU, Dr. Øyvind Bakke. He proposed this topic to me, giving me the opportunity to work on it in Trondheim. I appreciated the great amount of freedom I received along the way to work out ideas in greater depth. In our weekly Tuesday meetings, Øyvind always provided helpful feedback, and managed to help me out with the many questions I had. Furthermore, I want to express my gratitude to my supervisor in Delft, Dr. Rik Lopuhaä. Apart from our online meetings, there was no way to notice the geographical separation between us, as I was always welcome to ask questions, and received indispensable feedback

on my work. I am also grateful to Prof.Dr. Geurt Jongbloed, who as the responsible supervisor, assisted in many of the more administrative matters along the way. Finally, from the bottom of my heart, I thank my parents and brother Casper for their support during these months abroad, but also during all the years before that, long before I knew I would be where I am today.

Enjoy the reading!
Kos deg med lesingen!

Pim Keer
Trondheim, 1st of June 2023

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER

ONE

INTRODUCTION

In order to test the effectiveness of a new medicine for a certain disease, researchers have devised the following experiment. A number $n_{\cdot\cdot}$ of patients suffering from the disease have been split up at random into two groups. Group 1 is made up of $n_{1\cdot}$ patients who will receive the new medicine, while group 2 consists of $n_{2\cdot}$ patients who will not. After a predetermined testing period, the researchers will record for each of the two groups how many patients have recovered. It turns out that x_{11} persons in group 1 got better, and x_{21} persons in group 2. Although this a gross simplification of how a real medical trial would be performed, it introduces the main object of study of this text: the contingency table. Indeed, the researchers can summarise the results of their experiment as in Table 1.1. We call it a success if, after the testing period, a patient has healed. In total, $n_{\cdot 1}$ out of the $n_{\cdot\cdot}$ recovered.

	Success	No Success	
Group 1	x_{11}	x_{12}	$n_{1\cdot} = x_{11} + x_{12}$
Group 2	x_{21}	x_{22}	$n_{2\cdot} = x_{21} + x_{22}$
	$n_{\cdot 1} = x_{11} + x_{21}$	$n_{\cdot 2} = x_{12} + x_{22}$	$n_{\cdot\cdot} = n_{1\cdot} + n_{2\cdot} = n_{\cdot 1} + n_{\cdot 2}$

Table 1.1: A 2×2 contingency table.

Based on this table, researchers would want to conclude whether or not the medicine has an influence on the healing process (for now it is of no interest whether this effect is positive or negative for recovery). In other words, is there an association between recovering from the illness and receiving the medicine?

As we will see in Chapter 2, numerous statistical methods exist to answer this question whenever the table has 2 rows and 2 columns. We will split these up into methods using asymptotic approximations and methods based on the exact probability distributions found in the table. The latter can furthermore be subdivided into conditional and unconditional methods, based on which marginal totals are considered fixed. From the description of our medical experiment, it may seem logical that the $(n_{1\cdot}, n_{2\cdot})$-margin should be seen as fixed. However, it is less clear whether or not one should fix the $(n_{\cdot 1}, n_{\cdot 2})$-margin too. Interestingly, this seemingly simple problem has been, since its inception, surrounded by controversy and debate. We will try to navigate the vast literature that has been generated around this in Chapter 3.

Let us consider another example. This thesis has mostly been written while visiting the Norwegian University of Science and Technology (NTNU) in Trondheim, Norway. This university offers three different ways in which one can graduate as a mathematician. First of all, there is the – in the Netherlands common – path of taking a 3-year bachelor programme in Mathematical Sciences, followed by a 2-year master programme in Mathematical Sciences. Alternatively, one can choose for a 5-year teacher education. Finally, the technical universities may also award the legally protected title of "sivilingeniør" (which should not be confused with the term "civil engineer" in English) after completion of a 5-year programme focusing on mathematics and physics. One example of a question that might come up is whether the choice of programme a mathematics student makes is somehow linked to that student's gender. Dr. Bakke kindly provided us with data from the Department of Mathematical Sciences at NTNU from the year 2007, showing how many students were enrolled in each study programme that year, together with the male-to-female ratio. This data is shown in Table 1.2.

	3y. BSc + 2y. MSc	5y. teacher	5y. siv. ing.	
Male	4	0	23	27
Female	0	2	5	7
	4	2	28	34

Table 1.2: Male/female distribution across the different mathematics programmes at NTNU in 2007.

We can ask a similar question to the one we asked with Table 1.1, How can we find out if there is a significant – we will later see what that means formally – difference in the choice of programme between the male and female student populations?

Table 1.2 consists of three columns. This is not an uncommon phenomenon. Indeed, we can easily think of ways in which Table 1.1 from our first, medical experiment can be extended. Instead of having just two patient groups, the researchers might have chosen to introduce a third group which received a placebo, for example. Furthermore, they might have recorded something else than just a binary response of healed / not healed. After the testing period, the researchers could have encountered patients who have fully recovered, or not recovered at all, while others still have some symptoms, or even are experiencing some detrimental side effects.

Depending on the experiment we are doing, we can encounter tables with any number of rows and columns. Although a study of the 2×2 contingency table only would already be able to fill numerous theses, we will discuss in Chapter 4 how the statistical methods from Chapter 2 can be applied to the general $r \times c$ contingency table. Both asymptotic methods and exact conditional methods have been studied in great depth, but the exact unconditional methods have received far less attention. We will encounter some of the reasons why this may be. Furthermore, we will introduce a couple of alternative methods that try to address some of the problems that come up when generalising methods from the 2×2 case to larger tables. In Chapter 5, we will assess how well the proposed alternatives perform in comparison to the existing approaches with regard to computation time and statistical power. The alternative methods certainly come with their own

set of considerations and difficulties, which we will discuss in Chapter 6. Finally, conclusions will be drawn in Chapter 7.

CHAPTER

TWO

TESTS ON 2×2 CONTINGENCY TABLES

Many different solution methods exist to determine whether an association is present in a contingency table, but they can all ultimately be labelled as either asymptotic or exact methods. The exact methods can furthermore be divided into conditional and unconditional methods. We will first take a look at the asymptotic approach. This is the classical approach to hypothesis testing on contingency tables. Via the asymptotic test, we will also reiterate some fundamental concepts of hypothesis testing, as well as introduce some notions we will need later on. However, we will shift our focus rather quickly to the exact methods. These will turn out to be more suited in small-sample situations, which are not uncommon in medical and biological contexts. Contingency tables are very natural objects in both settings [1]. The discrimination between conditional and unconditional methods has led to a – yet to be settled – debate. In Chapter 3, we will try to summarise this discussion as accurately as possible. In particular, we will pay attention to the many interesting and nuanced, statistical and philosophical, concepts and arguments that divide statisticians up to this day.

2.1 The Classical Neyman–Pearson Approach

The main idea behind any asymptotic test is to come up with some test statistic, of which we know the asymptotic distribution under the null hypothesis. For large enough sample sizes, this (often more simple) asymptotic distribution will be a good approximation for the (often more complex) actual distribution of the test statistic. We can then decide whether to reject the null hypothesis based on how probable the value of the test statistic is, under this asymptotic distribution under the null hypothesis.

In the context of our 2×2 contingency table, the test statistic will be a function of the quantities in that table. However, if we intend to determine an asymptotic distribution, we have to agree on some assumptions on these table quantities. Many such assumptions are possible, but here we will treat $n_{1\cdot}$ and $n_{2\cdot}$ as fixed, while we treating x_{11} and x_{21} as realisations of two random variables, X_{11} and X_{21} respectively. Note that knowledge of x_{11} and x_{21}, together with $n_{1\cdot}$ and $n_{2\cdot}$ fully determines the table. In order to say something useful about the asymptotic distribution of the test statistics which will follow, we furthermore assume that X_{11} and

X_{21} are binomially distributed with parameters $(n_{1\cdot}, \theta_1)$ and $(n_{2\cdot}, \theta_2)$ respectively, where $\boldsymbol{\theta} := (\theta_1, \theta_2) \in \Theta := [0,1]^2$ are of course unknown. This assumption allows to rewrite the original research question in the format of a statistical hypothesis test $H_0 \colon \boldsymbol{\theta} \in \Theta_0$ against $H_1 \colon \boldsymbol{\theta} \in \Theta_1$, where $\Theta_0 = \{(\theta_1, \theta_2) \in \Theta : \theta_1 = \theta_2\}$ and $\Theta_1 = \Theta \setminus \Theta_0$, or alternatively

$$H_0 \colon \theta_1 = \theta_2 = \theta, \quad H_1 \colon \theta_1 \neq \theta_2, \tag{2.1}$$

for some unknown $\theta \in [0,1]$. The null hypothesis can be interpreted as the recovery probability of a patient in group 1 being equal to the recovery probability of a patient in group 2. That is, the recovery probability does not depend on the patient receiving the medicine or not. The alternative hypothesis states that the probabilities of recovery are different per group, i.e., the medicine has an effect on the healing process.

Although this binomial assumption seems quite natural, one can argue this is actually quite a strong claim to make. Indeed, recall that for a binomial trial, one should need a set of independent Bernoulli experiments with the same probability of success in each experiment. In the context of our medical example, it could be that the reactions of different patients to the medicine are not entirely independent; perhaps some patients are relatives. Besides that, the probability of recovery of a given patient, in so far that we can speak of one, might not be the same for each patient. We come back to this in Section 3.6.

That being said, assuming that X_{11} and X_{21} have a binomial distribution anyway, there is a wide array of possible test statistics to choose from. One of the more popular test statistics is Pearson's chi-square test statistic [2]. As a general goodness-of-fit test, it is perfectly suited for the setting of contingency tables. The test statistic is constructed as the sum over all 4 table cells of the quantity $(O-E)^2/E$, where O is the observed number of occurrences (so X_{11}, X_{21}, $X_{12} := n_{1\cdot} - X_{11}$ and $X_{22} := n_{2\cdot} - X_{21}$), and E the expected number of occurrences under the null hypothesis (so $n_{1\cdot}\theta$, $n_{2\cdot}\theta$, $n_{1\cdot}(1-\theta)$, and $n_{2\cdot}(1-\theta)$ respectively). Since θ is unknown, we will estimate it with the maximum likelihood estimator $\hat{\theta} = (X_{11} + X_{21})/(n_{1\cdot} + n_{2\cdot})$. Later on, we will spend more time on the different ways to get rid of this unknown θ. Concretely, the above description gives the following test statistic:

$$\chi^2(X_{11}, X_{21}) := \sum_{i=1}^{2} \left(\frac{\left(X_{i1} - n_{i\cdot}\hat{\theta}\right)^2}{n_{i\cdot}\hat{\theta}} + \frac{\left(X_{i2} - n_{i\cdot}(1-\hat{\theta})\right)^2}{n_{i\cdot}(1-\hat{\theta})} \right), \tag{2.2}$$

which can be written more compactly as

$$\chi^2(X_{11}, X_{21}) = \frac{(X_{11}X_{22} - X_{12}X_{21})^2 n_{\cdot\cdot}}{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}} \tag{2.3}$$

[3]. From this expression it is easy to see that whenever $n_{\cdot 1}$ or $n_{\cdot 2}$ is zero, the denominator is zero and the chi-square test statistic is undefined. If this is the case, we define $\chi^2(X_{11}, X_{21}) = -\infty$. For completeness, we mention yet another representation of writing this chi-square statistic, also often used in the literature, referred to as the squared Z-pooled statistic [4], or squared Z statistic with pooled

variance estimator.

$$Z_{\mathrm{p}}^2(X_{11}, X_{21}) := \frac{\left(\frac{X_{11}}{n_{1\cdot}} - \frac{X_{21}}{n_{2\cdot}}\right)^2}{\frac{X_{11}+X_{21}}{n_{1\cdot}+n_{2\cdot}}\left(1 - \frac{X_{11}+X_{21}}{n_{1\cdot}+n_{2\cdot}}\right)\left(\frac{1}{n_{1\cdot}} + \frac{1}{n_{2\cdot}}\right)}. \tag{2.4}$$

It is also worth mentioning that there exists a Z statistic with unpooled variance estimator as well, given by

$$Z_{\mathrm{u}}^2(X_{11}, X_{21}) := \frac{\left(\frac{X_{11}}{n_{1\cdot}} - \frac{X_{21}}{n_{2\cdot}}\right)^2}{\frac{1}{n_{1\cdot}} \cdot \frac{X_{11}}{n_{1\cdot}}\left(1 - \frac{X_{11}}{n_{1\cdot}}\right) + \frac{1}{n_{2\cdot}} \cdot \frac{X_{21}}{n_{2\cdot}}\left(1 - \frac{X_{21}}{n_{2\cdot}}\right)}. \tag{2.5}$$

We will encounter this variant later on. It can be shown, under the further assumptions that X_{11} and X_{21} are independent, that $\chi^2(X_{11}, X_{21}) = Z_{\mathrm{p}}^2(X_{11}, X_{21})$ converges in distribution to a chi-square distribution with one degree of freedom under the null hypothesis, whenever $n_{1\cdot}, n_{2\cdot} \to \infty$ such that the ratio $n_{2\cdot}/n_{1\cdot}$ converges to a positive constant [5]. That is, $n_{1\cdot}$ and $n_{2\cdot}$ grow to infinity at the same rate.

Based on this convergence result, it is straightforward to construct a test. The chi-square test consists of computing the value of the test statistic $\chi^2(x_{11}, x_{21})$ for the observed value of $\mathbf{X} = (X_{11}, X_{21})$. The larger the value of $\chi^2(x_{11}, x_{21})$, the less probable it becomes that this value could have been observed from a random sample from the χ_1^2-distribution [1]. We therefore reject the null hypothesis at a predetermined significance level $\alpha \in [0, 1]$ if $\chi^2(x_{11}, x_{21})$ lies in the critical region $K^\alpha := [\chi^2_{1,1-\alpha}, \infty)$, where $\chi^2_{1,1-\alpha}$ is the $1-\alpha$-quantile of the χ_1^2-distribution. This is of course the classical Neyman–Pearson approach to hypothesis testing. We first fix the significance level $\alpha \in [0, 1]$, choose an appropriate test statistic $T(\mathbf{X})$, in this case $\chi^2(\mathbf{X})$, and then choose the critical region K^α such that the probability of making a Type I error (wrongly rejecting H_0) is at most α, i.e. $\sup_{\boldsymbol{\theta} \in \Theta_0} \beta(\boldsymbol{\theta}) \leq \alpha$, where $\beta(\boldsymbol{\theta}) := P_{\boldsymbol{\theta}}(T(\mathbf{X}) \in K^\alpha)$ is the power function. This function is ideally 0, if $\boldsymbol{\theta} \in \Theta_0$, and 1, if $\boldsymbol{\theta} \in \Theta_1$. This is of course never possible due to randomness. Ensuring that the probability of making a Type I error is always bounded by α, we can then make K^α as large as possible in order to minimise the probability of a Type II error (wrongly not rejecting H_0). Finally, as a way to quantify the evidence against H_0, we can define the p-value corresponding to an observation $\mathbf{x} = (x_{11}, x_{21})$ as

$$p_T(\mathbf{x}) := \inf\{\alpha \in [0, 1] : T(\mathbf{x}) \in K^\alpha\}, \tag{2.6}$$

which the smallest significance level for which we would reject H_0 [6].

We can also look at the Neyman–Pearson approach to hypothesis testing slightly differently. Instead of constructing the critical region for a given significance level and rejecting H_0 if the observed value of the test statistic lies within

[1] One can convince oneself that $\chi^2(x_{11}, x_{21})$ will be large whenever $x_{11}/n_{1\cdot}$ and $x_{21}/n_{2\cdot}$ are far apart, indicating that there might be a difference between θ_1 and θ_2. This also explains why we set $\chi^2(x_{11}, x_{21}) = -\infty$ whenever $\chi^2(x_{11}, x_{21})$ is undefined. Realise that this only happens when we have $x_{11} = x_{21} = 0$ or $x_{11} = n_{1\cdot}$ and $x_{21} = n_{2\cdot}$. In both cases, the corresponding table does not provide strong evidence against the null hypothesis, and so we want the test statistic value to be as small as possible in order to avoid rejection when we observe such a table. Note that we might just as well have defined $\chi^2(x_{11}, x_{21})$ as any negative number instead.

this critical region, we can also design tests based on the p-value. This perspective will turn out rather useful later on. Given an observation $\mathbf{x}$, we can equivalently define the p-value as

$$p_T(\mathbf{x}) := P_{H_0}(T(\mathbf{X}) \geq T(\mathbf{x})), \tag{2.7}$$

in the case that we would reject the null hypothesis for large values of T (as is the case with the chi-square test). We would now reject H_0 if $p_T(\mathbf{x}) \leq \alpha$. This observation allows us to think of the p-value itself as a test statistic (and consequently of $p_T(\mathbf{X})$ as a random variable). We can thus turn our original test (T, K^α) into a *p-value test* $(p_T, [0, \alpha])$, where we reject H_0 whenever $p_T(\mathbf{x}) \leq \alpha$. In order for this p-value test to have significance level α, we should require the p-value to be valid.

Definition 2.1. A p-value $p(\mathbf{X})$ is called valid if, for all $\alpha \in [0, 1]$ and all $\boldsymbol{\theta} \in \Theta_0$,

$$P_\theta(p(\mathbf{X}) \leq \alpha) \leq \alpha \tag{2.8}$$

[7].

Notice the abuse of notation we will employ throughout Chapter 2: $P_\theta(A)$ indicates the probability of event A under the assumption that $\theta_1 = \theta_2 = \theta$. Saying that (2.8) should hold for all $\boldsymbol{\theta} \in \Theta_0$ is thus the same as saying it should hold for all $\theta \in [0, 1]$. A p-value that is valid automatically has a number of nice properties which we will state the moment we need them.

Apart from that, the p-value test perspective is very convenient in the discrete setting of contingency tables. As long as we are able to rank the table outcomes from less probable to most probable under the null hypothesis (whatever that may mean), we can compute the p-value of an observation $\mathbf{x}$ as the sum of the occurrence probabilities of each table outcome more extreme than $\mathbf{x}$. As we will see later, this is the main idea behind the unconditional tests. Before we get there though, let us first see how we can improve on the bad approximation by the asymptotic testing approach.

The most evident objection against the classical, asymptotic approach is that the approximation by the asymptotic distribution might not be justified when the samples remain small. This might in particular lead to the probability of making a Type I error being larger than α, or in terms of the p-value: the p-value not being valid. The poor asymptotic approximation is illustrated by Brataas [8], who mentions that when the group sizes shrink to around 5, the approximation breaks down. A similar order of magnitude is indicated by Fisher [9], who states that, as a rule of thumb, one should not employ the asymptotic test whenever the expected number in each cell of the table is less than 5. Yates [10] investigated how well the χ^2 distribution approximated the exact distribution, which we will derive in the next Section. He noted that the "discrepancies are primarily due to the fact that χ is a continuous distribution, whereas the distribution it is endeavouring to approximate is discontinuous". To this end, Yates suggests to use the continuity-corrected test statistic

$$\chi_c^2(X_{11}, X_{21}) := \frac{\left(|X_{11}X_{22} - X_{12}X_{21}| - \frac{1}{2}N\right)^2 n_{..}}{n_{1.}n_{2.}n_{.1}n_{.2}}, \tag{2.9}$$

and shows by examples that this continuity correction leads to a far better approximation of the exact distribution, in particular for tables where $\theta_1 = \theta_2 = \theta$

close to 1/2. As we will see in Chapter 5, the chi-square test with Yates' continuity correction will perform very similarly to the exact test we will see in the next Section.

2.2 The Exact Conditional Approach: Fisher's Exact Test

For small sample sizes, a more suitable alternative to the chi-square test with Yates' continuity correction might be to use so-called exact methods. They are exact in the sense that they work with the actual distributions appearing in the problem, without any large-sample approximations. We will now give a slightly quicker derivation of Fisher's exact test, which he originally described in his *Statistical Methods for Research Workers* [9].

In the setting of Table 1.1, making the binomial assumption, Fisher argues that under the null hypothesis that $\theta_1 = \theta_2 = \theta$, the probability of x_{11} successes in group 1 is

$$\binom{n_{1\cdot}}{x_{11}}\theta^{x_{11}}(1-\theta)^{x_{12}}. \tag{2.10}$$

Mutatis mutandis, the same expression can be written for the probability of x_{21} successes in group 2. Consequently, Fisher states that the probability of observing Table 1.1 is simply the product

$$P_\theta(\mathbf{X} = \mathbf{x}) = \binom{n_{1\cdot}}{x_{11}}\binom{n_{2\cdot}}{x_{21}}\theta^{n_{\cdot 1}}(1-\theta)^{n_{\cdot 2}} \tag{2.11}$$

In his derivation, Fisher implicitly assumed that the $(n_{1\cdot}, n_{2\cdot})$-margin is fixed, something which need not be the case when looking solely at Table 1.1 without any further context, as we will see in Section 2.3.

Fisher now realised that in order to get rid of the unknown parameter θ, the factor $\theta^{n_{\cdot 1}}(1-\theta)^{n_{\cdot 2}}$ should be the same for all tables with the same $(n_{\cdot 1}, n_{\cdot 2})$-margin. Therefore, the probability of a given table outcome $(X_{11} = x_{11}, X_{21} = x_{21})$ (note how this fully determines the table if $(n_{1\cdot}, n_{2\cdot})$ is fixed), conditional on the value of the $(n_{\cdot 1}, n_{\cdot 2})$-margin and under the null hypothesis, is given by

$$\begin{aligned}
&P_\theta(X_{11} = x_{11}, X_{21} = x_{21} \mid X_{11} + X_{21} = n_{\cdot 1}) \\
&\quad = \frac{P_\theta(X_{11} = x_{11}, X_{21} = n_{\cdot 1} - x_{11})}{P_\theta(X_{11} + X_{21} = n_{\cdot 1})} \\
&\quad = \frac{\binom{n_{1\cdot}}{x_{11}}\binom{n_{2\cdot}}{n_{\cdot 1}-x_{11}}\theta^{n_{\cdot 1}}(1-\theta)^{n_{\cdot\cdot}-n_{\cdot 1}}}{\binom{n_{\cdot\cdot}}{n_{\cdot 1}}\theta^{n_{\cdot 1}}(1-\theta)^{n_{\cdot\cdot}-n_{\cdot 1}}} \\
&\quad = \frac{\binom{n_{1\cdot}}{x_{11}}\binom{n_{2\cdot}}{n_{\cdot 1}-x_{11}}}{\binom{n_{\cdot\cdot}}{n_{\cdot 1}}}.
\end{aligned} \tag{2.12}$$

In the second equality, we used that the sum of two binomially distributed random variables with number of trials n and m and with the same success probability θ is again binomial with number of trials $n + m$ and success probability θ.

Notice how $P_\theta(X_{11} = x_{11}, X_{21} = x_{21} \mid X_{11} + X_{21} = n_{\cdot 1})$ no longer depends on θ, so we can just write $P(X_{11} = x_{11}, X_{21} = x_{21} \mid X_{11} + X_{21} = n_{\cdot 1})$ instead. In other

words, the test statistic $T(\mathbf{X}) = X_{11} + X_{21}$ is sufficient for θ. This could have also been seen directly by the factorisation theorem. Indeed,

$$P_\theta(\mathbf{X} = \mathbf{x}) = g_\theta(T(\mathbf{x}))h(\mathbf{x}) = \theta^{T(\mathbf{x})}(1-\theta)^{n_{\cdot\cdot} - T(\mathbf{x})} \cdot \binom{n_{1\cdot}}{x_{11}}\binom{n_{2\cdot}}{x_{21}}.$$

The conditional exact test now uses the probability (2.12), which is the probability mass function of the hypergeometric distribution with parameters $(n_{\cdot\cdot}, n_{1\cdot}, n_{\cdot 1})$, for all further inference.

The issue of how to define p-values in the case of the two-sided test (2.1) comes with an entire history of discussion of its own [3]. For a two-sided p-value we want to somehow add the null probabilities of tables that are more extreme than or just as extreme as the observed table. However, since the null distribution is the hypergeometric distribution, which is discrete and potentially asymmetric, how we can we define these more extreme tables? Several definitions exist based on how one answers this question [11]. However, all these definitions will have the following form:

$$p_F(\mathbf{x}) = p_F(x_{11}, x_{21}) = \sum_{i \in E_\mathbf{x}} P(X_{11} = i \mid X_{11} + X_{21} = n_{\cdot 1}), \tag{2.13}$$

where $n_{\cdot 1} = x_{11} + x_{21}$, where $P(X_{11} = i \mid X_{11} + X_{21} = n_{\cdot 1})$ is given by (2.12) and where the set $E_\mathbf{x}$ of outcomes contains all table outcomes for x_{11} that can be seen as more extreme than x_{11}. It is this set $E_\mathbf{x}$ that varies across definitions.

One possibility is to define $E_\mathbf{x}$ as the set of all table outcomes which have a null probability that is smaller than or equal to the null probability of $\mathbf{x} = (x_{11}, x_{21})$. That is,

$$E_\mathbf{x} := \{i \in \{0, \ldots, n_{1\cdot}\} : T(i, n_{\cdot 1} - i) \leq T(x_{11}, x_{21})\},$$

where

$$T(x_{11}, x_{21}) = P(X_{11} = x_{11} \mid X_{11} + X_{21} = n_{\cdot 1})$$

This is the definition used in the `R` implementation of Fisher's exact test. Note that we can view T as a test statistic for Fisher's exact test, such that we reject the null hypothesis when observing $\mathbf{x}$ such that $T(\mathbf{x})$ is small.

An alternative has been proposed by Mehrorta, Chan and Berger [12]. A table outcome would now be defined as more extreme than $\mathbf{x}$ if the null probability of ending up as far or further down the tail of the distribution (either in the left or the right tail) than that outcome is smaller than that same probability for the outcome $\mathbf{x}$. More concretely, we again have

$$E_\mathbf{x} := \{i \in \{0, \ldots, n_{1\cdot}\} : T(i, n_{\cdot 1} - i) \leq T(x_{11}, x_{21})\},$$

but now with

$$T(x_{11}, x_{21}) := P(X_{11} \leq x_{11} \mid X_{11} + X_{21} = n_{\cdot 1}) \wedge P(X_{11} \geq x_{11} \mid X_{11} + X_{21} = n_{\cdot 1}). \tag{2.14}$$

Once again, we can view T as a test statistic for which small values will lead to rejection of H_0. The set $E_\mathbf{x}$ is thus nothing else than the collection of tables which have a test statistic value smaller than or equal to the test statistic value of the observed $\mathbf{x}$, for some test statistic. With this in mind, it is possible to show an important result we will use later on, namely that Fisher's exact test produces a valid p-value.

Proposition 2.2. The Fisher p-value defined in (2.13) is valid.

To show this, we will make use of the following result, which is stated as Theorem 8.3.27 and proven in Casella and Berger [7].

Lemma 2.3. Let $T(\mathbf{X})$ be a test statistic such that large values of T give evidence that H_1 is true. For each sample point $\mathbf{x}$,

$$p(\mathbf{x}) := \sup_{\boldsymbol{\theta} \in \Theta_0} P_{\boldsymbol{\theta}}(T(\mathbf{X}) \geq T(\mathbf{x})) \tag{2.15}$$

is a valid p-value.

This lemma is very useful, as it directly proves the validity of a whole class of p-values, so-called supremum p-values or more formally p-values obtained via the supremum method (where we maximise with respect to the nuisance parameter). In Section 2.4, when we want to get rid of the unknown parameter θ without conditioning, we will make use of such p-values.

Proof of Proposition 2.2. Using the test statistic for Fisher's exact test defined in (2.14), the Fisher p-value (2.13) for a table outcome $\mathbf{x} = (x_{11}, x_{21})$ can be written as

$$\begin{aligned} p_F(x_1, x_2) &= \sum_{i \in E_{x_{11}}} P(X_{11} = i \mid X_{11} + X_{21} = n_{\cdot 1}) \\ &= \sum_{T(i, n_{\cdot 1} - i) \leq T(\mathbf{x})} P(X_{11} = i \mid X_{11} + X_{21} = n_{\cdot 1}) \\ &= P(T(\mathbf{X}) \leq T(\mathbf{x}) \mid X_{11} + X_{21} = n_{\cdot 1}). \end{aligned}$$

Clearly, since we no longer have a dependence on θ, we can write

$$p_F(x_1, x_2) = \sup_{\theta \in [0,1]} P(T(\mathbf{X}) \leq T(\mathbf{x}) \mid X_{11} + X_{21} = n_{\cdot 1}).$$

Although Lemma 2.3 only speaks of test statistics which reject H_0 for large values of the test statistic, it trivially generalises to test statistics $T(\mathbf{X})$ that reject H_0 for small values too, by using $-T(\mathbf{X})$ in the proof. Also conditioning on $X_{11} + X_{21} = n_{\cdot 1}$ does not alter the result, $T(\mathbf{X})$ merely has a different distribution than the unconditional distribution, but the validity of this supremum p-value still holds. We conclude that $p_F(\mathbf{X})$ is a valid p-value, given that $X_{11} + X_{21} = n_{\cdot 1}$. That is, for all $\alpha \in [0, 1]$ and all $\theta \in \Theta_0$,

$$P_\theta(p_F(\mathbf{X}) \leq \alpha \mid X_{11} + X_{21} = n_{\cdot 1}) \leq \alpha.$$

Of course, this is true for all $n_{\cdot 1} = 0, \ldots, n_{\cdot\cdot}$. Hence, this "conditional validity" can easily be shown to lead to the unconditional validity of $p_F(\mathbf{X})$. Indeed, for all $\alpha \in [0, 1]$ and all $\theta \in \Theta_0$, we have by the law of total probability that

$$\begin{aligned} P_\theta(p_F(\mathbf{X}) \leq \alpha) &= \sum_{n_{\cdot 1}=0}^{n_{\cdot\cdot}} P_\theta(p_F(\mathbf{X}) \leq \alpha \mid X_{11} + X_{21} = n_{\cdot 1}) P_\theta(X_{11} + X_{21} = n_{\cdot 1}) \\ &\leq \sum_{n_{\cdot 1}=0}^{n_{\cdot\cdot}} \alpha P_\theta(X_{11} + X_{21} = n_{\cdot 1}) \\ &= \alpha, \end{aligned}$$

where we used the conditional validity of $p_F(\mathbf{X})$ in the second line. □

2.3 Intermezzo: Contingency Tables as Outcomes from Urn Experiments

The key observation to make with Fisher's exact test is that by considering both margins of the table fixed, we remove any dependence on θ. Barnard remarks that based on the contingency table alone, it is possible to form several different abstract pictures [13].

2.3.1 The Independence Trial

One of these is the one implicitly assumed by Fisher's exact test. To put it in Barnard's words, this test corresponds to an experiment where we have $n_{1\cdot}$ balls marked with 1, and $n_{2\cdot}$ balls marked with 2. These $n_{\cdot\cdot} = n_{1\cdot} + n_{2\cdot}$ balls are put into an urn. Afterwards, they are withdrawn randomly from the urn and placed one by one in a row of $n_{\cdot\cdot}$ boxes, $n_{\cdot 1}$ of which are marked by "Success" and $n_{\cdot 2}$ by "No Success". Realise that for this experiment to be executed it is necessary to know both $(n_{1\cdot}, n_{2\cdot})$ and $(n_{\cdot 1}, n_{\cdot 2})$ beforehand. With this picture in mind, the probability of Table 1.1 occurring is, just as we saw with the derivation of Fisher's test, equal to

$$\frac{\binom{n_{1\cdot}}{x_{11}}\binom{n_{2\cdot}}{x_{21}}}{\binom{n_{\cdot\cdot}}{n_{\cdot 1}}}. \tag{2.16}$$

Indeed, the numerator is the exact number of ways in which we can choose x_{11} out of the $n_{1\cdot}$ balls marked 1 and x_{21} out of the $n_{2\cdot}$ balls marked 2, which we divide by the denominator; the total number of ways in which we can label $n_{\cdot 1}$ out of $n_{\cdot\cdot}$ boxes with the label "Success". This is nothing else than the probability of observing Table 1.1, conditional on the fact that we should have $n_{\cdot 1}$ "Success" boxes. Barnard called this experiment the 2×2 independence trial. Notice that although (2.16) is the same expression as (2.12), the way we reached those expressions was different in both cases. In the derivation of (2.12), we assumed the existence of the constant success probabilities θ_1 and θ_2 in the two respective groups. However, we made no such assumption in order to derive (2.16). As Barnard puts it in a later paper,

> "The 2×2 independence trial, by contrast, is concerned with a situation where we have a collection of experimental units allocated at random into two categories and we observe whether or not a feature such as "cured" arises just as often in the one category as in the other." [14]

The introduction of the success probabilities θ_1 and θ_2 was mostly done in order to illustrate how Fisher's exact test does not have to deal with any nuisance parameters. However, it is clear from the above that we do not need to assume such constant success probabilities in order to use Fisher's exact test. We will return to this in Section 3.5.

2.3.2 The Double Dichotomy

Another abstract picture we can draw, called the double dichotomy by Barnard, consists of one urn, containing balls with two labels each. The first label is either

1 or 2, and the second label is either "A" or "B". We drop the "Success" / "No Success" labels for this setting in order to stress that contingency tables could also arise from an experiment where we are testing the association of two properties, instead of comparing the amount of successes between two groups. An example could be testing whether there is a relation between sex and handedness, in which case "Success" / "No Success" is a less appropriate labelling. If we suppose there is a "very large" number of balls in the urn, and the proportion of balls labelled "1A" (or "1B", "2A", "2B") is given by θ_{1A} (or θ_{1B}, θ_{2A}, θ_{2B}), the probability of observing Table 1.1 is then just given by the multinomial expression

$$\frac{n_{..}!}{x_{1A}!x_{1B}!x_{2A}!x_{2B}!}\theta_{1A}^{x_{1A}}\theta_{1B}^{x_{1B}}\theta_{2A}^{x_{2A}}\theta_{2B}^{x_{2B}}, \tag{2.17}$$

where we should substitute $(x_{11}, x_{12}, x_{21}, x_{22})$ by $(x_{1A}, x_{1B}, x_{2A}, x_{2B})$. Note that this picture corresponds to fixing none of the table margins beforehand. We have no idea what $x_{1A} + x_{1B}$ and $x_{1A} + x_{2A}$ will be before we start the experiment. Again, realise that we are making a similar binomial (or in this case multinomial) assumption, by considering a very large number of balls. Like so, withdrawing a number of balls from the urn (without replacement!) does not alter the proportions θ_{1A}, θ_{1B}, θ_{2A} or θ_{2B} significantly during the experiment.

2.3.3 The Comparative Trial

Finally, it should not come as a surprise that the final abstract picture one can draw from the contingency table is the one where exactly one table margin is fixed. This received the name of "2×2 comparative trial" by Barnard. In this case the fixed margin is always the sample size margin, i.e. the $(n_{1\cdot}, n_{2\cdot})$-margin. This picture corresponds to the medical experiment we described at the start of this chapter. As we will see later on, it is this picture that will serve as the basis for Barnard's CSM test, which is an exact unconditional test. Let us consider two urns, one marked 1 and one marked 2. Each urn contains again a "very large" number of balls (such that $n_{..}$ drawings without replacement does not alter the experiment conditions), either marked "Success" or "No Success". We draw at random $n_{1\cdot}$ balls from urn 1 and $n_{2\cdot}$ balls from urn 2 (fixing the $(n_{1\cdot}, n_{2\cdot})$ margin), and count afterwards for each urn how many balls are labelled "Success". If we denote by θ_1 and θ_2 the respective proportions of "Success" balls in urns 1 and 2, the probability of ending up with Table 1.1 is given by

$$\binom{n_{1\cdot}}{x_{11}}\theta_1^{x_{11}}(1-\theta_1)^{n_{1\cdot}-x_{11}}\binom{n_{2\cdot}}{x_{21}}\theta_2^{x_{21}}(1-\theta_2)^{n_{2\cdot}-x_{21}}. \tag{2.18}$$

Notice that in the derivation of Fisher's exact test, we also used this abstract picture at first, after which we conditioned on the second margin to finally end up with (2.16).

For the sake of completeness, regarding the nomenclature, it is worth noting that the name "comparative trial" is not deemed correct by everyone. Yates [3] states that in a – medical – comparative trial, individuals need not to be chosen at random from a larger population. They can instead be chosen specifically because they are deemed as suitable test subjects. Yates argues that in the experiment that Barnard calls a "comparative trial", the two groups under consideration are in fact

samples from two larger populations, and prefers to call that experiment "samples from two binomials". In this text, we will however stick with the terminology adopted by Barnard.

From Equations (2.16), (2.17) and (2.18), we see that depending on how we read the contingency table, we end up with different probabilities of observing a given table outcome. This should of course not come as a surprise, as the three mentioned equations represent a completely different experiment. One might perhaps expect that depending the type of experiment performed, one should choose the according "urn-and-balls model" and base all further inference based on this. However, this touches upon the very essence of the debate we will describe in Chapter 3. We will therefore let this issue rest for a moment, and first introduce the test Barnard has built based on this last abstract picture; where the $(n_{1\cdot}, n_{2\cdot})$ table margin is fixed.

2.4 The Exact Unconditional Approach: Barnard's CSM Test

The CSM test, introduced by Barnard [13], considers the "comparative trial" we described in Section 2.3.3. In that setting, we made once again a binomial assumption, such that θ_1 and θ_2 can be interpreted as the binomial success probabilities for each of the groups. The null and alternative hypotheses of the CSM test are just as in Equation (2.1); we want to test whether or not the two groups have the same success probability $\theta_1 = \theta_2 = \theta$. Barnard's idea is to order all possible table outcomes, from the most incompatible with the null hypothesis, in some sense, to the most compatible with the null hypothesis. A table outcome can be incompatible with the null hypothesis, in the sense that if we assume the null hypothesis to be true, it is highly improbable for that particular table outcome to occur. By creating an ordering of the possible table outcomes, we can construct a p-value test, as we described in 2.1. Indeed, to put it in the classical setting of hypothesis testing, we did nothing else than constructing a test statistic $T(\mathbf{X})$ for $\mathbf{X} = (X_{11}, X_{21})$ the number of successes in groups 1 and 2, and then compute the p-value as the probability under the null hypothesis of obtaining a more extreme value of that test statistic. For a table outcome $\mathbf{X} = \mathbf{x}$, the test statistic takes the value $T(\mathbf{x}) \in \mathbb{N}$, which is the rank of $\mathbf{x}$ in the ordering. A more extreme value of $T(\mathbf{X})$ is a smaller value of $T(\mathbf{X})$, since more extreme table outcomes come earlier in the ordering. Thus, for an observation $\mathbf{x}$, we can look at its (two-sided) tail probability $p_\theta(\mathbf{x}) = P_\theta(T(\mathbf{X}) \leq T(\mathbf{x}))$. Since we are working in a discrete setting, computing this probability amounts to enumerating over all possible table outcomes $\mathbf{y}$ with $T(\mathbf{y}) \leq T(\mathbf{x})$, i.e.

$$p_\theta(\mathbf{x}) = \sum_{T(\mathbf{y}) \leq T(\mathbf{x})} P_\theta(\mathbf{X} = \mathbf{y}). \tag{2.19}$$

Under the null hypothesis, we can easily compute how probable it is for a certain table outcome $\mathbf{y} = (y_{11}, y_{21})$ to be observed; this is nothing else than Equation (2.18) with $\theta_1 = \theta_2 = \theta$, i.e.

$$P(\mathbf{y}; \theta) := P_\theta(\mathbf{X} = \mathbf{y}) = \binom{n_{1\cdot}}{x_{11}} \binom{n_{2\cdot}}{x_{21}} \theta^{n_{\cdot 1}} (1-\theta)^{n_{\cdot 2}}. \tag{2.20}$$

Normally, the corresponding p-value test would simply amounts to rejecting the null hypothesis whenever $p_\theta(\mathbf{x})$ is smaller than the given significance level α. However, we cannot refer to $p_\theta(\mathbf{x})$ as a p-value, since it is still a function of the unknown parameter θ.

2.4.1 Constructing the ordering

We want to concretely define how to order the table outcomes. What do we mean by saying that $\mathbf{y}$ is a more extreme observation than $\mathbf{x}$, i.e. $T(\mathbf{y}) \leq T(\mathbf{x})$? We would like to say that a table outcome is extreme if its occurrence probability (under H_0) $P(\cdot;\theta)$ is small. Saying that $\mathbf{y}$ is more extreme than $\mathbf{x}$ then means that $P(\mathbf{y};\theta) \leq P(\mathbf{x};\theta)$. However, $P(\cdot;\theta)$ is a function of $\theta \in [0,1]$, which is unknown to us. Consequently, $p_\theta(\mathbf{x})$ in (2.19) is actually also a function of θ. If we would have set ourselves a significance level α, it can now occur that we might reject the null hypothesis $H_0\colon \theta_1 = \theta_2 = \theta$ only for values of θ within a certain interval such that $p_\theta(\mathbf{x}) \leq \alpha$, while our test would remain inconclusive for θ outside that interval. Barnard mentions that maybe some day, researchers will be happy with a result such as (2.19). He also mentions however that this is not yet the case, and that we should instead strive to somehow translate $p_\theta(\mathbf{x})$, and thus the $P(\cdot;\theta)$-functions into something independent of θ, such that we can get a straightforward "reject/do not reject"-statement at a given significance level [13].

Thus, how can one rank functions, in the presence of the so-called *nuisance parameter* θ? Barnard proposes to make use of certain functionals of $P(\cdot;\theta)$, which would remove the dependence on θ. These functionals should be such that if $P(\mathbf{x};\theta) < P(\mathbf{y};\theta)$ for all $\theta \in [0,1]$ and some table outcomes $\mathbf{x}, \mathbf{y}$, the functional applied to $P(\mathbf{x};\theta)$ should give a smaller number than the functional applied to $P(\mathbf{y};\theta)$. In his paper, Barnard eventually suggests to maximise the sum of the most extreme $P(\cdot;\theta)$-functions as given in (2.19). Instead of (2.7), which assumes we do not have to deal with a nuisance parameter, our new p-value would then be

$$p(\mathbf{x}) = \sup_{\theta\in[0,1]} p_\theta(\mathbf{x}) = \sup_{\theta\in[0,1]} \sum_{T(\mathbf{y})\leq T(\mathbf{x})} P(\mathbf{y};\theta) = \sup_{\theta\in[0,1]} P_\theta(T(\mathbf{X}) \leq T(\mathbf{x})). \tag{2.21}$$

Notice that this is p-value has the same form as the one in defined in (2.15); $p(\mathbf{X})$ is a valid supremum p-value by Lemma 2.3. We have now answered our original question of how to define an ordering based on functions of the unknown nuisance parameter with another question. How do we define an ordering based on a maximum over these functions of the unknown nuisance parameter? Barnard introduced a number of additional conditions in order to decide how to sequentially build up the ordering. Before we discuss these here however, we want to show that we can create an ordering without these conditions, purely using what we have seen so far and one guiding principle. This will essentially tell us that Barnard's conditions, although quite intuitive, are just a matter of choice, and could easily be replaced by some other set of conditions, or could just be removed altogether. The only guiding principle we will need for now is that we want $p_\theta(\mathbf{x})$ to be a function of θ that is as constant as possible. Indeed, ideally we would have no dependence on θ and then $p_\theta(\mathbf{x})$ would be just a flat line as a function of θ, such that we can reject/not reject the null hypothesis for all values of $\theta \in [0,1]$, and

not just for an interval. Since $p(\mathbf{x}) = \sup_{\theta\in[0,1]} p_\theta(\mathbf{x})$, if $p_\theta(\mathbf{x})$ is as flat as possible, we will have intuitively that $p(\mathbf{x}) \approx p_\theta(\mathbf{x})$ for all $\theta \in [0, 1]$.

Based on this guiding principle, we can now construct our ordering inductively. Suppose we have already ordered k table outcomes $\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(k)}$. For the remaining outcomes, we can then compute what their potential p-values would be if they would be the next outcome in the ordering. As the actual next outcome in the ordering, we then choose the one which yields the smallest potential p-value, i.e.

$$\mathbf{x}_{(k+1)} = \underset{\mathbf{y}\in\Omega\setminus\{\mathbf{x}_{(1)},\ldots,\mathbf{x}_{(k)}\}}{\arg\min} \sup_{\theta\in[0,1]} \left\{ P(\mathbf{y};\theta) + \sum_{i=1}^{k} P\big(\mathbf{x}_{(i)};\theta\big) \right\}, \tag{2.22}$$

where $\Omega := \{\mathbf{x} = (x_{11}, x_{21}) \in \mathbb{Z}^2 : 0 \le x_{11} \le n_{1\cdot}, 0 \le x_{21} \le n_{2\cdot}\}$ is the set of all possible table outcomes. In case of multiple table outcomes having the same smallest potential p-value, we must come up with some tie-breaker rule, for which we refer the reader to Sections 2.5.1 and 2.5.2. What we have essentially done is computed for each outcome how $p_\theta(\mathbf{y})$ would look like as a function of θ if that outcome would be the next in the ordering. We then chose the next outcome in the ordering as the outcome for which $p_\theta(\mathbf{y})$ is the flattest function of θ. Indeed, saying that the outcome $\mathbf{y}$ which had the smallest possible maximum value for $p_\theta(\mathbf{y}) = P(\mathbf{y};\theta) + \sum_{i=1}^{k} P\big(\mathbf{x}_{(i)};\theta\big)$, is the same as saying that $p_\theta(\mathbf{y})$ is the flattest possible function of θ. Although not a proof, this idea is illustrated in Figure 2.1.

In the leftmost plot, one can see $p_\theta(\mathbf{x}_{(1)}) = P(\mathbf{x}_{(1)};\theta)$ up to and including $p_\theta(\mathbf{x}_{(7)})$ as functions of θ. Only the top one, $p_\theta(\mathbf{x}_{(7)})$, is indicated. The maximum of this curve, $p(\mathbf{x}_{(7)})$, is indicated by the dotted line. So far, this is our ordering. We are now considering two possible candidates to be next in the ordering, $\mathbf{y}_1$ and $\mathbf{y}_2$. In order to decide which one is next, we compute the what the functions $p_\theta(\mathbf{y}_1) = p_\theta(\mathbf{x}_{(7)}) + P(\mathbf{y}_1;\theta)$ and $p_\theta(\mathbf{y}_2) = p_\theta(\mathbf{x}_{(7)}) + P(\mathbf{y}_2;\theta)$ would look like if we would choose one of them as the next outcome in the ordering. The maximum of $p_\theta(\mathbf{y}_1)$ is around 0.6, and the maximum of $p_\theta(\mathbf{y}_2)$ is 1. Therefore, we set $\mathbf{y}_1 = \mathbf{x}_{(8)}$. Clearly, $p_\theta(\mathbf{y}_1)$ is a much flatter function of θ compared to $p_\theta(\mathbf{y}_2)$. We can now repeat this procedure in order to find $\mathbf{x}_{(9)}$ and so on. Note that at the end, we will have gone through all possible outcomes, so our stack of curves will have filled up the entire plot.

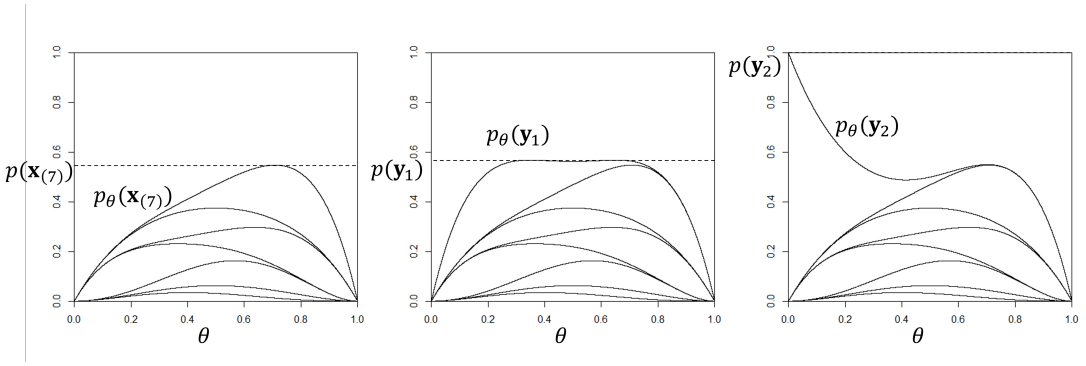


Figure 2.1: Plots of how $p_\theta(\cdot)$ would look like if $\mathbf{y}_1$ (middle) or $\mathbf{y}_2$ (right) was to be chosen as the next table in the ordering, where until now the first seven outcomes (left) have been determined.

Note that this smallest potential p-value also becomes the actual p-value for

$x_{(k+1)}$, i.e. $p(x_{(k+1)}) = p(y_1)$. This is indicated by Figure 2.1, which visualises $p_\theta(\cdot)$ as the sum of the $P(\cdot, \theta)$-functions, which are represented by the different curves stacked on top of each other. For example, in the leftmost plot, the highest curve represents $\theta \mapsto p_\theta(\mathbf{x}_{(7)})$; it is the sum of $P(\mathbf{x}_{(1)}; \theta)$ up to $P(\mathbf{x}_{(7)}; \theta)$. The six curves below indicate – from lowest to highest – the functions $p_\theta(\mathbf{x}_{(1)}) = P(\mathbf{x}_{(1)}; \theta)$ up to and including $p_\theta(\mathbf{x}_{(6)})$.

Using (2.22), we are now able to order all possible table outcomes. Note that setting $k = 0$ yields $\mathbf{x}_{(1)}$, the outcome $\mathbf{y} \in \Omega$ which has the smallest value for $\sup_{\theta \in [0,1]} P(\mathbf{y}; \theta)$. In terms of the test statistic $T(\mathbf{X})$, we set $T(\mathbf{x}_{(k)}) = k$, for each possible value of $k \in \mathbb{Z}$ (which is larger than 0 and at most $(n_{1\cdot} + 1)(n_{2\cdot} + 1)$, but possibly less if different table outcomes have the same maximum).

2.4.2 The C and S conditions

Realise that in order to build up the entire ordering, we need to repeat the above procedure until we have considered all table outcomes in Ω. Thus, in order to find $\mathbf{x}_{(k+1)}$, we need to compute and maximise $P(\cdot; \theta) + p_\theta(\mathbf{x}_{(k)})$ a total of at most $|\Omega| - k$ times. If we could somehow limit the amount of possible candidates for which we would need to do this procedure at each step, we would speed it up a lot. Barnard, who only had the computational power of the 1940s at his disposal, has done just that [13]. He introduced two conditions; the convexity (C) condition and the symmetry (S) condition, that would significantly reduce the set of candidates in which we should look to find the next outcome in the ordering. Together with the procedure we just described, which Barnard called the maximisation (M) condition, this resulted in Barnard's test; the CSM test.

In order to best explain the C and S conditions, we will represent the table outcome space Ω as a lattice. Indeed, given $(n_{1\cdot}, n_{2\cdot})$, Table 1.1 is uniquely determined if we know the pair (x_{11}, x_{21}). We can therefore summarise all possible table outcomes in a grid containing $(n_{1\cdot} + 1) \times (n_{2\cdot} + 1)$ points, just like Figure 2.2.

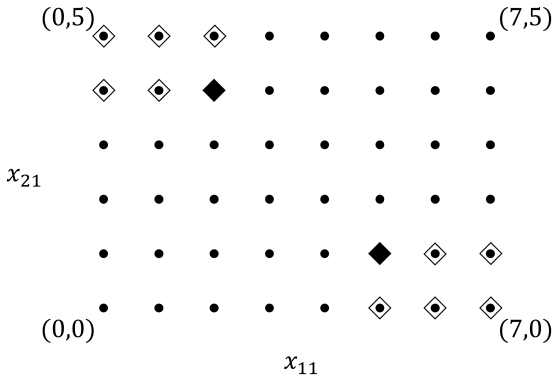


Figure 2.2: Sample space of a 2×2 contingency table with $n_{1\cdot} = 7$ and $n_{2\cdot} = 5$.

Let us start with the symmetry (S) condition. Let $n_{1\cdot} = 7$ and $n_{2\cdot} = 5$, and consider for a moment the outcome $\mathbf{x} = (2, 4)$, which is the upper-left point marked with a black diamond in the grid. This point corresponds to a contingency table with $x_{11} = 2$ successes in group 1 and $x_{21} = 4$ successes in group 2. Barnard argues that observing this table should provide the same evidence against the null hypothesis as a table with $n_{1\cdot} - x_{11} = 5$ successes in group 1 and $n_{2\cdot} - x_{21} = 1$

successes in group 2, i.e. the table where we interchanged successes and failures. Indeed, one could argue that instead of testing $\theta_1 = \theta_2$, we are testing $1 - \theta_1 = 1 - \theta_2$. Therefore, the two points indicated by a black diamond in Figure 2.2 should receive the same rank in the ordering. Remark that we can also have a different type of symmetry in the case that $n_{1\cdot} = n_{2\cdot}$. For example, if $n_{1\cdot} = n_{2\cdot} = 5$ the outcome $(2, 1)$ would provide the same evidence against the null hypothesis as $(3, 4)$ due to the S condition, but also as $(1, 2)$ (and by the S condition also as $(4, 3)$). This is because we could interchange the labels of group 1 and group 2, which would yield the same conclusion in the case that the alternative hypothesis is $H_1 \colon \theta_1 \neq \theta_2$.

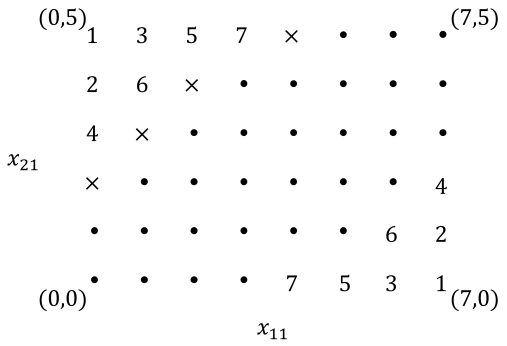
The other condition Barnard imposed was the convexity (C) condition, which states that if the table (x_{11}, x_{21}) lies above the diagonal connecting the outcomes $(0, 0)$ and $(7, 5)$, all tables (y_{11}, y_{21}) with $y_{11} \leq x_{11}$ and $y_{21} \geq x_{21}$ (at least one strict inequality) should receive a lower rank than (x_{11}, x_{21}). If (x_{11}, x_{21}) lies below the diagonal connecting the outcomes $(0, 0)$ and $(7, 5)$, all tables (y_{11}, y_{21}) with $y_{11} \geq x_{11}$ and $y_{21} \leq x_{21}$ (at least one strict inequality) should receive a lower rank than (x_{11}, x_{21}). The idea here is that, say, for the point $(2, 4)$, the points $(1, 4)$ and $(2, 5)$ should be stronger evidence against the null hypothesis than $(2, 4)$ itself. This is because these two points indicate a wider difference in the amount of successes of both groups. If in one group we observe even less successes than we already had, and in the other one more, we are intuitively more inclined to reject the null hypothesis that $\theta_1 = \theta_2$. By a transitivity argument, it should be clear that all points in the upper-left quadrant of $(2, 4)$ should be considered as more incompatible with the null hypothesis than $(2, 4)$. Due to the symmetry condition, the same can be said to all points in the lower-right quadrant of $(5, 1)$. All these points are marked with empty diamonds in Figure 2.2.

Barnard's CSM ordering can now be found in exactly the same way as the first ordering method we presented, except that now, by symmetry, we only need to look at the upper-left triangle of points in Figure 2.2, as the outcomes in the lower triangle will get the same rank as their respective symmetric counterparts. Then, by convexity, rank 1 should be given to the most upper-left point. From there, to determine which outcome should be next, we only need to consider the unordered points such that their nearest upper or left neighbour is already ordered. In this fashion, we slowly fill up the table with p-values starting from the $(0, 5)$ and $(7, 0)$ towards the diagonal between $(0, 0)$ and $(7, 5)$. We no longer need to compute $P(\mathbf{y}; \theta) + p_\theta(x_{(k+1)})$ at each step for each table outcome $\mathbf{y}$ which has not yet been given a rank, but only for the outcomes which we have to consider by the C and S conditions.

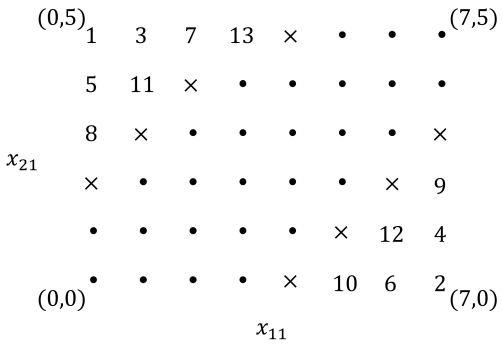
Note that we can also choose to omit either the C or the S condition to construct an ordering. In the case of omitting the C condition, this entails that each time we want to assign the next rank, we should look through all points in the upper triangle of the outcome space which have not been ranked yet. If we were to omit the S condition, we would have to look at both the upper and lower triangle of the outcome space; starting from the corners and working our way towards the main diagonal connecting $(0, 0)$ and $(n_{1\cdot}, n_{2\cdot})$. In the upper triangle, we only look at the points that have already ordered upper or left neighbours, while in the lower triangle we only consider the points of which the lower or right neighbours are already ranked. In Figure 2.3, the way in which we go through the outcome space is

visualised for the four ranking methods we considered so far; applying both C and S (2.3a), only C (2.3b), only S (2.3c) and applying neither C nor S (2.3d). In each lattice, we indicate all points we have already ordered by their respective ranks, and the points that we would consider next for the ordering by $\times$. The next point in line will be chosen according to (2.22). Of course, if the S condition is applied, we replace the term $P(x_{11}, x_{21}; \theta)$ by the term $P(x_{11}, x_{21}; \theta)+P(n_{1\cdot}-x_{11}, n_{2\cdot}-x_{21}; \theta)$ in (2.22), as we group a point and its symmetric counterpart together in the ordering.

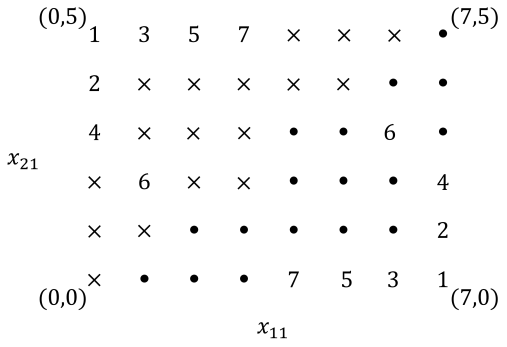
Thus, to recap, in Figure 2.3a, we apply both the C and the S conditions, we fill up the grid from both the $(0, 5)$ and $(7, 0)$ corners, giving each symmetric pair the same rank. Hence, both the "upper" and "lower" part will always contain the same number of grid points. In Figure 2.3b, we remove of the symmetry condition, so we no longer need to group symmetric points together. We thus have to look at both the upper-left points and lower-right points as completely separate. This is also the case in Figure 2.3d, except that now we also remove the convexity condition and so do not longer restrict ourselves to points which are adjacent to the ordered points. Finally, in Figure 2.3c, we only apply the symmetry condition and so can consider each point in the "upper triangle". The same ranks will be given to the respective symmetric counterparts in the "lower triangle". In practice, "jumps" as illustrated in Figure 2.3c and Figure 2.3d (rank 6 and 11 respectively) will be very rare. Even without enforcing the C condition, we will often end up with an ordering that satisfies the C condition (or at least does not have any of the aforementioned jumps) anyway.



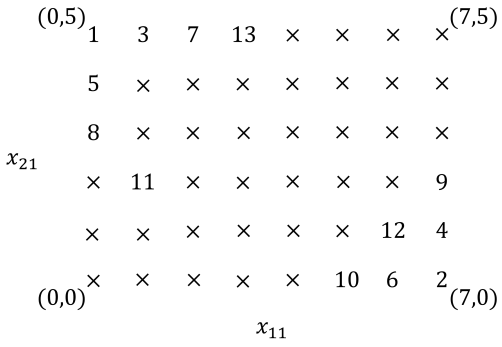
(a) Apply both C and S.



(b) Apply only C.



(c) Apply only S.



(d) Apply neither C nor S.

Figure 2.3: Each lattice shows which points we would consider next in our ordering according to the applied conditions. These points are indicated by $\times$.

2.5 Alternatives to the CSM Ordering

As we have just seen, one can create various unconditional tests by leaving out the C and/or S condition(s). Essentially, these conditions were only invoked in order to speed up the process of ordering the set of outcomes. The S condition furthermore made sure that tables which in our eyes present the same evidence against H_0 receive the same p-value. However, nothing prohibits us from coming up with different conditions, or even totally different orderings, as long as they are – in some sense – reasonable. Barnard's CSM test is in fact just a specific case of a p-value test, where the p-value is computed using the so-called supremum method. In such a test, we compute the p-value as

$$p(\mathbf{x}) = \sup_{\theta \in [0,1]} p_\theta(\mathbf{x}) = \sup_{\theta \in [0,1]} P_\theta(T(\mathbf{X}) \leq T(\mathbf{x})), \tag{2.23}$$

where in the case of Barnard's CSM test, the value of the test statistic $T(\mathbf{x})$ was the rank of the table outcome $\mathbf{x}$, according to the ordering we constructed for Barnard's CSM test. However, we might use any other test statistic too. Note that if we reject the null hypothesis for large values of the test statistic instead, we should turn around the inequality in (2.23). We have shown in Lemma 2.3 that this approach always produces a valid p-value, no matter the chosen test statistic.

2.5.1 The χ^2 test statistic

As an example of the supremum method, we could use the chi-square test statistic $\chi^2(\mathbf{X})$ defined in (2.2) instead of $T(\mathbf{X})$ in (2.23). For each possible table outcome, we compute the corresponding test statistic value. Afterwards, as we would reject the null hypothesis $H_0\colon \theta_1 = \theta_2 = \theta$ for large values of $\chi^2(\mathbf{x})$, we could define a p-value as in (2.23), with the inequality reversed. This approach has first been proposed by Suissa and Shuster [15], although they referred to χ^2 as Z_{p}^2. They also mentioned that one can use Z_{u}^2, but showed that if $n_{1\cdot} = n_{2\cdot}$, using Z_{u}^2 gives the same results as using Z_{p}^2.

The chi-square test statistic can also be used in a different way. In the context of the CSM test, instead of using the S condition defined earlier, we might also define another symmetry condition (S_χ) where table outcomes which have the same χ^2-value are grouped together. Note that if two tables are symmetric according to the S condition, they are also symmetric according to S_χ. Indeed, suppose we have the symmetric pair $\mathbf{x} = (x_{11}, x_{21})$, $\mathbf{y} = (y_{11}, y_{21}) = (n_{1\cdot} - x_{11}, n_{2\cdot} - x_{21})$, where $\mathbf{y}$ is just Table 1.1 with the columns flipped. That is, $y_{11} = x_{12}$, $y_{21} = x_{22}$, $y_{12} = x_{11}$, and $y_{22} = x_{21}$. But then (2.3) immediately shows that

$$\chi^2(\mathbf{x}) = \frac{(x_{11}x_{22} - x_{12}x_{21})^2 n_{\cdot\cdot}}{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}} = \frac{(y_{12}y_{21} - y_{11}y_{22})^2 n_{\cdot\cdot}}{n_{1\cdot}n_{2\cdot}n_{\cdot 2}n_{\cdot 1}} = \chi^2(\mathbf{y}).$$

Similarly, if $n_{1\cdot} = n_{2\cdot}$, the outcomes $\mathbf{z} = (z_{11}, z_{21}) = (n_{\cdot 1} - x_{11}, n_{\cdot 1} - x_{21})$ (Table 1.1 with the rows flipped) and $\mathbf{w} = (n_{1\cdot} - z_{11}, n_{2\cdot} - z_{21})$ (Table 1.1 with the rows and columns flipped) are also deemed as symmetric counterparts of $\mathbf{x}$. Since $z_{11} = x_{21}$, $z_{21} = x_{11}$, $z_{12} = x_{22}$, and $z_{22} = x_{12}$, we then have

$$\chi^2(\mathbf{x}) = \frac{(x_{11}x_{22} - x_{12}x_{21})^2 n_{\cdot\cdot}}{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}} = \frac{(z_{21}z_{12} - z_{22}z_{11})^2 n_{\cdot\cdot}}{n_{2\cdot}n_{1\cdot}n_{\cdot 2}n_{\cdot 1}} = \chi^2(\mathbf{z}).$$

From the equality of $\chi^2(\mathbf{y})$ and $\chi^2(\mathbf{x})$, it then trivially follows that $\chi^2(\mathbf{w}) = \chi^2(\mathbf{x})$ too. The reverse implication is however not necessarily true; as long as the factor $(x_{11}x_{22} - x_{12}x_{21})^2$ and the marginal totals remain constant, we might construct tables which are symmetric to $\mathbf{x}$ according to S_χ, but not according to S. A trivial example of this can be found in the space of tables with $(n_{1\cdot}, n_{2\cdot}) = (10, 10)$ (or any outcome space with $n_{1\cdot} = n_{2\cdot}$). Here, all tables with $x_{11} = x_{21}$ necessarily also have $x_{12} = x_{22}$ and thus $x_{11}x_{22} - x_{12}x_{21}$ is automatically zero. Consequently, tables like $(9, 9)$ and $(5, 5)$, which are definitely not symmetric according to Barnard's S, are symmetric according to S_χ. A less trivial example occurs at, amongst others, $(n_{1\cdot}, n_{2\cdot}) = (3, 6)$. There the tables $(1, 5)$, $(2, 6)$, and $(3, 3)$ all have a chi-square value of $9/4$, but are clearly not symmetric according to S.

Splitting up the set of outcomes into (equivalence) classes of symmetric tables according to S thus leads to more classes than when splitting according to S_χ. This has computational implications. The fewer groups of tables one needs to go through, the faster the maximisation procedure will be. However, also note that if we have fewer, larger groups of tables, we also have less possible ways in which to construct a sum of $P(\cdot;\theta)$-functions that is as flat as possible. To make sense of this intuitively, consider again Figure 2.1. If the symmetry groups are larger, we will have a small number of "thick layers" which we are able to use to make a stack of functions that is as flat as possible under the given significance level α. However, with smaller symmetry groups, we will have a larger number of "thin layers", which allows us to "fine-tune" the stack of layers a bit more. This leads us to expect that in general, symmetry conditions that result in fewer symmetry groups, will yield in a quicker, but less powerful test, as it will be more difficult to construct a stack of $P(\cdot;\theta)$-functions as close to α as possible, which potentially misses out on some power.

A final application of the chi-square test statistic is as a possible tie-breaker. It could happen that while building up the ordering, we encounter at a certain point that two possible candidates minimise the supremum in (2.22). In that case, we could compute the value of the χ^2 test statistic for both tables and pick as the next table in the ordering the one with the largest value. Such a tie happens for example when ordering the outcome space of 2×2 tables with group sizes $(n_{1\cdot}, n_{2\cdot}) = (9, 7)$. After ordering 34 out of the 40 possible symmetry groups of tables, there are six candidate groups to be the next one in the ordering and receive rank 35. Two of these, the tables $(x_{11}, x_{21}) = (4, 4)$ (with its symmetric counterpart $(5, 3)$) and $(x_{11}, x_{21}) = (5, 4)$ (with its symmetric counterpart $(4, 3)$) both reach the same maximum

$$\sup_{\theta\in[0,1]} \left\{ P((4,4);\theta) + P((5,3);\theta) + \sum_{i=1}^{34} P\big(\mathbf{x}_{(i)};\theta\big) \right\} = 0.7865601,$$

$$\sup_{\theta\in[0,1]} \left\{ P((5,4);\theta) + P((4,3);\theta) + \sum_{i=1}^{34} P\big(\mathbf{x}_{(i)};\theta\big) \right\} = 0.7865601,$$

which is also the lowest maximum out of the 6 maxima for all possible candidate groups. We thus have a tie: which of these two groups should receive rank 35? If we use the chi-square test statistic as a tie-breaker, we first compute the sum of

test statistic values for both groups,

$$\chi^2(4,4) + \chi^2(5,3) = 0.508,$$
$$\chi^2(5,4) + \chi^2(4,3) = 0.008.$$

Since the group consisting of the tables $(4,4)$ and $(5,3)$ has the largest sum of test statistic values, it is considered as "more extreme" in a chi-square sense. Therefore, these two tables will receive rank 35.

2.5.2 Using the mean value of $P(\cdot;\theta)$

Another p-value test is inspired on one of the functionals Barnard introduced in his CSM paper [13]. Although he only considered $\sup_{\theta\in[0,1]} \sum P(\cdot;\theta)$ to get rid of θ in the rest of his paper, he also mentioned the mean value $\int_0^1 P(\cdot;\theta)\mathrm{d}\theta$. The idea here is to use to mean value as a measure for how (un)likely a table outcome is under the null hypothesis. The smaller the mean value, the more extreme the test outcome. Important to note is that the mean values of the $P(\cdot;\theta)$-functions are easily computed, as we recognise a beta function with integer arguments:

$$\begin{aligned}\int_0^1 P(\mathbf{x};\theta)\mathrm{d}\theta &= \int_0^1 \binom{n_{1\cdot}}{x_{11}}\binom{n_{2\cdot}}{x_{21}}\theta^{n_{\cdot 1}}(1-\theta)^{n_{\cdot 2}}\mathrm{d}\theta \\ &= \binom{n_{1\cdot}}{x_{11}}\binom{n_{2\cdot}}{x_{21}} B(n_{\cdot 1}+1, n_{\cdot 2}+1) \\ &= \binom{n_{1\cdot}}{x_{11}}\binom{n_{2\cdot}}{x_{21}} \frac{n_{\cdot 1}!n_{\cdot 2}!}{(n_{\cdot\cdot}+1)!}. \end{aligned} \tag{2.24}$$

Apart from basing an ordering solely on this mean value, we can instead, just as with the chi-square test statistic, use it to determine equivalence classes of "symmetric" tables. Tables with the same value of (2.24) will be grouped together. We will refer to this symmetry condition as the $\mathrm{S_V}$ condition, indicating we are using the area (or **V**olume for larger tables as we will see in Chapter 4) under the $P(\mathbf{x};\theta)$-function. Tables which are symmetric according to Barnard's S condition are symmetric according to $\mathrm{S_V}$. A small computation can convince the reader that swapping columns (and rows if $n_{1\cdot} = n_{2\cdot}$) does not change the value (2.24). The reverse implication is in general not true, take for example $(n_{1\cdot}, n_{2\cdot}) = (4,3)$. Both $(3,3)$ and $(3,2)$ have a $P(\cdot;\theta)$-function with mean value $1/14$. Furthermore, this same counterexample can be used to realise that $\mathrm{S_V}$ does not imply S_χ. The counterexamples from 2.5.1 also serve as evidence that S_χ does not imply $\mathrm{S_V}$ either. However, in the case that $n_{1\cdot} = n_{2\cdot}$, we do suspect that $\mathrm{S_V}$ implies S. Although not a proof, we have confirmed that $\mathrm{S_V}$ and S yield the same symmetry groups in all outcome spaces up to $n_{1\cdot} = n_{2\cdot} = 150$. We will come back to this in Section 4.2.4

Alternatively, we can use this quantity as a tie-breaker as well and choose the table with the smallest mean value as the next one in the ordering. If we would apply this method to the tie discussed in Section 2.5.1, we would find that

$$\int_0^1 P((4,4);\theta)\mathrm{d}\theta + \int_0^1 P((5,3);\theta)\mathrm{d}\theta = 0.040,$$
$$\int_0^1 P((5,4);\theta)\mathrm{d}\theta + \int_0^1 P((4,3);\theta)\mathrm{d}\theta = 0.045.$$

Therefore, we would again assign rank 35 to the tables $(4,4)$ and $(5,3)$.

2.5.3 Boschloo's test

Let us finish this section with another well-known special case of the supremum method; Boschloo's test, which makes use of the p-value of another test as a test statistic. Originally, this test was designed as an unconditional improvement of Fisher's exact test, and is by construction uniformly more powerful than Fisher's test [16]. Similar tests have been proposed by MacDonald, Davis and Milliken [17] and Crans and Shuster [18]. Recall that for Fisher's exact test, we consider as our set of possible outcomes, all tables which had the same margin totals. Conditional on $X_{11} + X_{21} = n_{\cdot 1}$, we found that X_{11} had a hypergeometric distribution. Based on this, we could then construct the largest possible critical region $K^{\alpha}_{n_{\cdot 1}}$ such that $\alpha_{n_{\cdot 1}} := P\big((X_{11}, X_{21}) \in K^{\alpha}_{n_{\cdot 1}} \mid X_{11} + X_{21} = n_{\cdot 1}\big) \leq \alpha$. Consider the example in Figure 2.4. There we again visualise the outcome space Ω of (X_{11}, X_{21}) with $n_{1\cdot} = 7$ and $n_{2\cdot} = 5$. Suppose we observe the outcome $(2, 1)$ (i.e. $n_{\cdot 1} = 3$). Then Fisher's exact test, for some significance level α, will only consider all outcomes lying on the dotted line $n_{\cdot 1} = 3$, and we will end up with the critical region $K^{\alpha}_3 = \{(0, 3)\}$, as indicated by the black diamond on the $n_{\cdot 1} = 3$ line.

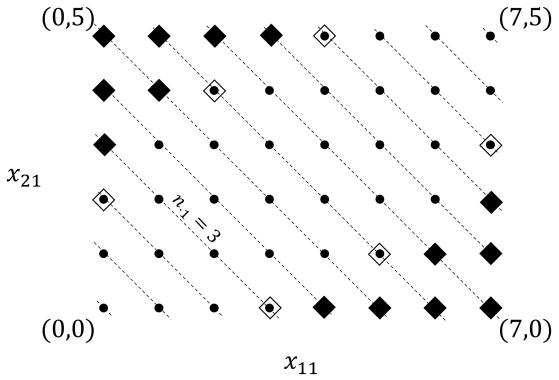


Figure 2.4: The outcome space Ω for an experiment where X_{11} and X_{21} are binomially distributed with parameters $(7, \theta_1)$ and $(5, \theta_2)$ respectively. The diagonal lines link table outcomes with equal values of $n_{\cdot 1}$.

Boschloo argued that the significance level α is in fact a conditional level, in the sense that the probability of wrongly rejecting H_0, conditional on $X_{11}+X_{21} = n_{\cdot 1}$ is at most α. One can also associate an unconditional size α' with α; the probability of wrongly rejecting H_0 when any arbitrary table outcome is observed (with possibly another value for $n_{\cdot 1}$, i.e. lying on another dotted line in Figure 2.4). By the law of total probability, this unconditional size is easily seen to be a function of the nuisance parameter θ:

$$\alpha'(\theta) = P_\theta((X_{11}, X_{21}) \in K^{\alpha}) = \sum_{n_{\cdot 1}=0}^{n_{\cdot\cdot}} \alpha_{n_{\cdot 1}} P_\theta(X_{11} + X_{21} = n_{\cdot 1}), \tag{2.25}$$

where $K^{\alpha} := \bigcup_{n_{\cdot 1}=0}^{n_{\cdot\cdot}} K^{\alpha}_{n_{\cdot 1}}$ is the union of all Fisher critical regions, where we run through all possible values of $X_{11} + X_{21}$. In Figure 2.4, this corresponds to the set of all black diamonds. Boschloo observes from a numerical experiment that this unconditional size is often much smaller than α, even at its maximum $\alpha'_{\max} := \max_{\theta \in [0,1]} \alpha'(\theta)$. As we will see later, this is in line with the often stated critique that Fisher's exact test is very conservative. Boschloo's idea to increase

the unconditional size, such that it lies closer to the level α (but still smaller than α), is to raise the conditional level. That is, by performing Fisher's exact test at a higher level, say $\gamma > \alpha$, each critical region $K^{\alpha}_{n_{\cdot 1}}$ would increase to $K^{\gamma}_{n_{\cdot 1}}$. In our example, this raising led to the addition of $(3, 0)$ to the critical region K^{α}_3, giving $K^{\gamma}_3 = \{(0, 3), (3, 0)\}$. This is indicated in Figure 2.4 by the empty diamond at $(3, 0)$. Doing this for all values of $n_{\cdot 1}$ gives the larger unconditional critical region K^{γ}, and thus also a higher unconditional size. Note that it need not to be the case that each conditional critical region $K^{\gamma}_{n_{\cdot 1}}$ is strictly larger than $K^{\alpha}_{n_{\cdot 1}}$ (recall that we are dealing with discrete distributions!). By numerical computation, Boschloo could find for a given table the largest value γ' such that $\sup_{\theta \in [0,1]} P_\theta\big((X_{11}, X_{21}) \in K^{\gamma'}\big) \leq \alpha$. Alternatively, we can write the restriction for γ' in terms of the Fisher p-value (2.13). The raised conditional level γ' is the largest value such that

$$\sup_{\theta \in [0,1]} P_\theta(p_F(X_{11}, X_{21}) \leq \gamma') \leq \alpha. \tag{2.26}$$

In his paper, Boschloo computed the γ'-values for tables of sample sizes up to 50 [16]. For general tables, finding γ' would require iteratively raising the conditional level until a suitable value has been found. However, as mentioned without proof in [12], it turns out that this method is equivalent to Barnard's test where the ordering is determined by the p-values of Fisher's exact test. The big advantage of this is that we can write Boschloo's test as a p-value test, instead of as a procedure which only returns a critical region. Indeed, in its current form, Boschloo's test does not really have a clear-cut p-value, other than maybe the corresponding Fisher p-value, which we should compare with the raised level γ' in order to decide whether or not to reject.

Proposition 2.4. Boschloo's test of rejecting the null hypothesis whenever $p_F(\mathbf{x}) \leq \gamma'$, where γ' is the largest number such that (2.26) holds, is the same test as the p-value test where we reject H_0 whenever $p_B(\mathbf{x}) \leq \alpha$, where

$$p_B(x_{11}, x_{21}) = \sup_{\theta \in [0,1]} P_\theta(p_F(X_{11}, X_{21}) \leq p_F(x_{11}, x_{21})). \tag{2.27}$$

We will call $p_B(\mathbf{x})$ the Boschloo p-value. Before beginning with the proof, let us state a useful property of valid p-values.

Lemma 2.5. Assume that $p(\mathbf{X})$ is a valid p-value. Then $p(\mathbf{0}) = p(\mathbf{n}) = 1$, where $\mathbf{0}$ and $\mathbf{n}$ are the table outcomes $(0, 0)$ and $(n_{1\cdot}, n_{2\cdot})$, respectively.

Proof of Lemma 2.5. Suppose that $p(\mathbf{X})$ is a valid p-value. We will show that $p(\mathbf{0}) = p(\mathbf{n}) = 1$. Recall that a p-value is valid whenever $P_\theta(p(\mathbf{X}) \leq \alpha) \leq \alpha$ for all $\theta \in [0, 1]$ and $\alpha \in [0, 1]$. In particular, for $\theta = 0$ and $\alpha \in [0, 1]$, $P_0(p(\mathbf{X}) > \alpha) \geq 1 - \alpha$. However, since $P_0(\mathbf{X} = \mathbf{x}) = \mathbb{1}_{\{\mathbf{x}=\mathbf{0}\}}$, we must have for all $\alpha \in [0, 1]$ that

$$\begin{aligned} P_0(p(\mathbf{X}) > \alpha) &= \sum_{\mathbf{x} \in \Omega} P_0(p(\mathbf{X}) > \alpha \mid \mathbf{X} = \mathbf{x}) P_0(\mathbf{X} = \mathbf{x}) \\ &= P_0(p(\mathbf{X}) > \alpha \mid \mathbf{X} = \mathbf{0}) \geq 1 - \alpha. \end{aligned} \tag{2.28}$$

Now suppose that $p(\mathbf{0}) < 1$. Then there exists an $\alpha \in [p(\mathbf{0}), 1)$. But then, if $\mathbf{X} = \mathbf{0}$, $p(\mathbf{X}) \leq \alpha$ and so $P_0(p(\mathbf{X}) > \alpha \mid \mathbf{X} = \mathbf{0}) = 0 < 1 - \alpha$, which contradicts (2.28). Thus we need $p(\mathbf{0}) = 1$. Repeating this argument for $\theta = 1$ yields $p(\mathbf{n}) = 1$. □

Proof of Proposition 2.4. Firstly, one key observation should be made. Although one can vary γ continuously within $[0, 1]$, it suffices to consider only the possible Fisher p-values as possible values for γ. Indeed, suppose that we have found γ' such that it is the largest value such that (2.26) holds. Moreover, let us order the Fisher p-values corresponding to each possible table outcome in Ω in a sequence $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{((n_{1\cdot}+1)(n_{2\cdot}+1))} = 1$. This last inequality follows from Lemma 2.5, and is a nice property of valid p-values.

Now, if $\gamma' < p_{(1)}$, the unconditional critical region would be empty. Hence, Boschloo's procedure leads to an empty critical region. By (2.26), we know that $\sup_{\theta\in[0,1]} P_\theta\big(p_F(X_{11}, X_{21}) \leq p_{(1)}\big)$ must be greater than α. For $\mathbf{a}, \mathbf{b} \in \Omega$, we clearly have

$$p_F(\mathbf{a}) \leq p_F(\mathbf{b}) \iff \sup_{\theta\in[0,1]} P_\theta(p_F(\mathbf{X}) \leq p_F(\mathbf{a})) \leq \sup_{\theta\in[0,1]} P_\theta(p_F(\mathbf{X}) \leq p_F(\mathbf{b})). \tag{2.29}$$

Therefore, if $\mathbf{a}$ is such that $p_{(1)} = p_F(\mathbf{a})$, we also have that $p_B(\mathbf{a})$ is the smallest possible Boschloo p-value. But then, all the Boschloo p-values are greater than α, and so the critical region from the supremum method with the Fisher p-values as test statistic is also empty. Hence the two test procedures yield the same conclusion. Remark that the right-to-left implication in (2.29) is only true because $p_F(\mathbf{a})$ and $p_F(\mathbf{b})$ are Fisher p-values themselves, and thus possible values of $p_F(\mathbf{X})$. If we would substitute these two numbers by arbitrary constants c_1 and c_2, the right-to-left implication does not hold as we might have $p_{(j)} \leq c_2 < c_1 < p_{(j+1)}$ for some $j \in \{1, \ldots, (n_{1\cdot} + 1)(n_{2\cdot+1}) - 1\}$, such that on the right-hand side we have

$$\sup_{\theta\in[0,1]} P_\theta(p_F(\mathbf{X}) \leq c_1) = \sup_{\theta\in[0,1]} P_\theta(p_F(\mathbf{X}) \leq c_2),$$

while $c_1 > c_2$.

If γ' were to be 1, that would mean that the whole sample space would be the unconditional critical region. By (2.26), since $p_F(X_{11}, X_{21}) \leq 1$ no matter the value of θ, we would realise that $\gamma' = 1$ only occurs if $\alpha = 1$ too. In this case both Boschloo's procedure and Barnard's test with the Fisher p-value ordering lead to the same critical region, being the whole sample space, as all p-values are always smaller than or equal to 1.

Therefore, let us disregard these two trivial cases and suppose that $\gamma' \in [p_{(1)}, 1)$. We can then find $k \in \{1, \ldots, (n_{1\cdot} + 1)(n_{2\cdot} + 1) - 1\}$ such that $p_{(k)} \leq \gamma' < p_{(k+1)}$. But then, since the outcome space is discrete, we clearly have for all $\theta \in [0, 1]$ that

$$\sup_{\theta\in[0,1]} P_\theta(p_F(X_{11}, X_{21}) \leq \gamma') = \sup_{\theta\in[0,1]} P_\theta\big(p_F(X_{11}, X_{21}) \leq p_{(k)}\big),$$

and that

$$\sup_{\theta\in[0,1]} P_\theta(p_F(X_{11}, X_{21}) \leq \gamma') < \sup_{\theta\in[0,1]} P_\theta\big(p_F(X_{11}, X_{21}) \leq p_{(k+1)}\big),$$

since $\mathbf{x}_{(k+1)}$ has a nonzero probability of being observed for $\theta \in (0, 1)$. By (2.26), we also know that $\sup_{\theta\in[0,1]} P_\theta\big(p_F(X_{11}, X_{21}) \leq p_{(k+1)}\big) > \alpha$. Consequently, we might just as well use $p_{(k)}$ as our raised conditional significance level instead of γ', as it will include exactly the same amount of table outcomes in the critical region.

However, the above means that if we observe an outcome $\mathbf{x}$, we will reject the null hypothesis using Boschloo's procedure if and only if:

$$\begin{aligned}
& p_F(\mathbf{x}) \leq \gamma' = \max\left\{\gamma : \sup_{\theta\in[0,1]} P_\theta(p_F(\mathbf{X}) \leq \gamma) \leq \alpha\right\} \\
\iff & p_F(\mathbf{x}) \leq \max_{\mathbf{y}\in\Omega}\left\{p_F(\mathbf{y}) : \sup_{\theta\in[0,1]} P_\theta(p_F(\mathbf{X}) \leq p_F(\mathbf{y})) \leq \alpha\right\} \\
\iff & \exists \mathbf{y} \in \Omega : p_F(\mathbf{x}) \leq p_F(\mathbf{y}) \text{ and } \sup_{\theta\in[0,1]} P_\theta(p_F(\mathbf{X}) \leq p_F(\mathbf{y})) \leq \alpha \\
\iff & p_B(\mathbf{x}) = \sup_{\theta\in[0,1]} P_\theta(p_F(\mathbf{X}) \leq p_F(\mathbf{x})) \leq \alpha,
\end{aligned}$$

which is equivalent to rejecting the null hypothesis using the supremum method with the Fisher p-value as a test statistic. In the last equivalence, the reverse implication is satisfied if we take $\mathbf{y} = \mathbf{x}$. □

CHAPTER

THREE

HOW TO GET RID OF THE NUISANCE PARAMETER

In the previous section, we have discussed three types of tests for contingency tables. In each of these tests, we have encountered in one way or another the unknown parameter θ; the common value of the probabilities θ_1 and θ_2 under the null hypothesis. In particular, for each of these tests, we employed a different method to get rid of the unknown θ. For the asymptotic chi-square test, we estimated θ away by replacing it with the maximum likelihood estimator $\hat{\theta}$. In the case of Fisher's exact test, we avoided dealing with θ because we conditioned on both table margins. Finally, Barnard's CSM test (and the variants where we left out the C or S conditions) removes the dependence on θ by maximisation. These three techniques are just some of the possibilities when eliminating nuisance parameters, as illustrated by Basu [19].

3.1 Barnard and Fisher's initial correspondence

As can already be seen from the correspondence between Barnard and Fisher in *Nature* throughout 1945 [20]–[22], the introduction of the CSM test sparked a lot of discussion about the way one should eliminate the nuisance parameter in exact tests. Estimating θ away in the case of asymptotic testing is a common practice, and does not come with much controversy. However, in the case of exact testing, whether or not one should condition on the $(n_{\cdot 1}, n_{\cdot 2})$-margin, has been debated up to this day.

In 1947, Barnard elaborated on his CSM test, which he claimed was a more powerful alternative to Fisher's exact test [13]. An often expressed criticism of Fisher's test is indeed that it is a "very conservative" test. This has been mentioned, amongst others, by Pearson [23], Gail and Gart [24] and D'Agostino, Chase and Belanger [25], who pointed out via computational examples that the size of the exact test is in many cases, in particular with small sample sizes, a lot smaller than the nominal significance level. They either explicitly discourage the use of the test, or limit its appropriateness to certain experimental setups. For example, in 1947 too, Pearson treated in great detail Barnard's three abstract pictures we have seen in 2.3: the comparative trial, independence trial and double dichotomy. He described the independence trial in a medical context, where $n_{1\cdot}$ out of $n_{\cdot\cdot}$ individuals would receive one treatment whereas the remaining $n_{2\cdot} = n_{\cdot\cdot} - n_{1\cdot}$

individuals would receive another treatment. A certain reaction "A" would be observed with x_{11} and x_{21} individuals respectively. The null hypothesis would then be, in the words of Pearson, that "while some individuals show reaction "A" and some do not, the result would be the same whichever treatment were applied as far as these $n_{\cdot\cdot}$ individuals are concerned". In other words, under the null hypothesis, there will always be $n_{\cdot 1} = x_{11} + x_{21}$ individuals who will react and $n_{\cdot 2} = n_{\cdot\cdot} - n_{\cdot 1}$ who will not, were we to repeat the experiment. It is for that reason that Pearson deems Fisher's approach of fixing all table margins as appropriate. According to Pearson, a repetition in this setting would entail solely a random reassignment of the treatments over all individuals. He also notes that a repetition might not be possible at all, as the reaction "A" could be for instance death, and the experiment only concerns the selected $n_{\cdot\cdot}$ individuals.

The distinction between the different experiment types which may lead to the same contingency table, is what according to Pearson led to the controversy between Fisher and Barnard. Later that year, Barnard published a second article, in which it became clear that the disagreement was rooted in the – in Fisher's eyes – ill-defined concept of a reference set in the Neyman–Pearson theory of hypothesis testing [26]. For one to be able to talk of a significance level, which is nothing else than a probability of a certain set of outcomes occurring, one should refer to a set of possible outcomes, over which one can compute this probability. Barnard describes a modified version of an example given to him by Fisher, in which it is not clear what the reference set should be.

Consider a bag of flower seeds; each flower known to be either white or purple. We want to test the null hypothesis that the proportion of white (w) and red (r) seeds in the bag is equal. That is, if we let ψ be the probability of a seed turning into a white flower, we want to test $H_0 \colon \psi = 1/2$ against $H_1 \colon \psi \neq 1/2$. To do this, we take $n = 10$ seeds out of the bag at random and plant them. There is a nonzero probability of the seed not coming out and hence not growing into a flower. If we let φ be the probability of a seed coming out, then the number S of seeds coming out is can be assumed to be binomially distributed with n trials and success probability φ. Only $S = 9$ of the 10 seeds eventually come out and lead to a flower. All of these 9 flowers are white. If we define C to be the number of white flowers, then $C \mid S = s$ is binomially distributed with s trials and success probability ψ. Depending on what we think should be the reference set, we can now come up with two different p-values for this observation. We could say that we should compute the p-value solely based on the possible outcomes of 9 seeds leading to flowers. This would imply we could reject the null hypothesis at the level $2 \cdot 2^{-9} = 2^{-8}$, since the outcome of 9 white flowers under the null hypothesis is just as unlikely as the outcome of 9 red flowers. However, we could also take as our reference set all possible outcomes of a hypothetical repetition of the experiment; in which another number of seeds might have led to flowers. In that case, we would reject the null hypothesis at the level $2^{-8}\varphi^9$ instead. Because of this, someone with green fingers (φ close to 1) would end up with a higher significance level than someone who is not a good gardener (φ closer to 0). Thus a worse gardener would be rewarded with stronger evidence against H_0, which makes no sense according to Fisher. Therefore, he argues, one should condition on anything which bears no information about the parameter of interest, ψ.

The same line of reasoning should be used in the setting of contingency tables,

to answer the question whether or not one should condition on the total number of successes/failures $(n_{\cdot 1}, n_{\cdot 2})$. Fisher was of the opinion that, just as the number of seeds which lead to a flower in the example, the total number of successes holds no information on the success probabilities in both experiment groups. Although Barnard initially attempted to resolve this issue in the framework of the CSM test in his second 1947 paper [26], he remained unsatisfied about this solution. Private correspondence with Fisher ultimately led Barnard in 1949 to reject his CSM test [27].

3.2 Ancillarity

The flower example actually contains the main idea in favour of conditioning on both table margins. It is the view of proponents of conditioning, which include Fisher, Yates, Cox and thus also Barnard, that when performing inference one should always condition on all **almost** *ancillary statistics*. The term *ancillary statistic* has originally been coined by Fisher [28]. Since then, a lot of effort has been spent defining and researching the concept (see for example the review by Ghosh, Reid and Fraser [29] and the references therein), with many different definitions as a consequence. We employ here the most general definition, as stated by Reid [30].

Definition 3.1. Given a parametric model $f_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\theta})$ for a random vector $\mathbf{X}$, we say that a statistic $\mathbf{T} := t(\mathbf{X})$ is *ancillary* for the parameter $\boldsymbol{\theta}$ if the distribution of $\mathbf{T}$ does not depend on $\boldsymbol{\theta}$.

In the case of the flower example, if we set $\mathbf{X} = (C, S)^T$, $\boldsymbol{\theta} = (\psi, \varphi)^T$, and $\mathbf{T} = S$, since the amount of seeds S that comes out contains no information about the probability ψ of a seed turning into a white flower, we can write

$$f_{\mathbf{X}}(c, s; \psi, \varphi) = f_{C|S}(c \mid s; \psi) f_S(s; \varphi) = \binom{s}{c} \psi^c (1-\psi)^{s-c} \binom{n}{s} \varphi^s (1-\varphi)^{n-s}, \quad (3.1)$$

That is, S is an ancillary statistic for ψ, and we should condition all inference about ψ on S [31]. We will return to this "Conditionality Principle" in Section 3.4. This example, where we condition on an ancillary statistic in order to make the inference about ψ as relevant to the data as possible, is the standard application of ancillary statistics.

However, in the presence of nuisance parameters, ancillary statistics are often used to try to remove the nuisance parameters [29], as is the case with hypothesis testing on contingency tables. If, in the flower example, we would regard φ as a nuisance parameter, we could see that by conditioning on S, we managed in (3.1) to isolate a part of $f_{\mathbf{X}}$ not involving φ, on which we could base our inference about ψ. Because the remainder f_S does not depend on ψ anymore, we do not lose any information on ψ by only considering $f_{C|S}$.

Notice the addition **almost** we made just before defining ancillarity. It turns out that, in the setting of Table 1.1, the margin $(n_{\cdot 1}, n_{\cdot 2})$ is actually not an ancillary statistic for the parameters (θ_1, θ_2). To show this, we follow the reasoning of Little [31]. In his paper, he rewrites the null hypothesis in (2.1) to $H_0 \colon \psi = 1$, where $\psi := \theta_1(1-\theta_2)/(\theta_2(1-\theta_1))$ is the odds ratio. Furthermore, he defines the odds

product $\varphi := \theta_1\theta_2/((1-\theta_1)(1-\theta_2))$. This reparametrisation allows us to rewrite the probability of observing Table 1.1,

$$\begin{aligned} f_{X_{11},X_{21}}(x_{11},x_{21};\theta_1,\theta_2) &:= P_{\theta_1,\theta_2}(X_{11}=x_{11},X_{21}=x_{21}) \\ &= \binom{n_{1\cdot}}{x_{11}}\theta_1^{x_{11}}(1-\theta_1)^{n_{1\cdot}-x_{11}}\binom{n_{2\cdot}}{x_{21}}\theta_2^{x_{21}}(1-\theta_2)^{n_{2\cdot}-x_{21}}, \end{aligned}$$

in terms of one parameter which is of interest to us for our hypothesis test; the odds ratio ψ, and one nuisance parameter φ. Using that $\sqrt{\psi\varphi} = \theta_1/(1-\theta_1)$ and that $\sqrt{\varphi/\psi} = \theta_2/(1-\theta_2)$, a quick computation shows that

$$f_{X_{11},X_{21}}(x_{11},x_{21};\psi,\varphi) = \binom{n_{1\cdot}}{x_{11}}\frac{\sqrt{\psi\varphi}^{x_{11}}}{\left(1+\sqrt{\psi\varphi}\right)^{n_{1\cdot}}}\binom{n_{2\cdot}}{x_{21}}\frac{\sqrt{\varphi/\psi}^{x_{21}}}{\left(1+\sqrt{\varphi/\psi}\right)^{n_{2\cdot}}}, \tag{3.2}$$

If $N_{\cdot 1} := X_{11} + X_{21}$ were an ancillary statistic for ψ, we would be able to factorise $f_{X_{11},N_{\cdot 1}}(x_{11},n_{\cdot 1};\psi,\varphi) = f_{X_{11}|N_{\cdot 1}}(x_{11} \mid n_{\cdot 1};\psi)f_{N_{\cdot 1}}(n_{\cdot 1};\varphi)$, just as in (3.1). However, using that $x_{21} = n_{\cdot 1} - x_{11}$, one can quickly see from (3.2) that

$$\begin{aligned} f_{X_{11},N_{\cdot 1}}(x_{11},n_{\cdot 1};\psi,\varphi) &= \binom{n_{1\cdot}}{x_{11}}\binom{n_{2\cdot}}{n_{\cdot 1}-x_{11}}\frac{\psi^{x_{11}}}{\left(1+\sqrt{\psi\varphi}\right)^{n_{1\cdot}}}\frac{\sqrt{\varphi/\psi}^{n_{\cdot 1}}}{\left(1+\sqrt{\varphi/\psi}\right)^{n_{2\cdot}}} \\ &= f_{X_{11}|N_{\cdot 1}}(x_{11} \mid n_{\cdot 1};\psi)f_{N_{\cdot 1}}(n_{\cdot 1};\psi,\varphi), \end{aligned} \tag{3.3}$$

where

$$f_{X_{11}|N_{\cdot 1}}(x_{11} \mid n_{\cdot 1};\psi) = \frac{\binom{n_{1\cdot}}{x_{11}}\binom{n_{2\cdot}}{n_{\cdot 1}-x_{11}}\psi^{x_{11}}}{\sum_{i=0\vee n_{\cdot 1}-n_{2\cdot}}^{n_{1\cdot}\wedge n_{\cdot 1}}\binom{n_{1\cdot}}{i}\binom{n_{2\cdot}}{n_{\cdot 1}-i}\psi^{i}}, \tag{3.4}$$

is a well-defined probability density function which no longer depends on φ but where

$$f_{N_{\cdot 1}}(n_{\cdot 1};\psi,\varphi) = \frac{\sqrt{\varphi/\psi}^{n_{\cdot 1}}\sum_{i=0\vee n_{\cdot 1}-n_{2\cdot}}^{n_{1\cdot}\wedge n_{\cdot 1}}\binom{n_{1\cdot}}{i}\binom{n_{2\cdot}}{n_{\cdot 1}-i}\psi^{i}}{\left(1+\sqrt{\psi\varphi}\right)^{n_{1\cdot}}\left(1+\sqrt{\varphi/\psi}\right)^{n_{2\cdot}}} \tag{3.5}$$

is still a function of ψ. Although we can factor out the nuisance parameter φ from the distribution of $X_{11} \mid N_{\cdot 1}$, the distribution of $N_{\cdot 1}$ is dependent on ψ. Hence $N_{\cdot 1}$ is not ancillary for ψ. Compare this to the flower example, where the second factor in (3.1) does no longer depend on ψ [1].

The reason the $(n_{\cdot 1}, n_{\cdot 2})$-margin is seen as almost ancillary, is nicely illustrated by Kalbfleisch and Sprott [32]. Just as Little, they split up the density $f_{X_{11},N_{\cdot 1}}(x_{11},n_{\cdot 1};\psi,\varphi)$ into a part which only depends on ψ, and another part containing both ψ and the nuisance parameter, where "any information contained in ψ is inextricably tied up with the unknown nuisance parameter". Instead of using the odds product φ as the nuisance parameter, they used θ_2, such that θ_1 could be written as $\psi\theta_2/(1-\theta_2+\psi\theta_2)$. We will repeat their argument using ψ and φ

[1] Kalbfleisch and Sprott [32] mention that although the distribution of $N_{\cdot 1}$ still depends on ψ, it would have still been possible for the distribution to carry no information on ψ whenever the nuisance parameter φ would be completely unknown. This would be the case whenever this distribution would only depend on a parameter $\phi(\psi,\varphi)$ that is an injective function of φ for each ψ, so that ψ is not identifiable. They argue that this cannot be achieved here however; (3.5) definitely contains some information on ψ.

instead, which should not – and will not – yield any different conclusions. The residual factor (3.5) contains some information about ψ. Therefore, by using only (3.4) for inference about ψ, we lose that information and Kalbfleisch and Sprott refer to the likelihood $\ell(\psi; x_{11}, n_{\cdot 1})$ coming from $f_{X_{11}|N_{\cdot 1}}(x_{11} \mid n_{\cdot 1}; \psi)$ as an approximate conditional likelihood. In order to say something about the amount of information that is lost by only considering (3.4) for inference, Kalbfleisch and Sprott look at what they define as the maximum relative likelihood of ψ

$$R_M(\psi) := \frac{\sup_{\varphi \in [0,\infty)} f_{N_{\cdot 1}}(n_{\cdot 1}; \psi, \varphi)}{\sup_{\psi, \varphi \in [0,\infty)} f_{N_{\cdot 1}}(n_{\cdot 1}; \psi, \varphi)}. \tag{3.6}$$

The argument now goes as follows. If this $R_M(\psi)$ is an almost constant function of ψ, the approximate conditional likelihood $\ell(\psi; x_{11}, n_{\cdot 1})$ will be very close to the actual likelihood coming from (3.3). Not knowing anything about the nuisance parameter φ could be seen as being able to choose any value for φ we want. In particular, for any value of ψ, we can choose φ as the MLE $\hat{\varphi}(\psi)$ of φ given ψ. If $R_M(\psi)$ is almost constant, this would make each value of ψ equally likely. That is, $R_M(\psi)$ and thus also $f_{N_{\cdot 1}}(n_{\cdot 1}; \psi, \varphi)$ carry little information on ψ if we do not know anything about φ. This would then justify conditioning on $(n_{\cdot 1}, n_{\cdot 2})$.

Kalbfleisch and Sprott investigated $R_M(\psi)$ for a specific contingency table with data from Chinn, Noell and Smith [33] on the effectiveness of dramamine in preventing seasickness, given in Table 3.1.

	Not Sick	Sick	
Dramamine	31	3	34
Placebo	18	12	30
	49	15	64

Table 3.1: Effectiveness of dramamine in preventing seasickness [33].

On the left in Figure 3.1, we have reproduced Figure 1 from Kalbfleisch and Sprott [32], which shows $R_M(\psi)$ and the relative conditional likelihood $CR(\psi) = \ell(\psi; x_{11}, n_{\cdot 1}) / \sup_{\psi \in [0,\infty)} \ell(\psi; x_{11}, n_{\cdot 1})$. We obtain the same picture whether we use the parametrisation (ψ, θ_2) or the parametrisation (ψ, φ). This should not come as a surprise, as $CR(\psi)$ and $R_M(\psi)$ are functions of only ψ anyway. For $R_M(\psi)$ we "maximised away" the nuisance parameter and we are taking ratios of likelihoods so there should not be any difference. Kalbfleisch and Sprott found the $R_M(\psi)$-function sufficiently constant in order to conclude that (3.5) carried a negligible amount of information on ψ, and that consequently conditioning on $(n_{\cdot 1}, n_{\cdot 2})$ is justified.

However, when changing the data to the contingency table shown in Table 3.2, repeating the same procedure yields the right plot in Figure 3.1. Although on a different scale, the $CR(\psi)$-function looks rather similar to the one of the original data. However, it is quite a stretch to claim that $R_M(\psi)$ is also an approximately constant function of ψ right now. The argument in favour of conditioning does not seem to hold up in this case. Whether or not the $(n_{\cdot 1}, n_{\cdot 2})$-margin can be seen as ancillary seems to be dependent on the specific contingency table under consideration.

Finally, as a small side note, it is worth mentioning that in the case of a 2×2 independence trial and a 2×2 comparative trial, the sample sizes $n_{1\cdot}$ and $n_{2\cdot}$ are

	Not Sick	Sick	
Dramamine	31	3	34
Placebo	2	28	30
	33	31	64

Table 3.2: Alternative outcome to the effectiveness study of dramamine.

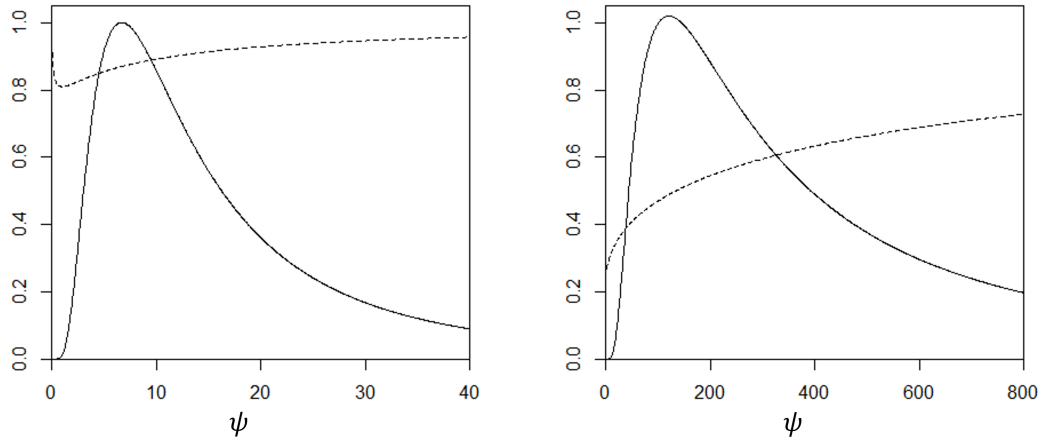


Figure 3.1: $CR(\psi)$ (solid) and $R_M(\psi)$ (dashed) for Tables 3.1 (left) and 3.2 (right).

fixed. In particular, their distribution (as far as we can speak of one) does not depend on (θ_1, θ_2), and hence $(n_{1\cdot}, n_{2\cdot})$ can be seen as an ancillary statistic for (θ_1, θ_2). After all, conditioning on the $(n_{1\cdot}, n_{2\cdot})$-margin does not result in controversy.

3.3 Berkson's dispraise

For Berkson, the fact that $N_{\cdot 1}$ is not ancillary for ψ is one of the main reasons to denounce the conditional exact test [34]. In his 1978 paper, he compares the size and power of Fisher's exact test with that of the chi-square test (with and without Yates' correction) for given nominal significance levels, for a one-sided alternative hypothesis, with different values of $n_{1\cdot} = n_{2\cdot}$. Just as with many of the critiques of Fisher's test, Berkson stresses the conservative behaviour of the exact test. The size of the test is, for many of the situations Berkson worked out, "considerably smaller" than the nominal significance level.

Apart from this numerical comparison, Berkson considers the ancillarity argument. He refers to Fisher's 1935 paper, in which the latter said that

> "if it be admitted that these marginal frequencies by themselves supply no information on the point at issue, namely, as to the proportionality of the frequencies in the body of the table, we may recognise the information they supply as wholly ancillary; and therefore recognise that we are concerned only with the relative probabilities of occurrence of the different ways in which the table can be filled in, subject to these

> marginal frequencies." [35]

Berkson questions what is meant by "information that is wholly ancillary", and continues to argue that the $(n_{\cdot 1}, n_{\cdot 2})$-margin contains information about the counts in the table.

First, he points to one of his other writings [36], in which he looks at all possible outcomes of the contingency table with $n_{1\cdot} = n_{2\cdot} = 5$ and computes the Fisher p-values in order to perform the one-sided test $H_0\colon \theta_1 = \theta_2$ against $H_1\colon \theta_1 > \theta_2$ at the level $\alpha = 0.05$. Because the test is one-sided, he only looks at the tables where $\hat{\theta}_1 \geq \hat{\theta}_2$, i.e., $x_{11}/n_{1\cdot} \geq x_{21}/n_{2\cdot}$ and thus $x_{11} \geq x_{21}$. This is a total of $|\{(x_{11}, x_{21}) : 0 \leq x_{21} \leq x_{11} \leq 5\}| = 21$ tables. In the cases that $n_{\cdot 1} = 0$ or $n_{\cdot 1} = 10$, the table is trivially exactly determined, as we need $x_{11} = 0$ or $x_{11} = 5$ respectively. The corresponding Fisher p-value is 1, and there is no significant difference in the observed success proportions (they are either both 0 or both 5). In total, for 8 out of the 11 possible values of $n_{\cdot 1}$, Fisher's exact test does not indicate a significant difference at the $\alpha = 0.05$ level, while for the remaining three values of $n_{\cdot 1}$, namely 4, 5 and 6, there was a 1/3 probability of observing a significant difference (for the tables $(4, 0)$, $(5, 0)$ and $(5, 1)$). This, for Berkson, was evidence that the $(n_{\cdot 1}, n_{\cdot 2})$-margin bears information on the proportionality of θ_1 and θ_2.

As a second argument for that, Berkson also cites Plackett [37], in which the latter tries to estimate $(\lambda, \phi) := (\log(\theta_1(1-\theta_2)/(\theta_2(1-\theta_1))), \log(\theta_2/(1-\theta_2)))$, which is the linear logistic transformation of (θ_1, θ_2), via the likelihood approach. Plackett shows that $\lambda \in \{0, \pm\infty\}$ is required in order for the score function to be zero. At $\lambda = 0$ (and $\phi = \log(n_{\cdot 1}/n_{\cdot 2})$, so $\theta_1 = \theta_2 = n_{\cdot 1}/n_{\cdot\cdot}$), the likelihood reaches a saddle point, with two rising edges as $\lambda \to \pm\infty$, which are asymptotically horizontal. These asymptotic maxima correspond to

$$\theta_1 = \frac{n_{\cdot 1}}{n_{1\cdot}} \mathbb{1}_{\{n_{\cdot 1} \leq n_{1\cdot}\}} + \frac{n_{\cdot 2}}{n_{2\cdot}} \mathbb{1}_{\{n_{\cdot 1} > n_{1\cdot}\}},\ \theta_2 = 0,\ \text{for } \lambda = \infty, \tag{3.7}$$

and

$$\theta_1 = 0,\ \theta_2 = \frac{n_{\cdot 1}}{n_{2\cdot}} \mathbb{1}_{\{n_{\cdot 1} \leq n_{2\cdot}\}} + \frac{n_{\cdot 2}}{n_{1\cdot}} \mathbb{1}_{\{n_{\cdot 1} > n_{2\cdot}\}},\ \text{for } \lambda = -\infty. \tag{3.8}$$

Because these ML estimates are purely functions of the margins, Berkson said that it "would hardly be possible if the marginal totals contained no information for judging whether $\theta_1 = \theta_2$" [34]. Plackett himself on the other hand, was less convinced. He felt that his findings were actually corroborating Fisher's "intuitive view that the likelihood function provides little information about λ" [37]. For example, he mentioned that for $n_{1\cdot}, n_{\cdot 1}, n_{\cdot\cdot} \to \infty$ in fixed proportions, the score equation is satisfied for all values of λ. Furthermore, Plackett shows that "application of likelihood ratio tests for $H_0\colon \lambda = 0$ has been inconclusive".

Interestingly enough, Berkson does also sketch a line of reasoning in which he sees how someone could prefer the exact test. In particular, according to Berkson,

> "it may be argued that, once the marginal totals have been observed, it does not matter what might have happened, or what will happen in the next experiment or what plan was in the mind of the experimenter when he obtained the data, we judge on the evidence before us." [34]

This is reflected by Plackett's ML estimators (3.7) and (3.8), which do not care about how we came to the observations we observed; they are the same regardless. This seems to agree with Fisher; who also sees the marginal totals as a "given", on which one should condition in order to perform any further inference. Failing to do so would take into account experiment outcomes which are not relevant. Consequently, the table outcomes given the marginal totals are hypergeometrically distributed, and Berkson agrees that Fisher's exact test with randomisation is the appropriate one. It seems that Berkson's issue with Fisher's position lies more in the statement that the marginal totals contain (almost) no information about the table, instead of the act of conditioning itself.

The randomised exact test, where we perform Fisher's exact test in combination with an independent Bernoulli experiment in order to exactly reach the desired level of significance, has been shown by Tocher [38] to be the uniformly most powerful unbiased (UMPU) test for testing $H_0 \colon \theta_1 = \theta_2$ against a one-sided alternative ($H_1 \colon \theta_1 > \theta_2$ or $H_1 \colon \theta_1 < \theta_2$). Recall that a test with significance level α and power function $\beta(\boldsymbol{\theta})$ is called uniformly most powerful unbiased (UMPU) if for all other unbiased tests of level α with power function $\beta'(\boldsymbol{\theta})$, we have $\beta(\boldsymbol{\theta}) \geq \beta'(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \Theta_1$. A test is unbiased if there are no values of $\boldsymbol{\theta} \in \Theta_1$ for which acceptance of H_0 is more likely than for some cases in which H_0 is actually true. That is, $\beta(\boldsymbol{\theta}) \leq \alpha$ for $\boldsymbol{\theta} \in \Theta_0$ and $\beta(\boldsymbol{\theta}) \geq \alpha$ for $\boldsymbol{\theta} \in \Theta_1$ [39].

Although one might interpret the randomised version of Fisher's exact test being UMPU as the response to the power argument often used by proponents of unconditional tests, one should keep in mind that randomised tests are unacceptable for practitioners. The conclusion of the test should not be dependent on some random event unrelated to the actual experiment. Furthermore, one should also recall that a test being UMPU is not the same thing as a test being uniformly most powerful (UMP). Indeed, Suissa and Shuster [40] object to striving for a test to be UMPU as if it is some kind of optimality criterion. They show that when testing the equality of two binomial proportions, one can in fact come up with a biased test that is more powerful than the randomised version of Fisher's exact test.

3.4 The Conditionality Principle

Essentially, Berkson brings up the question to which extent one should accept the Conditionality Principle (CP), which states, according to Little [31], that one should always condition on all ancillary statistics for a parameter in order to do inference on that parameter. It seems that Little argues that should one be in favour of the conditional test, then one should agree with CP. However, accepting CP has implications of its own. Birnbaum [41] proved that if one considers this Principle to be true, and additionally accepts the Sufficiency Principle, defined in Definition 3.2, one should logically also accept the controversial Likelihood Principle, defined in Definition 3.3. The definitions are due to van der Meulen [6].

Definition 3.2 (Sufficiency Principle; SP)**.** If T is a sufficient statistic for the parameter ϑ, then two observations $\mathbf{x}$ and $\mathbf{x}'$ which satisfy $T(\mathbf{x}) = T(\mathbf{x}')$ must lead to the same inference on ϑ.

This Principle essentially motivates the definition of a sufficient statistic; all the information the data carries about ϑ can be summarised with a sufficient test statistic.

Definition 3.3 (Likelihood Principle; LP)**.** All the information an observation $\mathbf{x}$ carries about ϑ is fully contained in the likelihood function $L(\vartheta; \mathbf{x})$. If for two observations $\mathbf{x}$ and $\mathbf{x}'$ depending on ϑ, possibly via different experiments, there exists a constant c such that $L(\vartheta; \mathbf{x}) = cL(\vartheta; \mathbf{x}')$ for every possible ϑ, both observations lead to the same inference on ϑ.

The LP does not agree very well with the significance testing we have done so far. Little mentions the classical example of testing the null hypothesis $H_0 \colon \vartheta = 1/2$, with ϑ the success probability of independent Bernoulli trials. We can now come up with two experiments; one with a fixed number of 6 trials in which we measure the amount of successes, and one where we perform trials until we have reached one success. If in the former experiment we observe one success, and if in the latter we need to perform 6 trials, both experiments will yield the likelihood function $\vartheta(1-\vartheta)^5$. The first (binomial) experiment is easily shown to yield a p-value of $7/64$, while the second (negative binomial) experiment yields a p-value of $1/32$. This clearly violates the Likelihood Principle.

Although not explicitly stating it, it seems that Little implies that proponents of the conditional test, essentially apply the CP and must therefore also accept the LP, which completely denounces classical hypothesis testing (including the conditional test itself). It is unclear how serious one should view this apparent contradiction. On one hand, as one is actually allowing conditioning on almost ancillary statistics with the conditional test, we doubt whether Birnbaum's result would still hold. Indeed, in the proof of the result, Birnbaum considered two experiments which yielded proportional likelihoods, and introduced a mixture experiment which picked one of the two experiments depending on the value of a binary ancillary statistic. However, if now instead of an ancillary statistic, we are also allowed to base the mixture experiment on an almost ancillary statistic, conditioning on the outcome of this statistic need not yield one of the two original experiments.

On the other hand, Birnbaum's theorem itself has generated a fair amount of debate of its own. Doubt has been cast on the validity and applicability of Birnbaum's formulations of the CP and SP, questioning for example what it means to lead to the same inference about θ [2]. Statisticians not so eager to accept the LP have shown that slightly different formulations of CP do not imply LP when furthermore assuming SP. For example, Durbin [42] proved that Birnbaum's result does not hold if we require to only condition on ancillary statistics which are functions of a minimal sufficient statistic. Another example is the proposal by Kalbfleisch [43], who showed that in order to avoid accepting LP, one could apply a CP that only required conditioning on ancillary statistics which are an actual part of the experimental setup. Meanwhile, statisticians willing to accept LP have come up with modifications to make Birnbaum's statement more precise, amongst others Wechsler, de B. Pereira and Marques F. [44] and Grossmann [45]. Note that Bayesian statistics is entirely based on the likelihood function, so accepting

[2] In his article, Birnbaum[41] stated the three Principles in terms of a concept called *evidence*, also leaving a certain amount of room for interpretation.

LP is a logical thing to do for statisticians of the Bayesian school. Little [31] even finishes his paper by arguing that Bayesian methods are the most suitable to deal with nuisance parameters as in the setting of contingency tables.

The debate is far from settled, with criticisms as recent as 2013 by Evans [46] and 2014 by Mayo [47], and replies by amongst others Peña and Berger [48]. This whole discussion, which revolves around the fundamental principles of statistical inference, and in particular the Conditionality Principle, and the way in which Birnbaum's result plays into this, is very interesting. However, we feel that treatment of this rabbit hole would better be suited for a thesis of its own, and will therefore not dive deeper into the topic here.

Before heading back to the main discussion of this Chapter though, we would like to highlight the perspective of Helland [49]. His article looks into several examples which should make the reader realise that the validity of CP is maybe not as obvious as one might initially think. Helland motivates this by mentioning the work of Evans, Fraser and Monette [50], [51], which shows that in many settings CP alone is actually equivalent to LP. Like this, Helland offers an escape to all those statisticians in the impossible position of not accepting LP while finding CP (and SP) rather logical.

The key point of the paper is that since LP is such a universal statement (applying to any experiment, no matter the setup and underlying assumptions, as only the likelihood function matters), CP is an equally universal statement. Helland however argues via examples that the universality of CP is "unreasonable". We cannot say beforehand that we should condition on all ancillary statistics, but we should take into account the "target population" and the "questions of interest". This view goes directly against LP, but more importantly for us, it also questions the position of Fisher by asking when it is appropriate to condition on ancillary statistics. Helland proceeds by working out an example in which we have two patients who followed the same treatment. Before and after the treatment we measure some response X_h that is normally distributed with mean μ_h and known variance σ^2. Here $h \in \{1, 2\}$ is an index representing the patient. Because both patients had the same treatment, Helland is interested in the mean of the means $\vartheta = (\mu_1 + \mu_2)/2$. In other words, we are interested in inference on ϑ, where $\mu_1 = \vartheta - \delta$ and $\mu_2 = \vartheta + \delta$, where δ is a nuisance parameter. He now argues that neither experiments E_1 nor E_2, which measure X_1 and X_2 respectively, can provide us with any information on ϑ that does not depend on the unknown δ. However, the mixture experiment of choosing either E_1 or E_2 with probability $1/2$ does tell us something about ϑ, as the response X we measure in E is an unbiased estimator for ϑ. Alternatively, the only difference between E and a random sampling is that we have a nonzero σ^2; were we to let $\sigma \to 0$, then we would just pick one of the two patients as our random sample. This should carry some information about ϑ, in contrast to choosing a fixed sample. Thus, even though the Bernoulli random variable $H \in \{1, 2\}$ which decides which experiment to choose is ancillary, we seem to lose information by conditioning on it and performing either E_1 or E_2 instead of E. Later on, Helland explicitly mentions the debate regarding conditioning in contingency tables, first agreeing with Little [31] that the question is not about whether or not to condition on ancillary, but on approximately ancillary, statistics. Second, as we will see in Section 3.7, he shares the view of Greenland [52] that the question contingency tables try to answer is too broad to confidently say whether

or not to condition in every possible situation. One should look on a case-by-case basis, and not blindly follow CP, or an approximate ancillary variant of it.

3.5 Reactions to Berkson's work

The 1978 paper by Berkson has often been cited by other opponents of the conditional test. For example, Kempthorne [53] agrees for the most part with Berkson, and argues additionally that Fisher's exact test should be used for both the independence trial and the double dichotomy, while Barnard's CSM test is appropriate for the comparative trial. Upton [54] also advocates against Fisher's exact test. He discusses 22 different tests for the 2×2 comparative trial. Interesting is that he did not include Barnard's CSM test because of its high computational burden. Most of the tests considered were variants of the chi-square test (and hence asymptotic), with a few exact tests such as Fisher's test or Boschloo's test. Based on several criteria related to the size of the test and its possible dependency on the nuisance parameter, Upton came to the conclusion that for a comparative trial, Fisher's exact test should not be used. He instead argues in favour of using a scaled version of Pearson's chi-square test statistic. Although not the subject of his paper, Upton agrees with Kempthorne that, for other experimental setups like the independence trial and the double dichotomy, Fisher's test should be used.

On the other hand, proponents of the conditional test also seemed to treat Berkson's paper as an important representative of those who were against conditioning. In 1979, Barnard [14] wrote a short note in reply to Berkson's work. In that note, he first reiterates the difference between the comparative trial and the independence trial, and points out that Berkson confuses the two terms in one of his examples. This example is essentially the same as the medical trial we discussed in Chapter 1, which resulted in Table 1.1. Berkson considers a clinical trial with $n_{\cdot\cdot} = 70$ participants of which half is randomly assigned to a treatment A and the other half to treatment B. Of interest is how many participants recovered for each treatment group. The results of this hypothetical study are shown in Table 3.3. According to Berkson, this clinical trial is a comparative trial in which we compare

	Cured	Not Cured	
A	30	5	35
B	24	11	35
	54	16	70

Table 3.3: The 2×2 contingency table from Berkson's example [34].

the recovery probabilities for each treatment. This is again under the assumption that we can actually speak of such probabilities, something which Barnard does not think we should do. He mentions that the reason to randomise the groups in the first place is to take into account the differences between the participants that might affect their chance of recovery. Under the null hypothesis that treatments A and B have the same effect, participants cured by A would also have recovered by B. Therefore, argues Barnard, the amount of 54 cured participants was predetermined and if we want to test whether A and B are significantly different, we should consider as our reference set only the table outcomes where the

54 cured people are distributed at random over the two treatments. That is, the $(54, 16)$-margin in Table 3.3 should also be kept fixed. This clinical trial is thus in fact an independence trial, for which we should be using Fisher's exact test, even according to Berkson. This same argument has been set forth by Mehta and Hilton [55].

Regarding comparative trials, Barnard acknowledges that "the situation is less clear" [14]. Barnard ascribes the differences between the results of the conditional and unconditional/asymptotic tests to the discreteness of the set of attainable p-values, instead of to a negligible loss of information due to conditioning on the almost ancillary $(n_{\cdot 1}, n_{\cdot 2})$-margin. We will return to this point in a bit.

Another reply to Berkson's work, as well as to the papers of, amongst others, Kempthorne and Upton, came from Yates [3] in 1984. This text has been applauded by proponents of the conditional testing approach, in particular Barnard [56] and Cox [57]. Both mention in their discussion of this paper that it would finally, after 70 years of arguments, settle the debate in favour of conditioning. After a brief historical account which finishes with Barnard's disavowal of his CSM test, Yates moves on to some numerical examples. With these examples, he wants to indicate that the $(n_{\cdot 1}, n_{\cdot 2})$-margin contains (almost) no information on whether or not $\theta_1 = \theta_2$. No matter how many margins were fixed beforehand, it seemed "obvious" to Yates (and Fisher) that one should condition on both.

In particular, Yates seemed to argue that in the cases where the $(n_{\cdot 1}, n_{\cdot 2})$-margin would say something about the difference between θ_1 and θ_2, the contents $(x_{11}, x_{12}, x_{21}, x_{22})$ of the table would be able to tell a lot more about the actual values of θ_1 and θ_2. For example, if $n_{\cdot\cdot}$ were to be very large and $n_{1\cdot} \approx n_{\cdot 1}$, like in Table 3.2, then we might suspect that there is a large difference between θ_1 and θ_2. Indeed, for example, if $\theta_1 = 1, \theta_2 = 0$ then we will always have $(n_{\cdot 1}, n_{\cdot 2}) = (n_{1\cdot}, n_{2\cdot})$. This corresponds with what we have seen in the right plot of Figure 3.1, which indicated that the conditioning on the margin did in fact lead to some loss of information. At the same time however, Yates argues that the large value of $n_{\cdot\cdot}$ entails that the empirical estimates $\hat{\theta}_1 = x_{11}/n_{1\cdot}$ and $\hat{\theta}_2 = x_{21}/n_{2\cdot}$ will already be rather informative on the actual values of θ_1 and θ_2, and that in fact the $(n_{\cdot 1}, n_{\cdot 2})$-margin does not tell us that much extra.

He continues by remarking that in a discrete setting, reducing the set of possible outcomes to a smaller reference set, inevitably reduces the level of significance of the more extreme outcomes. According to Yates, this combined with the "urge to find more powerful tests" explains why people are drawn to unconditional tests in the discrete setting, even though conditional testing has been widely accepted in the continuous setting. Related to this, is the use of nominal significance levels, such as 0.05 or 0.01, with discrete data in the first place. The classic example of this is tossing a coin ten times to test if it is biased, i.e. $H_0 \colon p = 1/2$ against $H_1 \colon p \neq 1/2$, where p is the probability of the coin giving heads. Under H_0, the probability of throwing 8 heads or more, or 2 heads or less is $2^{-10}(1 + 10 + 45 + 45 + 10 + 1) = 0.11$, while the probability of throwing 9 heads or more, or 1 head or less is $2^{-10}(1 + 10 + 10 + 1) = 0.021$. If we observe 9 heads, we would reject the null hypothesis at the $\alpha = 0.05$ level, even though this "preassigned magic level", as Cox [57] put it, can never be exactly equal to the size of the test. Thus, in the discrete case, one should therefore always report the actual significance level (read: p-value) instead. This makes it more difficult to speak of a "more powerful"

test in the first place. In the Neyman–Pearson approach for testing, it only makes sense to compare the power of two tests which have the same level of significance. This we can only achieve, in general, by making use of randomised tests, which is undesirable in practice.

Barnard [14] made a similar observation in his 1979 note, advocating for flexibility in significance levels when working with discrete distributions. He came back to this in a paper from 1989 [58], saying that we should not adhere to one fixed nominal significance level, to be used under all circumstances. For example, in planning an experiment, we would set ourselves a certain target sample size. However, if for some reason the actual sample size $n_{\cdot\cdot}$ turned out differently, Barnard would find it strange to not alter the pre-determined significance level we would have chosen too. If the actual sample size were to be smaller than the target sample size, Barnard argues, we should consider raising the significance level. By doing so, we would slightly increase the probability of making a Type I error, but at the same time lower the probability of a Type II error. Barnard thinks the same way about the $(n_{\cdot 1}, n_{\cdot 2})$-margin. If we would work with the actual sample size $n_{\cdot\cdot}$ instead of the target sample size, why would we not base our inference on the actually achieved values of $(n_{\cdot 1}, n_{\cdot 2})$ instead of all possible values of $(n_{\cdot 1}, n_{\cdot 2})$, if this comes at no (significant) loss of information? Just as Fisher and Yates, Barnard treats the $(n_{\cdot 1}, n_{\cdot 2})$-margin as if it were ancillary.

3.6 The debate after Yates' paper

Contrary to the hopes of Yates himself, Barnard and Cox, Yates' 1984 paper would not, in the words of Cox, "squash once and for all various misconceptions" regarding conditional tests. As pointed out by Upton [59] in the second paper he wrote on the subject, the discussion was far from over; numerous articles on the issue followed Yates' 1984 work. Let us mention some of them.

Suissa and Shuster [15] propose an exact unconditional test based on the unpooled Z_{u}^2 statistic defined in (2.5) (in other words the chi-square test statistic); essentially boiling down to a p-value test with p-value determined by (2.23) where $T = -Z_{\mathrm{u}}^2$ such that we reject H_0 for large values of T, along with a procedure to maximise with respect to the nuisance parameter. Their view on the debate is predominantly a pragmatic one. Again, the power argument shows up, now in the form that for a given Type I error probability and power, one needs smaller sample sizes using unconditional tests. Moreover, Suissa and Shuster point out that results of unconditional tests are more easily explainable. In contrast with a conditional p-value from for example Fisher's exact test, an unconditional p-value can simply be interpreted as the maximum possible probability of obtaining an at least as extreme observation, were we to repeat the experiment (with the group sizes $n_{1\cdot}$ and $n_{2\cdot}$ fixed) in a setting where H_0 is true. Of course, this assumes that a repetition of the experiment need not to have the same amount of successes. They however also acknowledge that the conditional test requires simpler and fewer computations.

A couple of years later, Barnard [58] wrote *On alleged gains in power from lower p-values*. As we mentioned in Section 3.5, he argued for a more flexible use of significance levels. If one observed a certain value of $n_{\cdot 1}$, one could choose a

significance level from the set of attainable p-values for that given $n_{\cdot 1}$. By appropriately choosing significance levels for each $n_{\cdot 1}$, Barnard managed to construct a test procedure which would yield the exact same conclusion as, say, the chi-square test. In Barnard's eyes, this addresses the apparent loss of power of Fisher's exact test. By getting rid of this idea that one should have a fixed significance level over all possible values of $n_{\cdot 1}$, Fisher's exact test no longer seems to "underrate the strength of evidence against H_0" [58].

This argument has been repeated by Upton [59] in a paper in which he elaborated why he changed his mind and was now in favour of conditioning. Upton denounced his previous line of reasoning that "if users of a test believe that it has a type I error equal to α, then their belief should not be too far from the truth" [59]. He argues that one should not stick to a pre-determined nominal significance level that is more often than not unattainable. Instead, one should only work with the set of significance levels that can be reached given the $(n_{\cdot 1}, n_{\cdot 2})$-margin. Like that, one can no longer view Fisher's exact test as conservative.

It seems that one's opinion on the conditioning matter boils down to how one wants to treat p-values, or more generally probabilities, in the setting of discrete distributions. We find the take of Camilli [60] on this very elucidating. On one hand, the Fisherian school of thought is of the opinion that researchers should base conclusions on their hypotheses purely on the available data (essentially the Likelihood Principle). On the other hand, the Neyman–Pearson school argues that conclusions should be drawn by comparing the observed result with all alternative results that could have been observed instead. With the latter approach, we fix a nominal significance level α, which should represent the amount of times that we make a Type I error in a hypothetical (infinite) repetition of the experiment. This is linked to the frequentist point of view that a probability of some experiment outcome is the proportion of occurrences of that outcome "in the long run", i.e., as the amount of experiments grows to infinity. According to Rice [61], this is the interpretation most preferred by "practising empiricists".

Fisher objected that one cannot infinitely (and identically) repeat an experiment. Even if a replication is carried out, the sample and/or experiment circumstances change. Although Fisher viewed a probability, just as the frequentists, as a "limiting ratio of a set of events", he required furthermore that one could not recognise subsets of events. Camilli mentions as an example, that before tossing a coin, we cannot recognise any set of circumstances which would affect the outcome of the coin toss. In particular, this rejects the Neyman–Pearson interpretation of the probability of an observed table outcome, where we define "recognisable subsets" based on the value of the $(n_{\cdot 1}, n_{\cdot 2})$-margin. Table outcomes with another value of this margin are not deemed relevant to find out whether or not there is an association between the two groups.

As an example, let us consider an experiment mentioned by Routledge [62], where fish and bacteria are added to six fish tanks. Three out of the six fish tanks were selected at random to be treated with ozone, and after a while it was recorded that only the untreated tanks contained some dead fish. This is shown in Table 3.4. One wanted to test the null hypothesis that treating the fish tank with ozone had no effect on the survival of the fishes inside those tanks, against the one-sided alternative hypothesis that treating the fish tank would improve the chances of survival of the fishes. It can easily be found that $1/20 = 0.05$ is the one-sided Fisher

	Some dead fish	No dead fish	
Treated	0	3	3
Untreated	3	0	3
	3	3	6

Table 3.4: The 2×2 contingency table from Routledge's example [62].

p-value, while the one-sided CSM p-value is $1/64 \approx 0.016$. The table configuration in Table 3.4, with $n_{\cdot 1} = 3$, is the only one for which Fisher's exact test rejects the null hypothesis. Fisher argued that although "unhelpful outcomes" such as the tables with $n_{\cdot 1} = 0$ (no dead fish) or $n_{\cdot 1} = 6$ (some dead fish in all tanks) can occur with positive probability, they should for sure not affect our judgement about the potential effect of treating the tanks, as is the case when using an unconditional test. These outcomes are unhelpful in the sense that, given $n_{\cdot 1} = 0$ or $n_{\cdot 1} = 6$, it is impossible to get any evidence in favour of the alternative hypothesis.

As a side note, Routledge goes even further by questioning the binomial model we have required so far for the unconditional test. We mentioned in Section 2.1 that viewing the medical experiment we considered back then as a binomial trial is a perhaps indefensible assumption. Routledge mentions that under such a model, the sum $X_{11} + X_{21}$ should also be binomially distributed. However, due to time pressure, or limited resources, one might alter the experimental procedure in such a way that observing a outcome with $X_{11}+X_{21} = 3$ (the only one which can lead to a significant conclusion!) becomes more probable than the assumed binomial probability. Such changes are certainly not always reported [62]. For example, it turned out that in the fish tank experiment, the experiment would have been repeated once more if the first try would not have been able to yield significant results, i.e., if $X_{11} + X_{21} \neq 3$. This alters the underlying probability model, rendering the binomial assumption, and so the unconditional test, inaccurate. Because of this, Routledge advocated for more precise descriptions of the experimental procedure. He proceeded to argue that this aim, amongst others, would get rid of much of the discrepancies between conditional and unconditional p-values. Since conditional tests do not require a – possibly complicated – analysis of how the contingency tables are distributed if we step away from the binomial assumption, he eventually advocated for the use of Fisher's exact test.

Back to Fisher, who only wanted to look at the data that has actually been observed. This restricted the sample space to only those contingency tables which had the same marginal totals. Camilli [60] points out that within this sample space, Fisher could be seen as a frequentist too, computing the (hypergeometric) probability of observing the observed table, or a more extreme one, just as with the Neyman–Pearson approach. On the other hand, Fisher could also be seen as a subjectivist, as he interpreted the computed probability as a "measure of reluctance to accept H_0 – to be defined by each individual – within the confines of one particular experiment" [60]. This differs from the Neyman–Pearson view in the sense that Fisher did not want to attach a probability to a hypothesis about the real world. In his own words, "tests of significance are based on *hypothetical* probabilities calculated from their null hypotheses. They do not generally lead to any probability statements about the real world" [63].

Camilli concludes by expressing his preference for Fisher's exact test, as the

CSM test is built around the Neyman–Pearson idea of long-run experiments. This is a perspective Camilli views as inadequate to say something about the meaning of the observed data "in the short run". Just as Barnard [14] proposed, Camilli recommends using Fisher's exact test in combination with a flexible nominal significance level. Even better, he would like to see the label "significant/not significant" be entirely replaced by just a mentioning of the significance level.

3.7 The current state of the debate and how this thesis fits in

The contingency table is the most straightforward and natural way to summarise the results from an experiment with categorical outcomes. It is therefore amazing that the debate on which type of significance test to use has been going on for well over a century by now. A quick search on Google Scholar for research articles in the fields of medicine or biology that have results in the form of a contingency table seems to indicate that the chi-square test and Fisher's exact test are popular under practitioners [64]. This makes it unclear if there is any room or justification for the exact unconditional test. By extension, as this thesis predominantly focuses on the unconditional approach, proponents of the conditional test may find the findings that will follow in Chapters 4 and 5 utterly useless.

However, up until this day, we have seen throughout this Chapter that several authors have argued in favour of unconditional tests. Even some authors, such as Macdonald [65] and Greenland [52], have voiced their preference for conditional tests, but also described some specific cases in which the unconditional test could be appropriate too. We would like to focus on the work of the latter author, who justified the use of conditional tests not via the usual ancillarity argument, but rather by deducing a logical absurdity from the use of unconditional tests under certain causal models.

First of all, Greenland repeats the argument from Barnard [14] that stated that unconditional tests are not appropriate for inference of the experiment set out in Chapter 1 (or in Table 3.3). In a hypothetical repetition of the experiment, we would be dealing with the same individuals and thus, under H_0, the treatment would have no effect and the total number of recoveries $n_{\cdot 1}$ will be the same. Hence, both margins should be treated as fixed. It is important to realise here that this argument only works because we are testing a hypothesis that only concerns the individuals involved. That is, we observe the entire population. If we would instead look at the experiment where we only record the results for a random sample of the population, this sample would change at every hypothetical repetition of the experiment: $n_{\cdot 1}$ is thus no longer the same each time.

Greenland argues however that even in this case of random sampling, the conditional test is the correct one. However, we should be very careful defining what we mean with random sampling. Greenland spent a great amount of time describing very precisely the data-generating model, for it is his view that a poor description of the underlying experimental model is the root cause for the whole debate on whether or not to condition. Shuster [66] seemed to agree with this explanation, something Greenland confirmed in personal communication.

The model we will now consider is often referred to as Rubin's model. Each

member $i \in \mathcal{P}$ of a target population $\mathcal{P}$ has a nonrandom response $r_i(x) \in \{1, 2\}$ if subjected to a treatment $x \in \{1, 2\}$. We want to test what Greenland calls the *Sharp Causal null hypothesis about the Population* (SCP),

$$H_{\text{SCP}} \colon \forall i \in \mathcal{P} \colon r_i(1) = r_i(2).$$

That is, under this null hypothesis, the response of each individual will be the same no matter the treatment given. Greenland stresses that $r_i(1)$ and $r_i(2)$ are "responses of the *same* individual under *different* conditions and that the hypothesis refers to a *single* population". In order to test this hypothesis, we would like to observe both $r_i(1)$ and $r_i(2)$. This is impossible in this model; we cannot subject i to both treatments simultaneously and will therefore only measure either $r_i(1)$ or $r_i(2)$ for each individual. This "missingness problem" is often tackled by randomisation. We randomly sample without replacement $n_{\cdot\cdot}$ individuals. Let us call this sample $\mathcal{S}$. For each individual, we then choose at random which of the $r_i(x)$ we wish to observe by assigning them to either treatment $x = 1$ or treatment $x = 2$. This effectively turns x into a random variable, determining which entry of the outcome vector $(r_i(1), r_i(2))$ will be missing for each individual in the sample. Note that the random variable x is not defined for the individuals not in $\mathcal{S}$.

The total number $n_{\cdot 1}$ of responses $r_i = 1$ in this sample $\mathcal{S}$ will then be a random variable. Again, the question is whether or not $n_{\cdot 1}$ should be fixed, as we are now no longer observing the total population. To answer this question, Greenland formulates the *Sharp Causal null hypothesis about the Sample* (SCS),

$$H_{\text{SCS}} \colon \forall i \in \mathcal{S} \colon r_i(1) = r_i(2).$$

Once the sample $\mathcal{S}$ is drawn, H_{SCS} is well-defined and concerns only the $n_{\cdot\cdot}$ observed individuals. This is thus the same setting as the one from Table 3.3, and we should condition on $n_{\cdot 1}$. How is H_{SCS} related to H_{SCP}? Greenland first notes that $H_{\text{SCP}} \implies H_{\text{SCS}}$. Indeed, if the given treatment has no effect on the response of any of the individuals in the population $\mathcal{P}$, then certainly it has no effect on the response of any of the individuals in the sample $\mathcal{S}$. The reverse implication $H_{\text{SCS}} \implies H_{\text{SCP}}$ does not hold. However, Greenland argues that it would make no sense to reject H_{SCP} without rejecting H_{SCS}. This would mean that we conclude that the choice of treatment did affect the response of some population member, while not concluding that the treatment choice had an effect on some sample member response. This is the same as saying that although we have not found evidence of an effect of the treatment choice from the observations in our sample, we will still conclude that there is an effect for someone in the population. Such reasoning defeats the whole purpose of trying to infer something about a whole population by only observing a sample.

Thus, there is no coherent way in which we can reject H_{SCP}, but not reject H_{SCS}, based on the same data and the same model. Since we used a conditional test on the sample in order to test H_{SCS}, the only way to avoid incoherencies is to also use the conditional test on the sample to test H_{SCP}, and to conclude that performing the unconditional test on the sample $\mathcal{S}$ must test for some other hypothesis instead.

Under Rubin's model, it can thus be argued that one should condition on both margins. However, what happens when we are not dealing with this causal

model? Shuster [66], a big proponent of unconditional tests himself, questions for example the applicability of Greenland's reasoning because it is based on the use of sharp null hypotheses, which is debatable in certain experimental settings, even though Greenland also showed that his reasoning also works for other, weaker hypotheses. But besides that, Greenland argues that in several cases, there is no logical argument, such as the one just made, against the use of unconditional tests. He stresses that this is not a reason to use unconditional tests, but merely that there is no "logically compelling argument for conditional tests" [52].

One of these cases where unconditional tests might be appropriate is the descriptive comparison of two superpopulations. Consider two disjoint superpopulations, labelled $x = 1$ and $x = 2$, and let p_1 and p_2 be the respective proportions of individuals i in each population who have a certain characteristic $r_i \in \{1, 2\}$ of interest. That is, $p_x = P(r_i = 1 \mid i \in x)$. We take independent random samples from each superpopulation and want to test what Greenland refers to as the *Descriptive null hypothesis for 2 Populations* (D2P),

$$H_{\text{D2P}} \colon p_1 = p_2.$$

The sampling distribution in this case is the same product-binomial distribution given in (2.20) (but then with p_1, p_2 instead of θ_1, θ_2) which also appears in the causal example. Again, hypothetical repetitions of experiment would yield different values of $n_{\cdot 1}$ at each repetition. In both experiments, we draw a sample in a way that what we observe for each individual is not related to the actual attribute of interest (being $(r_i(1), r_i(2))$ in the causal experiment and r_i in the descriptive experiment).

However, it is imperative to realise the difference between the two experiments. In the causal model, the statistical uncertainty comes from the "missingness problem". It is because we cannot observe both $r_i(1)$ and $r_i(2)$ for an individual i that we had to randomise and to introduce H_{SCS}. In the descriptive experiment, the only statistical uncertainty comes from trying to say something about the entire populations based on only the samples we took, just as in survey sampling. If we would observe the entire population, there would be no uncertainty and we could simply reject / not reject H_{D2P} with certainty. The descriptive experiment does not have the same structure, and it is thus not possible to have a construction that, by using an unconditional test, could lead to incoherencies between a hypothesis concerning only the sample and a hypothesis concerning the entire population, like we had in the causal example. The author would like to express his gratitude towards Prof. Greenland. He made a number of insightful comments in private correspondence, which definitely helped in correctly conveying his ideas in this text.

Apart from the philosophical considerations described in the previous paragraphs and in fact the whole of this Chapter, one should also look at the practical perspective. Of course, conditional (and asymptotic) tests are way faster than unconditional tests. However, we feel that within the realm of contingency tables with small sample sizes, the current state of readily available computational power made it entirely feasible to perform unconditional tests in acceptable time. Apart from some unconditional tests which are already quite fast, such as the ones using external test statistics like the one introduced by Suissa and Shuster [15], we will propose in Chapter 4 a test within the classical Neyman–Pearson framework that

performs similarly to Barnard's CSM test, but in a fraction of the time.

The question whether to condition on both table margins or only on the group size margin seems to still lack a definitive answer. Some argue that this "reveals a fundamental weakness in frequentist inference" [67], or accept that the question cannot be answered at all and instead recommend to "act like a Bayesian" [68]. It is within this context of disagreement and uncertainty that we find ourselves today. As long as a clear answer fails to appear, both approaches deserve to be further researched and developed. It is the – perhaps slightly biased – view of the author that this thesis adds to this development. This in particular by reaching out an opportunity to perform unconditional tests in larger tables, for everyone who deems their use in a specific setting appropriate.

CHAPTER

FOUR

LARGER TABLES

Until now, we have been predominantly focusing on the 2×2 contingency table. We can however enlarge the number of rows and/or columns to obtain larger tables. The first generalisation would be to compare r binomial distributions instead of 2. This would yield a $r \times 2$ table, where the corresponding null hypothesis would be $H_0 : \theta_1 = \theta_2 = \cdots = \theta_r = \theta$. Alternatively, we could compare multinomial distributions instead of binomial distributions. A c-nomial distribution is characterised by a number of n independent trials, where each trial leads to the occurrence of exactly one out of c mutually exclusive events $A_1, A_2, \ldots, A_c$, each with corresponding probabilities $\theta_1, \theta_2, \ldots, \theta_c$ such that $\sum_{j=1}^{c} \theta_j = 1$. The outcome would then be a vector $(x_1, x_2, \ldots, x_c)$, with probability

$$\frac{n!}{x_1! \cdots x_c!} \theta_1^{x_1} \cdots \theta_c^{x_c}. \tag{4.1}$$

If we wished to compare two c-nomial distributions, we would end up with a $2 \times c$ contingency table, where we would test the multiple null hypothesis

$$H_0 : \theta_{11} = \theta_{21} = \theta_1,\ \theta_{12} = \theta_{22} = \theta_2, \ldots,\ \theta_{1,c-1} = \theta_{1,c-1} = \theta_{c-1},$$

where θ_{ij} is of course the probability of result j occurring in group i. Note that the null hypothesis immediately implies that $\theta_{1c} = \theta_{2c} = \theta_c$ too. Combining these two generalisations leads to the $r \times c$ contingency table, which is shown in Table 4.1.

	A_1	A_2	$\cdots$	A_c	
Group 1	x_{11}	x_{12}	$\cdots$	x_{1c}	$n_{1\cdot}$
Group 2	x_{21}	x_{22}	$\cdots$	x_{2c}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Group r	x_{r1}	x_{r2}	$\cdots$	x_{rc}	$n_{r\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot c}$	$n_{\cdot\cdot}$

Table 4.1: An $r \times c$ contingency table.

In the 2×2 case, the two-dimensional vector (x_{11}, x_{21}) fully determined the table if we also knew $(n_{1\cdot}, n_{2\cdot})$. From now on, we will regard Table 4.1 as a real-

isation $\mathbf{x}$ of the random matrix

$$\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1c} \\ X_{21} & X_{22} & \cdots & X_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ X_{r1} & X_{r2} & \cdots & X_{rc} \end{pmatrix}, \tag{4.2}$$

where $(X_{i1}, X_{i2}, \ldots, X_{ic})$ has a multinomial distribution with $n_{i\cdot}$ trials and respective event probabilities $\theta_{i1}, \theta_{i2}, \ldots, \theta_{ic}$ for $i \in \{1, \ldots, r\}$. For such a table, we are interested in testing the multiple null hypothesis

$$\begin{aligned} H_0 : \theta_{11} &= \theta_{21} = \ldots = \theta_{r1} = \theta_1, \\ \theta_{12} &= \theta_{22} = \ldots = \theta_{r2} = \theta_2, \\ &\ldots, \\ \theta_{1,c-1} &= \theta_{2,c-1} = \ldots = \theta_{r,c-1} = \theta_{c-1} \end{aligned} \tag{4.3}$$

against the alternative hypothesis that at least one of the above equalities does not hold. Notice that for an $r \times c$ table, the number of common probabilities $\theta_1, \theta_2, \ldots, \theta_{c-1}$, and hence the number of nuisance parameters, will be equal to $c - 1$. Alternatively, if we let $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{ic})$ for $i = 1, \ldots, r$ (keep in mind that $\sum_{j=1}^{c} \theta_{ij} = 1$), we can write the null hypothesis as $H_0 \colon (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_r) \in \Theta_0$, where

$$\Theta_0 = \{(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_r) : \boldsymbol{\theta}_1 = \ldots = \boldsymbol{\theta}_r = \boldsymbol{\theta}\},$$

where $\boldsymbol{\theta}$ is the vector of $c - 1$ nuisance parameters (where the last entry is then uniquely defined via $\sum_{j=1}^{c} \theta_j = 1$). The overall parameter space is thus the Cartesian product of r $c - 1$-dimensional simplexes

$$\left\{ (\theta_{i1}, \ldots, \theta_{ic}) : \sum_{j=1}^{c} \theta_{ij} = 1 \right\}.$$

We will adopt the convention that unless stated otherwise, $\boldsymbol{\theta}$ (without a subscript) will always represent the common value of $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_r$ under H_0, i.e., the vector of nuisance parameters. These nuisance parameters are just numbers in $[0, 1]$, and will be indexed by a subscript. No confusion is possible however because of the boldface for the vectors of probabilities $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_r$. Finally, it is worth noting that one can even create higher dimensional tables, i.e. $d_1 \times d_2 \times \cdots \times d_s$ tables. We will not deal with the latter in this text.

In the situations where we would consider the use of an unconditional test justified, as discussed in Chapter 3, the question we would like to answer is how unconditional exact tests would generalise to larger $r \times c$ tables. Asymptotic and conditional tests have already been explored to a great extent for larger tables (see for example the book of Fagerland, Lydersen and Laake [1]). However, probably due to the large number of computations that unconditional tests will inevitably require, there does not seem to be a lot of effort put into developing exact unconditional methods for $r \times c$ tables [1]. Therefore, let us first briefly mention how the asymptotic and conditional tests extend to larger tables, and then take a look at how we might extend the supremum method (as well as the C and S conditions of Barnard's CSM test) to $r \times c$ contingency tables.

4.1 Extending the asymptotic and conditional tests

The chi-square test statistic given in (2.2) can easily be generalised to an $r \times c$ table by:

$$\chi^2(\mathbf{X}) = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(X_{ij} - n_{i\cdot}\hat{\theta}_j)^2}{n_{i\cdot}\hat{\theta}_j}, \quad \text{where } \hat{\theta}_j = \frac{n_{\cdot j}}{n_{\cdot\cdot}}. \tag{4.4}$$

If this expression is undefined, we again set $\chi^2(\mathbf{x}) = -\infty$. One can show that under the null hypothesis, $\chi^2(\mathbf{X})$ has an asymptotic chi-square distribution with $(r-1)(c-1)$ degrees of freedom. We will omit the proof here, as well as the exact requirements on the asymptotic growth of the group sizes $(n_{i\cdot})_{i=1\ldots,r}$, but one can get an idea of the proof in an article by Benhamou and Melot [69]. An argument that explains the number of degrees of freedom is that, given both row and column sums, we need $(r-1)(c-1)$ table values in order to fully determine the table. This argument does not attempt in any way to be rigorous.

Be that as it may, trusting that this asymptotic distribution is correct, we can reject the null hypothesis for large values of $\chi^2(\mathbf{X})$, and set the corresponding p-value as $p(\mathbf{x}) = P_{H_0}(\chi^2(\mathbf{X}) \geq \chi^2(\mathbf{x}))$ just as in (2.7). Again, this asymptotic approximation may not be justified for small sample sizes. Similar to the rule of thumb mentioned by Fisher, Cochran [70] suggested to only use this test whenever at least 80% of the expected cell counts $n_{i\cdot}\hat{\theta}_j$ is larger than 5, and all expected cell counts are larger than 1. Remark that this reduces to Fisher's rule for the 2×2 table.

Luckily, exact approaches can also be used whenever the chi-square method is not deemed appropriate. In particular, there exists a generalisation of Fisher's exact test to $r \times c$ tables, called the Fisher–Freeman–Halton exact test. In 1951, Freeman and Halton [71] showed that, fixing the table margins, i.e., conditional on $\mathbf{X}$ being in

$$M_{n_{i\cdot}}^{n_{\cdot j}} = \left\{ \mathbf{x} \in \Omega : \sum_{j=1}^{c} x_{ij} = n_{i\cdot}, i = 1, \ldots, r \text{ and } \sum_{i=1}^{r} x_{ij} = n_{\cdot j}, j = 1, \ldots, c \right\},$$

where Ω is the set of all possible table outcomes, the probability of observing Table 4.1 under H_0 is given by

$$P_{H_0}\big(\mathbf{X} = \mathbf{x} \mid \mathbf{X} \in M_{n_{i\cdot}}^{n_{\cdot j}}\big) = \frac{\prod_{i=1}^{r} n_{i\cdot}! \prod_{j=1}^{c} n_{\cdot j}!}{n_{\cdot\cdot}! \prod_{i=1}^{r} \prod_{j=1}^{c} x_{ij}!}. \tag{4.5}$$

This is the probability mass function of the multiple hypergeometric distribution. We can compute a p-value by summing the probabilities of all tables which are less or equally likely to occur under this distribution than the observed table, i.e.

$$p_{\text{FFH}}(\mathbf{x}) = \sum_{\mathbf{y} \in S_{\mathbf{x}}} P_{H_0}\big(\mathbf{X} = \mathbf{y} \mid \mathbf{X} \in M_{n_{i\cdot}}^{n_{\cdot j}}\big), \tag{4.6}$$

where

$$S_{\mathbf{x}} = \big\{ \mathbf{y} \in M_{n_{i\cdot}}^{n_{\cdot j}} : P_{H_0}\big(\mathbf{X} = \mathbf{y} \mid \mathbf{X} \in M_{n_{i\cdot}}^{n_{\cdot j}}\big) \leq P_{H_0}\big(\mathbf{X} = \mathbf{x} \mid \mathbf{X} \in M_{n_{i\cdot}}^{n_{\cdot j}}\big) \big\}.$$

This specific choice allowed Mehta and Patel [55] to construct a network algorithm that would speed up this otherwise lengthy computation. They proposed to construct a directed acyclic graph in which the set of all unique paths running from

some initial vertex to some terminal vertex would be isomorphic to $M_{n_{i\cdot}}^{n_{\cdot j}}$. The path corresponding to $\mathbf{y} \in M_{n_{i\cdot}}^{n_{\cdot j}}$ would be assigned a length of

$$\frac{n_{\cdot\cdot}!}{\prod_{i=1}^{r} n_{i\cdot}!} P_{H_0}\big(\mathbf{X} = \mathbf{y} \mid \mathbf{X} \in M_{n_{i\cdot}}^{n_{\cdot j}}\big).$$

The problem of computing the p-value would then amount to summing the lengths of all paths which have a smaller length than the path corresponding to the observed $\mathbf{x}$. Mehta and Patel argue that by translating the computation of the p-value to a network problem, the need to enumerate over all possible tables is removed, as it is possible to preemptively remove infeasible paths from consideration. A nice and detailed explanation of the procedure can be found in the book by Berry, Johnston and Mielke [72].

4.2 Generalising Barnard's (CS)M test

Extending the supremum method of finding p-values to larger tables is – conceptually – not a very hard thing to do. Also for an $r \times c$ table, we can list all possible table outcomes with given a $(n_{1\cdot}, \ldots, n_{r\cdot})$-margin. This set of outcomes Ω becomes rather big quite quickly, however. This is straightforward combinatorial computation; we can view the i-th row of the table as a balls-and-boxes problem of dividing $n_{i\cdot}$ indistinguishable balls over c distinguishable boxes. It is well-known that we can do this in

$$\binom{n_{i\cdot} + c - 1}{n_{i\cdot}}$$

different ways. Consequently, the total number of possible table outcomes is the product

$$\omega := |\Omega| = \prod_{i=1}^{r} \binom{n_{i\cdot} + c - 1}{n_{i\cdot}}. \tag{4.7}$$

Given this set of tables, we can then for example compute a test statistic of interest for all these tables, such as (4.4), and compute a p-value via the analogue of (2.21) for multiple nuisance parameters:

$$p(\mathbf{x}) = \sup_{\boldsymbol{\theta} \in \Theta_0} P_{\boldsymbol{\theta}}(T(\mathbf{X}) \geq T(\mathbf{x})),$$

or with the inequality sign reversed. Alternatively, speaking in terms of ordering the tables, we might work out an ordering sequentially just as with Barnard's (CS)M test, based on (2.22), such that the $k+1$-th table in the ordering can be defined once $\{\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(k)}\}$ are known:

$$\mathbf{x}_{(k+1)} = \operatorname*{arg\,min}_{\mathbf{y} \in \Omega \setminus \{\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(k)}\}} \sup_{\boldsymbol{\theta} \in \Theta_0} \left\{ P(\mathbf{y}; \boldsymbol{\theta}) + \sum_{i=1}^{k} P\big(\mathbf{x}_{(i)}; \boldsymbol{\theta}\big) \right\}, \tag{4.8}$$

The vectors in that equation should still be understood as representing the tables, but they will of course no longer be 2-dimensional vectors. Furthermore, the probability $P(\mathbf{x}; \theta)$ of observing the 2×2 table $\mathbf{x} = (x_{11}, x_{21})$ under the null hypothesis

that $\theta_1 = \theta_2 = \theta$ should be replaced by the equivalent probability of observing Table 4.1 under the null hypothesis (4.3), that is,

$$P(\mathbf{x};\boldsymbol{\theta}) := P_{\boldsymbol{\theta}}(\mathbf{X}=\mathbf{x}) = \prod_{i=1}^{r}\prod_{j=1}^{c}\frac{n_{i\cdot}!}{x_{ij}!}\theta_j^{x_{ij}} = \frac{\prod_{i=1}^{r} n_{i\cdot}!}{\prod_{i=1}^{r}\prod_{j=1}^{c} x_{ij}!}\prod_{j=1}^{c}\theta_j^{n_{\cdot j}}. \tag{4.9}$$

Constructing the ordering of Ω entails maximising a function with respect to the nuisance parameters (i.e., over a c-dimensional simplex) a total of $|\Omega| - k + 1$ at the k-th iteration. Therefore, introducing some larger dimensional analogue to the C and S conditions might drastically reduce the number of maximisations we need to perform.

4.2.1 The S Condition

In the 2×2 case, we interpreted a given table and the table where the columns were swapped as equally strong evidence against H_0. Equivalently, in the $r \times c$ case, we say that two tables carry equal evidence against H_0 if they are the same up to a permutation of the columns. If moreover, we have $n_{i\cdot} = n_{j\cdot}$ for $i, j \in I \subset \{1, \ldots, r\}$, we can also permute the rows with index in I and view the resulting tables as symmetric to the original table.

However, finding out which tables are symmetric counterparts of each other by switching around and comparing columns and rows can become a relatively costly operation. Therefore, if we could somehow reduce the requirements on rows and columns to just comparing a few numbers, there might be a significant computational gain. Therefore, we would like to introduce a slightly different definition of the S condition. Essentially, Barnard defined his S condition for 2×2 contingency tables by requiring that symmetric tables should provide the same evidence against the null hypothesis. We can interpret this requirement for general $r \times c$ in terms of the $P(\cdot;\boldsymbol{\theta})$-function. Two tables could be seen as equally unlikely under the null hypothesis if their $P(\cdot;\boldsymbol{\theta})$-functions are equal, up to a permutation of $\boldsymbol{\theta}$. What do we mean by that? Consider an example with 3×3 contingency tables with sample sizes $(n_{1\cdot}, n_{2\cdot}, n_{3\cdot}) = (5, 5, 4)$. Then (4.9) becomes

$$P(\mathbf{x};\boldsymbol{\theta}) = \frac{\prod_{i=3}^{3} n_{i\cdot}!}{\prod_{i=1}^{3}\prod_{j=1}^{3} x_{ij}!}\theta_1^{n_{\cdot 1}}\theta_2^{n_{\cdot 2}}\theta_3^{n_{\cdot 3}} =: K_{\mathbf{x}}\theta_1^{n_{\cdot 1}}\theta_2^{n_{\cdot 2}}\theta_3^{n_{\cdot 3}}.$$

We can say that two tables are equally likely under H_0 if their $P(\cdot;\boldsymbol{\theta})$-functions have the same value for $K_{\mathbf{x}}$ and if the column margins $(n_{\cdot 1}, n_{\cdot 2}, n_{\cdot 3})$ are composed of the same values, but maybe in a different order. In general, since $(n_{i\cdot})_{i=1,\ldots,r}$ is fixed, two tables $\mathbf{x}$ and $\mathbf{y}$ can be seen as symmetric according to what we will call the $\mathrm{S_P}$ condition if

$$\prod_{i=1}^{r}\prod_{j=1}^{c} x_{ij}! = \prod_{i=1}^{r}\prod_{j=1}^{c} y_{ij}!, \quad \text{and } (n^{\mathbf{x}}_{\cdot j})_{j=1,\ldots,c} \sim (n^{\mathbf{y}}_{\cdot j})_{j=1,\ldots,c}, \tag{4.10}$$

where by $\mathbf{a} \sim \mathbf{b}$ we mean that $\mathbf{a}$ and $\mathbf{b}$ are permutations of each other. It is easy to see that if two tables are symmetric according to Barnard's S condition, then they are also symmetric according to $\mathrm{S_P}$. The reverse implication is a bit more complicated. In the case of 2×2 tables, under the condition that $n_{1\cdot} = n_{2\cdot}$, it

seems that the S_P symmetry requirement coincides with Barnard's S. We state this based on the conjecture that

$$
\begin{aligned}
(4.10) \iff & \begin{cases} x_{11}!x_{12}!x_{21}!x_{22}! = y_{11}!y_{12}!y_{21}!y_{22}!, \\ x_{11}+x_{12} = y_{11}+y_{12}, \\ x_{21}+x_{22} = y_{21}+y_{22}, \\ x_{11}+x_{21} = y_{11}+y_{21} \text{ or } x_{11}+x_{21} = y_{12}+y_{22}, \end{cases} \\
\overset{?}{\Longrightarrow} & \begin{cases} x_{11}=y_{11},\ x_{12}=y_{12},\ x_{21}=y_{21},\ x_{22}=y_{22}, \text{ or} \\ x_{11}=y_{12},\ x_{12}=y_{11},\ x_{21}=y_{22},\ x_{22}=y_{21}, \text{ or} \\ x_{11}=y_{21},\ x_{12}=y_{22},\ x_{21}=y_{11},\ x_{22}=y_{12}, \text{ or} \\ x_{11}=y_{22},\ x_{12}=y_{21},\ x_{21}=y_{12},\ x_{22}=y_{11}. \end{cases}
\end{aligned}
$$

which is the same as saying that $\mathbf{x}$ and $\mathbf{y}$ are symmetric according to S. We have checked that this conjecture is true at least until $n_{1\cdot} = n_{2\cdot} = 150$.

For $n_{1\cdot} \neq n_{2\cdot}$, this reverse implication does not hold. Indeed, taking $(n_{1\cdot}, n_{2\cdot}) = (2, 11)$ serves as a counterexample. The tables $(1, 8)$ and $(2, 7)$ both have column margins $(n_{\cdot 1}, n_{\cdot 2}) = (9, 4)$ and the product of the factorials of the table entries is in both cases equal to $1!1!8!3! = 0!2!4!7! = 241920$. These tables are thus symmetric according to S_P, but clearly not according to S.

Also for larger tables, this symmetry condition leads to larger equivalence classes than Barnard's symmetry condition. For example, in the 3×3 case with equal group sizes, the two tables given in Table 4.2 are symmetric according to S_P, but not according to Barnard's S condition. Indeed, by only switching rows and columns, we can never go from the left table to the right table or vice versa.

	A_1	A_2	A_3	
Group 1	2	1	1	4
Group 2	0	3	1	4
Group 3	0	0	4	4
	2	4	6	12

	A_1	A_2	A_3	
Group 1	1	3	0	4
Group 2	1	1	2	4
Group 3	0	0	4	4
	2	4	6	12

Table 4.2: Two tables that are symmetric according to S_P, but not according to S.

Because we only need to perform some elementary computations to see whether S_P is satisfied, instead of comparing matrices, we shorten the computation time. Also, as for larger tables the symmetry classes are actually larger in the case of S_P, we will go through the space of possible outcomes more quickly. More on this computational aspect in Section 5.3.

4.2.2 The C condition

The C condition is a bit more tricky to define. Let us first limit ourselves to the $r \times 2$ case, where we are comparing r binomials, and therefore still only have one nuisance parameter. Although we just mentioned that we will from now on only look at table outcomes as matrices like in (4.2), it is worth noticing that $r \times 2$ tables could still be represented as r-dimensional vectors, given the $(n_{1\cdot}, \ldots, n_{r\cdot})$-margin.

Consequently, the set of outcomes Ω can be visualised as an r-dimensional lattice, in the same spirit as Figure 2.2. Because of that, we can easily define a similar convexity condition. For a table $\mathbf{x} = (x_{11}, \ldots, x_{r1})$, we view all the tables which lie further away from the (hypercube) diagonal connecting the outcomes $(0, \ldots, 0)$ and $(n_{1\cdot}, \ldots, n_{r\cdot})$. For example, in the 3×2 setting with $n_{1\cdot} = n_{2\cdot} = n_{3\cdot} = 5$, we would expect the table $(1, 2, 1)$ to provide less evidence against H_0 than the tables $(0, 2, 1)$, $(1, 1, 1)$ and $(1, 2, 0)$. With that, the CSM algorithm works in exactly the same way as in the 2×2 case; we start ordering at the most extreme points, being $n_{i\cdot}\mathbf{e}_i$ $(i = 1, \ldots, r)$ and their symmetric counterparts. Here $\mathbf{e}_i$ is the i-th unit vector. Afterwards we start looking at all nearest neighbours which we have not ordered already, respecting the convexity requirement, and slowly fill up the r-dimensional lattice.

There is no longer a nice graphical interpretation the moment we are considering tables with more than 2 columns. However, we can still apply the same idea of looking at the "most extreme" tables and subsequently considering the nearest neighbours of these most extreme tables first. We will go through all tables in Ω sequentially by only considering the nearest neighbours of tables we have already ordered to be the next in our ordering. This is not exactly the same as the condition we employed in the 2×2 case. For example, the method we just described would allow the outcome $(3, 4)$ to be considered next in Figure 2.3, in contrast to our original convexity condition. However, we would expect this less strict requirement to work well enough. Recall that we can enforce any conditions we would reasonably want; as we are just looking for ways to reduce the number of outcomes to consider at each iteration.

The only question left to answer is how to determine which tables are the most extreme. In the $r \times 2$ case, we considered all tables for which there was exactly one group k such that $x_{k1} = n_{k\cdot}$, and $x_{i1} = 0$ for $i \neq k$ as the most extreme tables. They contained the most possible evidence against H_0, in particular that $\theta_k \neq \theta_i$ for $i \neq k$. As we are interested in the multiple null hypothesis (4.3) for the $r \times c$ setting, we consider each table which has a column of $r - 1$ zero entries, and exactly one nonzero entry equal to sample size for that group. Of course, their symmetric counterparts should be treated as equally extreme.

4.2.3 The use of external test statistics

Just as we discussed in Section 2.5, we can introduce external test statistics to do a couple of things. If while ordering the tables, we encounter a tie between two candidates, we can use an external test statistic (chi square or the mean value of $P(\cdot; \boldsymbol{\theta})$) to determine which candidate to pick. We could also use the test statistic to create the ordering itself, and finally the external test statistic can be used to split up the outcome space in symmetry classes. Almost all these applications naturally extend to larger tables too. For example, Boschloo's test, which uses the Fisher p-value test statistic in the 2×2 case, is also easily defined for larger tables. We just need to use the Fisher–Freeman–Halton p-value. Furthermore, we can also compute the mean value test statistic, i.e., the integral $\int_{\Theta_0} P(\mathbf{x}; \boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}$ of $P(\cdot; \boldsymbol{\theta})$ over Θ_0. This is done in Proposition 4.1.

Proposition 4.1. The mean value of $P(\mathbf{x};\theta)$ as defined in (4.9) is equal to

$$\int_{\Theta_0} P(\mathbf{x};\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = \frac{\prod_{i=1}^{r} n_{i\cdot}! \prod_{j=1}^{c} n_{\cdot j}!}{(n_{\cdot\cdot}+c-1)! \prod_{i=1}^{r}\prod_{j=1}^{c} x_{ij}!}, \tag{4.11}$$

where the integral over Θ_0 should be understood as a multiple integral over the $c-1$-dimensional simplex

$$\Theta_0 := \left\{ \boldsymbol{\theta} \in [0,1]^c : \sum_{j=1}^{c} \theta_j = 1 \right\}. \tag{4.12}$$

Proof of Proposition 4.1. We will show by induction on c that

$$\int_{\Theta_0} \prod_{j=1}^{c} \theta_j^{n_{\cdot j}} \mathrm{d}\boldsymbol{\theta} = \frac{\prod_{j=1}^{c} \Gamma(n_{\cdot j}+1)}{\Gamma\left(\sum_{j=1}^{c} n_{\cdot j} + c\right)}. \tag{4.13}$$

Since $\Gamma(m+1) = m!$ for $m \in \mathbb{N}$, (4.13) immediately leads to (4.11) by plugging (4.9) into $\int_{\Theta_0} P(\mathbf{x};\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$.

To this end, let us first suppose that $c = 2$. Then $\Theta_0 \subset [0,1]^2$ is the line segment $\{(\theta_1, 1-\theta_1) : 0 \le \theta_1 \le 1\}$ and so

$$\begin{aligned}
\int_{\Theta_0} \prod_{j=1}^{c} \theta_j^{n_{\cdot j}} \mathrm{d}\boldsymbol{\theta} &= \int_0^1 \theta_1^{n_{\cdot 1}} (1-\theta_1)^{n_{\cdot 2}} \mathrm{d}\theta_1 \\
&= B(n_{\cdot 1}+1, n_{\cdot 2}+1) \\
&= \frac{\Gamma(n_{\cdot 1}+1)\Gamma(n_{\cdot 2}+1)}{\Gamma(n_{\cdot 1}+n_{\cdot 2}+2)},
\end{aligned}$$

which is what we wanted to show. Now suppose that (4.11) holds for $c = k \ge 2$. Then realise that for $c = k+1$, we can write

$$\int_{\Theta_0} \prod_{j=1}^{k+1} \theta_j^{n_{\cdot j}} \mathrm{d}\boldsymbol{\theta} = \int_0^1 \theta_{k+1}^{n_{\cdot,k+1}} \int_{\Theta_0'} \prod_{j=1}^{k} \theta_j^{n_{\cdot j}} \mathrm{d}\boldsymbol{\theta}' \mathrm{d}\theta_{k+1}, \tag{4.14}$$

where $\boldsymbol{\theta}' = (\theta_1, \ldots, \theta_k)$ and $\Theta_0' = \left\{ \boldsymbol{\theta}' \in [0,1]^k : \sum_{j=1}^{k} \theta_j = 1 - \theta_{k+1} \right\}$, which is a $(k-1)$-dimensional simplex given θ_{k+1}. To evaluate this integral, we perform the change of variables

$$\nu_j = \frac{\theta_j}{1-\theta_{k+1}}, \quad \text{for } j = 1, \ldots, k \text{ and } \mathrm{d}\boldsymbol{\nu} = (1-\theta_{k+1})^{-(k-1)} \mathrm{d}\boldsymbol{\theta},$$

where the $(k-1)$ in the exponent follows from the integration region being $(k-1)$-

dimensional. If we let $N_0 := \left\{\boldsymbol{\nu} \in [0,1]^k : \sum_{j=1}^k \nu_j = 1\right\}$, we can rewrite (4.14) as

$$\begin{aligned}
(4.14) &= \int_0^1 \theta_{k+1}^{n_{\cdot,k+1}} \int_{N_0} \left(\prod_{j=1}^k (1-\theta_{k+1}^{n_{\cdot j}})\nu_j^{n_{\cdot j}}\right)(1-\theta_{k+1})^{k-1} \mathrm{d}\boldsymbol{\nu}\mathrm{d}\theta_{k+1} \\
&= \int_0^1 \theta_{k+1}^{n_{\cdot,k+1}}(1-\theta_{k+1})^{\sum_{j=1}^k n_{\cdot j}+k-1}\left(\int_{N_0}\prod_{j=1}^k \nu_j^{n_{\cdot j}}\mathrm{d}\boldsymbol{\nu}\right)\mathrm{d}\theta_{k+1} \\
&= \int_0^1 \theta_{k+1}^{n_{\cdot,k+1}}(1-\theta_{k+1})^{\sum_{j=1}^k n_{\cdot j}+k-1}\frac{\prod_{j=1}^k \Gamma(n_{\cdot j}+1)}{\Gamma\left(\sum_{j=1}^k n_{\cdot j}+k\right)}\mathrm{d}\theta_{k+1} \qquad \text{(induction)} \\
&= \frac{\prod_{j=1}^k \Gamma(n_{\cdot j}+1)}{\Gamma\left(\sum_{j=1}^k n_{\cdot j}+k\right)} B\left(n_{\cdot,k+1}+1, \sum_{j=1}^k n_{\cdot j}+k\right) \\
&= \frac{\prod_{j=1}^k \Gamma(n_{\cdot j}+1)}{\Gamma\left(\sum_{j=1}^k n_{\cdot j}+k\right)} \frac{\Gamma(n_{\cdot,k+1}+1)\Gamma\left(\sum_{j=1}^k n_{\cdot j}+k\right)}{\Gamma\left(\sum_{j=1}^{k+1} n_{\cdot j}+k+1\right)} \\
&= \frac{\prod_{j=1}^{k+1} \Gamma(n_{\cdot j}+1)}{\Gamma\left(\sum_{j=1}^{k+1} n_{\cdot j}+k+1\right)},
\end{aligned}$$

which is exactly (4.13) with $c = k+1$. □

We should only be careful when using the chi-square test statistic for larger tables. Of course, we would work with (4.4) instead of (2.2). However, for tables with more than two columns, the chi-square test statistic shows some undesirable behaviour. For the sake of illustration, consider the two 2×3 tables in Table 4.3. According to the last paragraph of Section 4.2.2, we would consider the left table as a "most extreme" table. Intuitively, this also seems to be the strongest possible evidence against the null hypothesis. The right table on the other hand, seems to be the weakest evidence against the null hypothesis we could possibly imagine. However, because both tables have two zero entries in column A_3, they both have a chi-square test statistic value of $-\infty$. The extreme table on the left would be perceived as just as extreme as the table on the left. Of course, we could still use the chi-square test statistic to determine the ordering of tables. The only downside would be that what we would see as the most extreme tables would instead end up last in the ordering (and thus a p-value of 1). The test would still produce a valid p-value, but will be less powerful. All the other tables will still get a "reasonable" p-value.

	A_1	A_2	A_3	
1	5	0	0	5
2	0	3	0	3
	5	3	0	8

	A_1	A_2	A_3	
1	5	0	0	5
2	3	0	0	3
	8	0	0	8

Table 4.3: Two 2×3 tables with the same chi-square test statistic value.

The implications are a bit more severe if we wish to use the chi-square test statistic in a CSM-like test to determine the symmetry classes. Of course, both

tables in Table 4.3 will be put in the same symmetry class. However, if we would still decide to start our ordering with the most extreme tables, we would need to put this odd symmetry class first in our ordering. The associated p-value will be 1. Correspondingly, all the symmetry classes that come afterwards will have a p-value of 1 too. Two possible solutions for this issue would be to define some other reasonable (less extreme) starting point. This will yield a sensible ordering (and set of p-values) and put the odd symmetry class containing the tables in Table 4.3 at the end of the ordering. Alternatively, one might opt to artificially define different values for the chi-square test statistic depending on a case-by-case basis for these few special tables.

However, because of the large number of tests we will already be considering in Chapter 5, and because the chi-square approach will turn out to in general lead to a less powerful test in the case of 2 columns, we decided to just not use the chi-square test statistic to define symmetry classes in a CSM-like procedure whenever we have more than 2 columns.

4.2.4 A small recap on the different symmetry conditions

Just before we move on constructing some unconditional tests, let us briefly recap the different symmetry conditions we have encountered. Their – rather brief – introductions have been scattered over this Chapter, which might not benefit the clarity of this text.

In the 2×2 case, Barnard introduced a symmetry condition based purely on the swapping of columns (and possibly rows with equal margins). Although this condition generalised to larger tables, we introduced due to computational reasons a slightly different condition, $\mathrm{S_P}$, defined in (4.10). We will use this condition instead of Barnard's S condition, also for 2×2 tables (for which we suspect there is no actual difference). Apart from these two, we can also define symmetry conditions based on the (external) test statistics. Tables which have the same chi-square test statistic value (4.4), are said to be symmetric according to the S_χ condition. If tables have the same mean value of $P(\cdot; \boldsymbol{\theta})$, i.e., the same value of (4.11), they are said to be symmetric according to the $\mathrm{S_V}$ condition.

Are there any relations between this symmetry conditions? That is, does one imply the other in certain cases? We have already shown in Section 2.5.1 that in the 2×2 case, $\mathrm{S} \Longrightarrow \mathrm{S}_\chi$ but not $\mathrm{S}_\chi \Longrightarrow \mathrm{S}$. By this abuse of notation we mean that "if two tables are symmetric according to S, they are also symmetric according to S_χ". Similarly, as mentioned in Section 2.5.2, $\mathrm{S} \not\Longrightarrow \mathrm{S_V}$ but not $\mathrm{S_V} \Longrightarrow \mathrm{S}$ (although we suspect this to be true whenever $n_{1\cdot} = n_{2\cdot}$), and also $\mathrm{S_V} \not\Longrightarrow \mathrm{S}_\chi$ and $\mathrm{S}_\chi \not\Longrightarrow \mathrm{S_V}$.

For larger tables, we already noted with Table 4.2 that $\mathrm{S_P} \not\Longrightarrow \mathrm{S}$. It is also very easy to find tables that show that in general, $\mathrm{S_V} \not\Longrightarrow \mathrm{S_P}$, $\mathrm{S}_\chi \not\Longrightarrow \mathrm{S_P}$, $\mathrm{S_V} \not\Longrightarrow \mathrm{S}_\chi$ and $\mathrm{S}_\chi \not\Longrightarrow \mathrm{S_V}$. We already have some in the 3×2 case with group sizes $(n_{1\cdot}, n_{2\cdot}, n_{3\cdot}) = (3, 2, 2)$. For example, the tables $(x_{11}, x_{21}, x_{31}) = (3, 0, 0)$ and $(0, 2, 0)$ are symmetric according to S_χ, but not according to $\mathrm{S_V}$ or $\mathrm{S_P}$. Similarly, the tables $(3, 1, 2)$ and $(2, 1, 2)$ are symmetric according to $\mathrm{S_V}$, but not according to S_χ or $\mathrm{S_P}$.

Furthermore, we do clearly see that if $\mathbf{x}$ and $\mathbf{y}$ are symmetric according to $\mathrm{S_P}$, i.e., satisfying (4.10), that $\mathbf{x}$ and $\mathbf{y}$ should also have the same value for (4.11). In other words, $\mathrm{S_P} \Longrightarrow \mathrm{S_V}$. A similar relation does not hold between $\mathrm{S_P}$ and S_χ; the two tables in Table 4.2 serve as a counterexample since they do not have the same

chi-square test statistic value. However, it is true that $\mathrm{S} \Longrightarrow \mathrm{S}_\chi$. Indeed, because (4.4) is a sum over all table cells, swapping columns (or rows with equal group sizes) does not change the individual terms but merely the order in which they appear in the sum, not changing the chi square test statistic value.

We can summarise the above observations as in Figure 4.1.

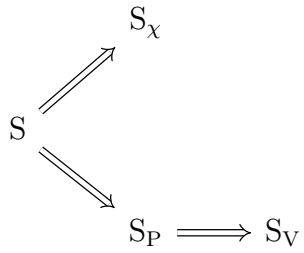


Figure 4.1: Chain of implications of the different symmetry conditions.

To conclude, recall the trade-off between speed and power we mentioned in Section 2.5.1. Given the chain of implications above, the conditions S_V and S_χ will in general divide the space of outcomes into fewer, larger symmetry classes than S_P. Except in the two 2×2 case, where we suspect that $\mathrm{S}_\mathrm{P} \Longrightarrow \mathrm{S}_\mathrm{V}$, we therefore expect that the tests using S_P will be slightly more powerful than those using S_V or S_χ.

4.3 Exact unconditional tests for $r \times c$ tables

Now that we have discussed how the building blocks of Barnard's CSM test would generalise to larger tables, we feel that we are able to construct an exact unconditional test for arbitrary $r \times c$ tables. We will propose two different approaches. The first approach could be seen as the direct generalisation of Barnard's CSM test. Here, we build up the ordering of tables sequentially, respecting the C and S conditions. We moreover introduce a number of adjustments in order to speed up an otherwise computationally intensive procedure. The second approach tries to perform an unconditional test more from a classical Neyman–Pearson perspective. By doing so, it turns out we can translate the problem of constructing a critical region for a given significance level α into a linear programming problem, providing a far quicker alternative to the CSM computation. Finally, we briefly discuss an idea that aims to reduce the $r \times c$ table to $c - 1$ $r \times 2$ tables. This will remove the need to perform maximisations over a $c - 1$-dimensional simplex of nuisance parameters, by instead solving $c - 1$ maximisation problems similar to that one encountered in the 2×2 setting.

4.3.1 Approach 1: Maximisation over the full simplex

As we have discussed in Section 4.2, Barnard's CSM test – and every other supremum test – essentially works in the same way for larger tables as for its 2×2 counterpart. There are however two complications when increasing the table dimensions. The first is that we need to consider a larger set of possible candidate outcomes at each step of the ordering. We have seen in the previous section that

defining equivalents for the C and S conditions severely decreases the number of table outcomes we should consider at each iteration.

The second one is that we should maximise the function

$$P(\mathbf{y};\boldsymbol{\theta}) + \sum_{i=1}^{k} P\big(\mathbf{x}_{(i)};\boldsymbol{\theta}\big) \tag{4.15}$$

no longer just over the unit interval $[0,1]$, but over the $c-1$-dimensional simplex Θ_0 as defined in (4.12). To perform this maximisation, we could either use a non-linear constrained optimisation method, or discretise Θ_0 in some way and simply evaluate the objective function in each point of our discretisation.

The advantage of using a discretised grid $\{\boldsymbol{\theta}^1,\ldots,\boldsymbol{\theta}^N\}$ of Θ_0 is that we can keep track of the summation $\sum_{i\leq k} P(\mathbf{x}_{(i)};\boldsymbol{\theta}^j)$ of all $P(\cdot;\boldsymbol{\theta}^j)$-functions over the already ordered points for all $j=1,\ldots,N$. This makes the maximisation in (2.22) a lot easier, as we can just store this "base layer" instead of evaluating it again and again, as would be the case when using a constrained optimisation method. Therefore, we opt to proceed with the discretisation method. It is important to note here however that the p-values we obtain using this method, are approximations and not the exact p-values. It is of course very unlikely that the maximum of (4.15) will be reached exactly in one the grid points. However, by choosing the grid "fine enough", we can assume that this approximate p-value will lie not too far from the actual p-value.

The question we are now left with is how to discretise the $c-1$-dimensional simplex Θ_0. One possibility is to first construct an (equidistant) grid on the hypercube $[0,1]^c$, and then transform it in order for it to lie in the simplex. If we map every grid point

$$(\theta_1,\ldots,\theta_c) \mapsto \left(\frac{\theta_1}{\sum_{j=1}^{c}\theta_j},\ldots,\frac{\theta_c}{\sum_{j=1}^{c}\theta_j}\right),$$

then we make sure that the entries of the image point sum up to 1. Although straightforward, this approach has two disadvantages. First of all, the number of grid points we have to consider grows exponentially with the number of columns c. Also, the transformed grid will no longer be equidistant, such that we might have a higher concentration of grid points in specific parts of the simplex. Consequently, this transformed grid could be too coarse to catch some maxima.

Because of these drawbacks, we propose to use a technique inspired from the Monte Carlo evaluation of integrals; so-called Quasi-Monte Carlo sequences, or low-discrepancy sequences. These are (multidimensional) sequences of points constructed in such a way that any of their subsequences will fill up $[0,1]^c$ "as best as possible". We elaborate on what this means in Appendix A. Using a low-discrepancy sequence removes the curse of dimensionality an equidistant grid has. Indeed, we are now able to have a subsequence with a fixed size, independent of the number of dimensions, that will evenly fill up $[0,1]^c$. We can act as if this entirely deterministic sequence is a sample from a uniform distribution on $[0,1]^c$, which we can then transform to something which looks like a uniform sample on the simplex. To this end, we state and prove the following result.

Proposition 4.2. If we let $\mathbf{U} = (U_1, \ldots, U_c)$ be a random vector uniformly distributed on $[0,1]^c$, then

$$\mathbf{W} := \left(\frac{-\log U_1}{-\sum_{k=1}^{c} \log U_j}, \ldots, \frac{-\log U_c}{-\sum_{k=1}^{c} \log U_j} \right) \tag{4.16}$$

is uniformly distributed on the simplex Θ_0 defined in (4.12). That is, $\mathbf{W}$ has a Dirichlet distribution with parameters $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_c) = (1, \ldots, 1)$ with density function:

$$f_{\mathbf{W}}(w_1, \ldots, w_c; \alpha_1, \ldots, \alpha_c) = \frac{\Gamma\left(\sum_{j=1}^{c} \alpha_j\right)}{\prod_{j=1}^{c} \Gamma(\alpha_j)} \prod_{j=1}^{c} w_j^{\alpha_j - 1} = \Gamma(c), \tag{4.17}$$

for $\mathbf{w} = (w_1, \ldots, w_c) \in \Theta_0$.

On a side note, observe that the normalising constant in (4.17) follows immediately from (4.13). The Dirichlet distribution can be seen as multivariate extension of the Beta distribution. Therefore, the normalising constant is often referred to as the multivariate Beta function:

$$B(\boldsymbol{\alpha}) := \frac{\prod_{j=1}^{c} \Gamma(\alpha_j)}{\Gamma\left(\sum_{j=1}^{c} \alpha_j\right)}. \tag{4.18}$$

Proof of Proposition (4.2). (Based on the proof of Theorem XI.4.1 in Devroye [73]) First of all, note that if $\mathbf{U}$ is uniformly distributed on $[0,1]^c$, then the U_j are independent and identically distributed uniform random variables on $[0,1]$ for $j = 1, \ldots, c$. Now define $V_j := -\log U_j$. Then V_j is exponentially distributed with parameter 1 for $j = 1, \ldots, c$. Indeed, by the inverse probability transform, we know that for V having a standard exponential distribution, $F_V(v) = 1 - e^{-v}$, so $F_V^{-1}(u) = -\log(1-u)$ and thus

$$-\log(U) \simeq -\log(1-U) = F_V^{-1}(U),$$

has a standard exponential distribution whenever U is uniformly distributed on $[0,1]$. The joint density function of $(V_1, \ldots, V_c)$ is therefore

$$f_{\mathbf{V}}(v_1, \ldots, v_c) = \prod_{j=1}^{c} e^{-v_j} = e^{-\sum_{j=1}^{c} v_j}.$$

If we now define $g : [0,\infty)^c \to [0,1]^{c-1} \times [0,\infty)$ as

$$g(v_1, \ldots, v_c) := \left(\frac{v_1}{s}, \ldots, \frac{v_{c-1}}{s}, s \right) = \mathbf{w}',$$

where $s = \sum_{k=1}^{c} v_k$, then $g^{-1}(w_1', \ldots, w_c') = \left(w_1' w_c', \ldots, w_{c-1}' w_c', w_c'(1 - \sum_{k=1}^{c-1} w_k') \right)$ and the density of $\mathbf{W}' = g(\mathbf{V})$ is given by

$$\begin{aligned} f_{\mathbf{W}'}(\mathbf{w}') &= f_{\mathbf{V}}\left(g^{-1}(\mathbf{w}')\right) \left| \frac{\partial g^{-1}(\mathbf{w}')}{\partial \mathbf{w}'} \right| \\ &= e^{-\sum_{j=1}^{c-1} w_j' w_c' - w_c'\left(1 - \sum_{k=1}^{c-1} w_k'\right)} (w_c')^{c-1} \\ &= e^{-w_c'} (w_c')^{c-1}, \end{aligned}$$

since

$$\left|\frac{\partial g^{-1}(\mathbf{w}')}{\partial \mathbf{w}'}\right| = \begin{vmatrix} w'_c & 0 & \cdots & w'_1 \\ 0 & w'_c & \cdots & w'_2 \\ \vdots & \vdots & \ddots & \vdots \\ -w'_c & -w'_c & \cdots & 1-\sum_{k=1}^{c-1} w'_k \end{vmatrix} = \begin{vmatrix} w'_c & 0 & \cdots & w'_1 \\ 0 & w'_c & \cdots & w'_2 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{vmatrix} = (w'_c)^{c-1}.$$

But then, integrating out the w'_c to obtain the joint density of $\left(\frac{V_1}{S}, \ldots, \frac{V_{c-1}}{S}\right)$:

$$f_{\left(\frac{V_1}{S},\ldots,\frac{V_{c-1}}{S}\right)}(w'_1, \ldots, w'_{c-1}) = \int_0^\infty e^{-w'_c}(w'_c)^{c-1}\mathrm{d}w'_c = \Gamma(c).$$

Since V_c is uniquely defined by $V_1, \ldots, V_{c-1}$ and S, $\mathbf{W} \simeq \left(\frac{V_1}{S}, \ldots, \frac{V_{c-1}}{S}\right)$ and we are done. □

On top of this Quasi-Monte Carlo grid, we will add some other points which do not belong to the low-discrepancy sequence. By construction, this sequence will never contain a boundary point of the simplex. However, it is desirable to include some of these boundary points. For many tables, the maximum of their corresponding $P(\cdot;\boldsymbol{\theta})$-functions will be reached at a $\boldsymbol{\theta}$-value which has one or more zero entries. For example, the sum of $P(\cdot;\boldsymbol{\theta})$-functions of the left 2×3 table given in Table 4.4 and its 5 symmetric counterparts is given by

$$2\big(\theta_1^3\theta_2^3 + \theta_1^3\theta_3^3 + \theta_2^3\theta_3^3\big),$$

which reaches a maximum of $1/32$ for $\boldsymbol{\theta} \in \{(1/2, 1/2, 0), (1/2, 0, 1/2), (0, 1/2, 1/2)\}$. We could therefore opt to construct another QMC grid on the edges of Θ_0. However, in practice, it seemed that using only the midpoints of each of the edges, such as $\{(1/2, 1/2, 0), (1/2, 0, 1/2), (0, 1/2, 1/2)\}$ for tables with three columns, already gave satisfactory results. Another example is given by the right table in Table 4.4. The sum of the $P(\cdot;\boldsymbol{\theta})$-functions of this table and its two symmetric counterparts is

$$\theta_1^6 + \theta_2^6 + \theta_3^6,$$

which reaches a maximum of 1 for $\boldsymbol{\theta} \in \{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}$. Of course, these are the three only tables for which $P(\cdot;\boldsymbol{\theta})$ evaluated at one of the vertices of the simplex Θ_0 will be something else than 0. However, it is important to include these $\boldsymbol{\theta}$-values into our grid too. As soon as these three tables become a candidate to be the next tables in the ordering, including $\boldsymbol{\theta} \in \{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}$ will immediately render these tables ineligible, for most candidates will have a maximum $P(\cdot;\boldsymbol{\theta})$-value smaller than 1.

	A_1	A_2	A_3	
1	3	0	0	3
2	0	3	0	3
	3	3	0	6

	A_1	A_2	A_3	
1	3	0	0	3
2	3	0	0	3
	6	0	0	6

Table 4.4: 2×3 tables with $\arg\max_{\boldsymbol{\theta}\in\Theta_0} P(\cdot;\boldsymbol{\theta})$ not in the interior of Θ_0.

Finally, we will also include $\boldsymbol{\theta}_m = (1/c, \ldots, 1/c)$, the midpoint of the simplex Θ_0 in the grid. This is because the sum of $P(\cdot;\boldsymbol{\theta})$-functions of some symmetry

classes will certainly reach a maximum at this point, exactly due to the symmetry within that class. For example, when applying the S_P condition, within a symmetry class, the sum of $P(\cdot;\boldsymbol{\theta})$-functions will return the same value for all permutations of a given $\boldsymbol{\theta}$. Clearly, $\|\boldsymbol{\theta}-\boldsymbol{\theta}_m\| = \|\pi(\boldsymbol{\theta})-\boldsymbol{\theta}_m\|$, where $\pi(\boldsymbol{\theta})$ is a vector obtained by permuting the entries of $\boldsymbol{\theta}$ according to a certain permutation. Therefore, it might be reasonable to expect that sometimes the maximum of the sum of $P(\cdot;\boldsymbol{\theta})$-functions is reached in $\boldsymbol{\theta}_m$.

4.3.2 Approach 2: A Packing Problem

Instead of working through the entire space of possible table outcomes using some kind of supremum test, we could also try to just construct a critical region K^α for a given significance level α. Surely, from this alone, we will not be able to derive p-values for individual tables, but if one is only interested in whether or not to reject the null hypothesis at a nominal significance level, having a critical region is enough. Setting aside all objections to this Neyman–Pearson way of testing, let us describe a method in which we can turn the problem of constructing the critical region K^α for a given α into a linear programming problem.

We would like to construct a critical region of table outcomes such that the probability of making a Type I error is less than α. Recall that under the null hypothesis, the probability of observing Table 4.1 was nothing else than $P(\mathbf{x};\boldsymbol{\theta})$ defined in (4.9). Our aim is thus to construct a critical region K^α such that

$$\sup_{\boldsymbol{\theta}\in\Theta_0} \sum_{\mathbf{x}\in K^\alpha} P(\mathbf{x};\boldsymbol{\theta}) \leq \alpha. \tag{4.19}$$

To minimise the Type II error, we also would like to make K^α as large as possible. That is, we want to make $\sum_{\mathbf{x}\in K^\alpha} P(\mathbf{x};\boldsymbol{\theta})$ as large as possible for $\boldsymbol{\theta}\in\Theta_1$. In order to convert this into a linear programming problem, we will slightly alter this aim in the hope it will still give reasonable power. This slightly altered objective is to include as many different tables into K^α as possible. The idea is that, as long as (4.19) is satisfied, creating the largest K^α possible will make $\sum_{\mathbf{x}\in K^\alpha} P(\mathbf{x};\boldsymbol{\theta})$ as close as possible to α for all $\boldsymbol{\theta}\in\Theta_0$, and potentially also very close to 1 for $\boldsymbol{\theta}\in\Theta_1$. We can already translate this into the following problem:

$$\begin{aligned} &\text{maximise } \mathbf{w}^T\mathbf{1} \\ &\text{subject to } \sum_{j=1}^{\omega} w_j P(\mathbf{x}^j;\boldsymbol{\theta}) \leq \alpha, \quad \text{for } \boldsymbol{\theta}\in\Theta_0, \\ &\qquad \mathbf{w}\in\{0,1\}^\omega, \end{aligned}$$

where we indexed all tables in Ω from 1 to ω and $\mathbf{w} = (w_1,\ldots,w_O)^T$ is a binary vector such that $w_j = 1$ whenever $\mathbf{x}^j \in K^\alpha$. Instead of checking that the first condition holds for all $\boldsymbol{\theta}\in\Theta_0$, we will again discretise Θ_0 and only consider a Quasi-Monte Carlo grid of N values for $\boldsymbol{\theta}$, labelled $\boldsymbol{\theta}^1,\ldots,\boldsymbol{\theta}^N$. This turns our problem into a so-called binary integer linear programming:

$$\begin{aligned} &\text{maximise } \mathbf{w}^T\mathbf{1} \\ &\text{subject to } A\mathbf{w} \leq \alpha\mathbf{1}, \\ &\qquad \mathbf{w}\in\{0,1\}^\omega, \end{aligned} \tag{4.20}$$

where $A = (a_{ij})$ is a $N \times \omega$ matrix with entries $a_{ij} = P(\mathbf{x}^j; \boldsymbol{\theta}^i)$ for $i = 1, \ldots, N$ and $j = 1, \ldots, \omega$. Although this linear program only approximates our actual problem, it allows us to make use of the vast collection of already-existing LP-solvers. This will turn out to significantly decrease computation time, as well as allow us to handle far bigger contingency tables. In this text, we solved all LP problems using the Gurobi Optimizer software [1]. The Gurobi software is proprietary, so its precise inner workings are unknown to us. However, as indicated on Gurobi's website [2], the common approach to solve binary integer linear programming programs (which are part of a larger class of problems called mixed integer linear programming problems) is to use a branch-and-bound algorithm. Mixed integer programming (MIP) problems are characterised by requiring that some of the variables should be integers.

The basic idea of the branch-and-bound approach is to "relax" the mixed integer programming to a linear programming by removing all the integrality constraints. This will yield (most probably) an optimal solution that does not satisfy the integrality constraints. We will now pick one of the variables v which should be integer, but takes the value $a \notin \mathbb{Z}$ instead, say. We can now solve two new LP problems: the one with the added restriction that $v \leq \lfloor a \rfloor$, and the one with the added restriction that $v \geq \lceil a \rceil$. With the obtained solutions, we can then again pick another variable which should be integer, and repeat the whole procedure. One can see that this will result in a tree of LP problems which get more and more restricted the further down the branches of the tree we go. Once we have found a solution that satisfies all the integrality conditions, we have found a feasible solution of the original MIP problem. We can now "prune" this branch of the tree and record the objective value this solution yields. If the original MIP is a maximisation problem, this objective value is a lower bound on the optimal objective value. Were we to find another feasible solution with a smaller objective value, we can also prune that branch as we already have a better solution. If we found another feasible solution with a larger objective value, we will update our lower bound. At the same time, note that the objective values of the solutions of all the relaxed LP sub-problems we considered so far serve as upper bounds for the optimal objective value. We can never obtain a better solution for the – more restricted – integer problem. We have found an optimal solution once the upper and lower bounds coincide.

Let us get back to statistics. One can wonder why the objective function (4.20) would yield sensible critical regions. An important realisation to make is that because we are maximising the number of tables in the critical region, we are effectively prioritising the inclusion of more extreme tables. To see this, it is instructive to interpret the problem of constructing the critical region as some kind of packing problem. This is visualised in Figure 4.2 for the simple 2×2 case with $(n_{1\cdot}, n_{2\cdot}) = (2, 2)$, but the same idea holds for general $r \times c$ tables too. Given the very high significance level $\alpha = 0.45$, we want to "fill up" the area underneath the level α with a selection of $P(\cdot; \boldsymbol{\theta})$-functions on the left of Figure 4.2. The objective of the LP formulation is to pack as many of these $P(\cdot; \boldsymbol{\theta})$-functions underneath the level α. A number of possible solutions are given to the right of Figure 4.2. The

[1] A free academic license of this solver is available at `https://www.gurobi.com/features/academic-named-user-license/`.

[2] `https://www.gurobi.com/resources/mixed-integer-programming-mip-a-primer-on-the-basics/`

optimal solution turns out to be the one in the top-right corner, i.e., the critical region $\{(1,0),(2,0),(1,2),(0,2)\}$, with an optimal objective value of 4. Clearly, this is because it uses the smaller $P(\cdot;\boldsymbol{\theta})$-functions, which do not take up that much space, and thus yield a larger objective value. This is true in general: the LP formulation (4.20) will naturally be inclined to include tables in the critical region which have small $P(\cdot;\boldsymbol{\theta})$-functions, i.e. $P(\cdot;\boldsymbol{\theta})$-functions taking "small" values for many values of $\boldsymbol{\theta}$. In other words, we will naturally be inclined to include the most extreme tables (in the case of the example, the tables $(2,0)$ and $(0,2)$). This may serve as an indication that the critical region obtained via (4.20) should have some resemblance to the one obtained via, say, a CSM-like test. Other, weirdly shaped critical regions will result a less effective packing of the area.

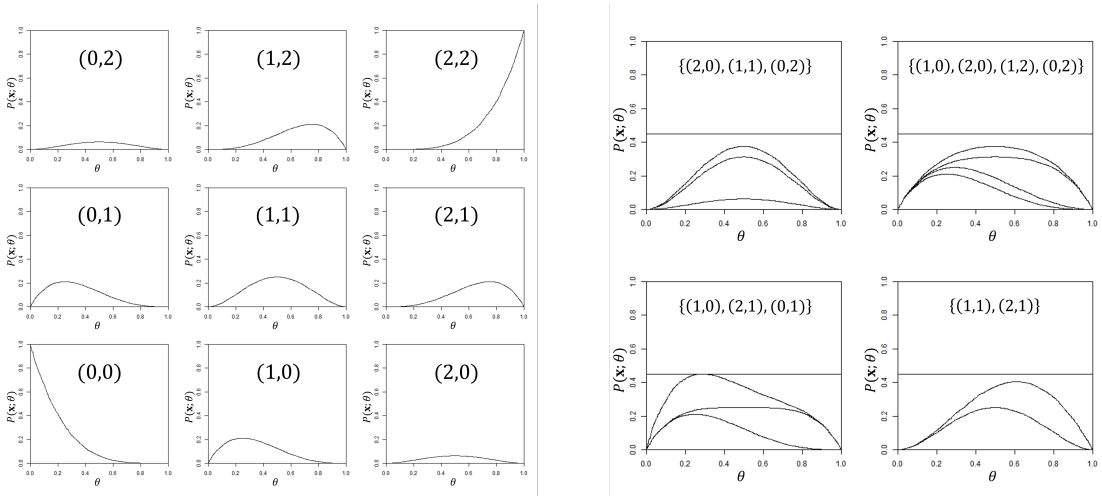


Figure 4.2: Pool of available $P(\cdot;\boldsymbol{\theta})$-functions for $(n_{1\cdot}, n_{2\cdot}) = (2,2)$ (left) and a number of solutions to the packing problem (right).

Note that we can also choose to incorporate some type of symmetry condition, just as with the supremum methods. Before solving (4.20), we might group tables together which are symmetric counterparts of each other, and then decide group by group whether or not we want to include it in the critical region. This has obviously a large effect on the size of the problem, going from an $N \times \omega$ matrix to a $N \times \omega'$ matrix, where ω' is the number of equivalence classes of tables that are symmetric counterparts of each other. The vector $\mathbf{w}$ then also becomes ω'-dimensional, each entry representing whether or not to include the corresponding group of tables. We should consequently also change the objective function to $\mathbf{w}^T\mathbf{g}$, where $\mathbf{g}$ is the vector containing the number of tables in each of the symmetry classes.

Let us make three remarks. First of all, as mentioned earlier, this approach only constructs a critical region for a given significance level α. It is thus not necessarily the case that $K^{\alpha_1} \subset K^{\alpha_2}$ whenever $\alpha_1 < \alpha_2$. Indeed, as (4.20) only looks for the "most efficient packing" of the (symmetry groups of) tables, it is possible that the tables in K^{α_1} are different from those in K^{α_2}. Secondly, when looking at tables with larger sample sizes, the entries a_{ij} can become quite small, especially for the more "extreme" tables. In order to partially prevent any rounding errors by the LP-solver, we may scale the matrix A by a factor k, still to be determined, and apply the condition $kA\mathbf{w} \leq k\alpha\mathbf{1}$ in (4.20) instead. We will use this idea in Section 4.3.2.2. Finally, the solution $\mathbf{w}$ of (4.20) is not necessarily unique. This

can be seen from Figure 4.2. Assume for a moment that the top-right solution in the right figure would not exist, and that instead the bottom-left solution would have been marked as optimal. The corresponding objective value would be 3. This objective value is however also achieved by the top-left selection of tables, making this too an optimal solution. If it turns out that there are two optimal solutions $\mathbf{w}^1$ and $\mathbf{w}^2$, we will adopt the convention that $\mathbf{w}^1$ will be the optimal solution the algorithm returns whenever $\max_{i=1,\ldots,N}(A\mathbf{w}^1)_i > \max_{i=1,\ldots,N}(A\mathbf{w}^2)_i$, i.e., whenever $\mathbf{w}^1$ results in the largest maximal size. In the case of Figure 4.2, this would mean we would pick the bottom-left solution as the optimal solution, as the maximal size comes very close to 0.45, while that of the top-left solution lies more around 0.40.

Usually, the branch-and-bound approach is able to find many optimal solutions, and the Gurobi optimiser can be programmed to look for a user-specified number of optimal solutions. The danger here is that if there turn out to be more optimal solutions than we asked Gurobi to look for, two LP problems which should yield the same solution after application of our decision rule on which optimal solution to pick, return different solutions. Suppose for example that we want to solve (4.20) two times, once with the constraint matrix A_1 and once with the constraint matrix A_2, where A_1 and A_2 are the same matrix, up to a permutation of the columns. This could happen if we construct the matrix based on two different symmetry conditions which actually yield the same symmetry classes (such as $\mathrm{S_P}$ and $\mathrm{S_V}$ in the 2×2 case). This permutation should not affect the optimal solution in any way, but the branch-and-bound approach could potentially construct a different search tree based on this difference in matrices. Consequently, it might find different optimal solutions first. If we asked the solver to look for more optimal solutions than there actually exist, this is no problem. The solver will just find all of them. However, if there are more optimal solutions then the number we asked to look for, the set of optimal solutions Gurobi returns will be different for the two instances of the same problem. If this is the case, our decision rule might yield different solutions as it maximises over a different pool of solutions. One way to remove this discrepancy is to make sure that the columns in A_1 and A_2 are in the same order. However, the problem remains that we might miss the optimal solution which should be chosen according to our decision rule, whenever there are more optimal solution than we actually look for. An interesting problem to look into would be to find out whether we can determine the size of the pool of optimal solutions beforehand, to make sure we do not leave out any optimal solutions.

4.3.2.1 A few other packing problems

The objective function $\mathbf{w}^T\mathbf{1}$ in (4.20) is by far the most simple one that we could come up with. In an attempt to more accurately represent the problem that supremum tests are trying to solve, we came up with a few other possible linear programming problems. Recall that the initial goal in Barnard's procedure, and every other supremum test, is to find a combination of tables such that the sum of their respective $P(\cdot;\boldsymbol{\theta})$-functions is a function of $\boldsymbol{\theta}$ that is as constant as possible, and as close to α as possible. One possible way to achieve that is to force

$$\min_{i=1,\ldots,N} \sum_{\mathbf{x}\in K^\alpha} P(\mathbf{x};\boldsymbol{\theta}^i) \tag{4.21}$$

to be as large as possible. The idea here is that maintaining the constraint that

$$\sum_{\mathbf{x}\in K^\alpha} P(\mathbf{x};\boldsymbol{\theta}^i) \leq \alpha$$

for every possible $\boldsymbol{\theta}^i$ would essentially "squeeze" $\sum_{\mathbf{x}\in K^\alpha} P(\mathbf{x};\boldsymbol{\theta}^i)$ in as small a range as possible. By introducing some auxiliary variables $z \in \mathbb{R}_{\geq 0}$ and $\mathbf{v} \in \mathbb{R}^N_{\geq 0}$, we can turn (4.20) into a linear program that solves this "maximin" problem:

$$\begin{aligned} &\text{maximise } z \\ &\text{subject to } z = \min\{v_i : i = 1,\ldots,N\}, \\ &\qquad \mathbf{v} \leq A\mathbf{w}, \\ &\qquad A\mathbf{w} \leq \alpha\mathbf{1}, \\ &\qquad \mathbf{w} \in \{0,1\}^\omega,\ \mathbf{v} \in \mathbb{R}^N_{\geq 0}. \end{aligned} \tag{4.22}$$

There is however an immediate difficulty with this approach. Consider for simplicity the 2×2 case with $(n_{1\cdot}, n_{2\cdot}) = (20, 20)$. If we would include in our grid of $\boldsymbol{\theta}$-values a vertex of Θ_0, say $\boldsymbol{\theta} = (1, 0)$, then one of the rows of the matrix A would consist of a single one, with the remaining entries all zero. Indeed, only the table $(x_{11}, x_{21}) = (20, 20)$ (or its associate symmetry group) has a positive probability of occurring for $\boldsymbol{\theta} = (1, 0)$. This table cannot be included in the critical region, as we would violate our size constraint. Consequently, one of the entries of $A\mathbf{w}$ will be zero, and so will one of the entries of $\mathbf{v}$. Since the entries of $\mathbf{v}$ must be non-negative, we will always have $z = \min\{v_i : i = 1,\ldots,N\} = 0$. This entails that basically all critical regions (which do not contain the table $(x_{11}, x_{21}) = (20, 20)$) will be optimal.

Thus, we cannot include any vertices of Θ_0 with this approach. Including a $\boldsymbol{\theta}$-value that lies close to one of these vertices also turns out to be a bad idea. The values of the $P(\cdot;\boldsymbol{\theta})$-functions at this $\boldsymbol{\theta}$-value will be rather small for most tables. Because the determination of the critical region is solely based on the values of $P(\cdot;\boldsymbol{\theta})$ at that $\boldsymbol{\theta}$-value, we might obtain a critical region containing one (not so extreme) table which happens to have a relatively large value of $P(\cdot;\boldsymbol{\theta})$, and a few other tables. This would not make for a very powerful test. It also seemed that the solver software had trouble dealing with these values. We will come back to this in Section 5.2.

Instead of working with the minimum over all entries of $\mathbf{v}$, an idea might be to only look at the k-th smallest entry. We have not worked out this approach further as it would entail the introduction of yet another parameter to decide on, but it would for sure be an interesting numerical study to find out what value of k would give the most powerful test.

Another approach is to maximise the area under the $\sum_{\mathbf{x}\in K^\alpha} P(\mathbf{x};\boldsymbol{\theta}^i)$-function. That is, we aim to maximise the integral of the power function over Θ_0. Again, since it is bounded from above by α, maximising its area hopefully will make it lie as close as possible to α for as many values of $\boldsymbol{\theta}^i$ as possible. In order to turn this aim into a suitable objective function for a linear programming formulation,

we will perform a Quasi-Monte Carlo approximation of the area, i.e.,

$$\begin{aligned}\int_{\Theta_0} \sum_{\mathbf{x}\in K^\alpha} P(\mathbf{x};\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} &= \int_{\Theta_0} \sum_{j=1}^{O} w_j P(\mathbf{x};^j \boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \\ &\approx \sum_{i=1}^{N}\sum_{j=1}^{O} w_j P(\mathbf{x};^j \boldsymbol{\theta}^i) \\ &= \sum_{i=1} \mathbf{w}^T A_i \\ &= \mathbf{w}^T A^T \mathbf{1},\end{aligned}$$

where A_i is the i-th row of the matrix A we also used in (4.20). More on this numerical integration procedure can be read in Appendix A. The according linear program is given by (4.23).

$$\begin{aligned}&\text{maximise } \mathbf{w}^T A^T \mathbf{1} \\ &\text{subject to } A\mathbf{w} \leq \alpha \mathbf{1}, \\ &\qquad \mathbf{w} \in \{0,1\}^\omega.\end{aligned} \tag{4.23}$$

As a side effect, we hoped that $\beta(\boldsymbol{\theta})$ will be large for $\boldsymbol{\theta} \in \Theta_1$. However, as we will see in Section 5.2, (4.23) seemed to be "overfitting" too much. It would manage to find a collection of tables for which $\sum_{\mathbf{x}\in K^\alpha} P(\mathbf{x};\boldsymbol{\theta}^i)$ is very close to α for many $\boldsymbol{\theta}$-values, but this set of tables did not form an "intuitive" critical region and had very small power for $\boldsymbol{\theta} \in \Theta_1$.

We can improve on the previous approach by recalling how we got to the idea of using a linear programming formulation in the first place. We were trying to maximise the power function for $\boldsymbol{\theta} \in \Theta_1$, while keeping the power below α for $\boldsymbol{\theta} \in \Theta_0$. Thus, instead of maximising the integral of the power over Θ_0, why not maximise the integral of the power over Θ_1 instead? We still use a Quasi-Monte Carlo grid of $\boldsymbol{\theta}^i$-values over Θ_0 to check that the size constraint is satisfied, but we can additionally create a grid of $\boldsymbol{\theta}^i$-values over Θ_1 that we can use to approximate the integral $\int_{\Theta_1} \sum_{\mathbf{x}\in K^\alpha} P(\mathbf{x};\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$ just as we did earlier. This we can write as the following linear program.

$$\begin{aligned}&\text{maximise } \mathbf{w}^T B^T \mathbf{1} \\ &\text{subject to } A\mathbf{w} \leq \alpha \mathbf{1}, \\ &\qquad \mathbf{w} \in \{0,1\}^\omega,\end{aligned} \tag{4.24}$$

where again, A is the matrix of $P(\mathbf{x};\boldsymbol{\theta}^i)$-values for the different tables and grid points in Θ_0. The matrix B also consists of $P(\mathbf{x};\boldsymbol{\theta}^i)$-values, but now the $\boldsymbol{\theta}^i$-values are grid points in Θ_1. Just keep in mind: a grid point in Θ_0 is a vector of c entries that sum up to 1, while a grid point in Θ_1 is a vector of $r \times c$ entries where the first c entries represent θ_1 (and must therefore sum to 1), the next set of c entries represent θ_2 (and again sum to 1), and so on.

4.3.2.2 A binary search

All the linear programming problems we have formulated so far, only give us a critical region for a given α. Let us now try to extend the LP-approach in order

to find a p-value for some observed table. One approach would just be to use again use an external test statistic to order all the tables in the obtained critical region, and then apply the maximisation procedure to compute the p-values. We could even perform a CSM-like test, only considering the tables in the critical region. However, these approaches would almost completely nullify the decrease in computation time we make by using the LP method, for it is the maximisation procedure itself that takes a lot of time. Furthermore, the critical region we obtain from an LP test might well be very similar to that we would have obtained with one of the supremum tests, containing in particular all the most extreme tables that the CSM-like method will start out with either way. It is only towards the end of the maximisation procedure, that the pool of tables to consider for the ordering will be meaningfully smaller. We might thus just as well have started immediately with a CSM-like test.

Therefore, we would like to stay away from using the already discussed methods of computing p-values, and instead try to come up with a solution based solely on solving LP problems (which can be done efficiently). To this end, realise that if we apply the LP approach for different values of α, we should be able to iteratively find the smallest value of α for which the table of interest is still in the critical region. This is exactly the p-value as defined in (2.6). We will describe two attempts. The first one, as we will see, does not entirely succeed at solving the problem at hand, but is still very instructive to mention.

Let us first fix a value of α, and set $\alpha_0 := \alpha$. By applying the LP-method, we can find the corresponding critical region K^{α_0}. We are interested in finding the p-value $p(\mathbf{x})$ of the observed table $\mathbf{x}$. Now, if $\mathbf{x} \notin K^{\alpha_0}$, we can choose to for example double α_0 until $\mathbf{x} \in K^{\alpha_0}$, or set $\alpha_0 = 1$ if $2\alpha_0 > 1$. Therefore, assume that $\mathbf{x} \in K^{\alpha_0}$. This tells us already that $p(\mathbf{x}) \leq \alpha_0$. Therefore, let us again apply the LP-method, but now with significance level $\alpha_1 := \alpha_0/2$. Furthermore, we limit our linear program to only use the tables in K^{α_0}. This enforces that $K^{\alpha_1} \subset K^{\alpha_0}$. We now check if $\mathbf{x} \in K^{\alpha_1}$. If so, we define $\alpha_2 := \alpha_1/2$ and compute K^{α_2}, using only the tables in K^{α_1}. This further restricts $p(\mathbf{x})$ to the range $(0, \alpha_1]$. If $\mathbf{x} \notin K^{\alpha_1}$ however, we define $\alpha_2 := \alpha_1 + \alpha_1/2$. Furthermore, we will solve a slightly modified version of (4.20). Since we know which tables are already contained in K^{α_1}, we only need to find the binary vector $\mathbf{w}$, indicating which (groups of) tables to include from $\mathcal{P} := K^{\alpha_0} \setminus K^{\alpha_1}$, which maximises $\mathbf{w}^T\mathbf{1}$ subject to

$$\sum_{\mathbf{y} \in K^{\alpha_1}} P(\mathbf{y}; \boldsymbol{\theta}^i) + \sum_{j=1}^{|\mathcal{P}|} w_j P(\mathbf{y}^j; \boldsymbol{\theta}^i) \leq \alpha_2,$$

for all $i \in \{1, \ldots, N\}$. We indexed all the tables in $\mathcal{P}$ with the index j. We can translate this to the linear program

$$\begin{aligned} &\text{maximise } \mathbf{w}^T\mathbf{1} \\ &\text{subject to } A_{\mathcal{P}}\mathbf{w} \leq \alpha_2\mathbf{1} - \mathbf{b}, \\ &\qquad \mathbf{w} \in \{0,1\}^{|\mathcal{P}|}, \end{aligned} \tag{4.25}$$

where

$$\mathbf{b} := \left(\sum_{\mathbf{y} \in K^{\alpha_1}} P(\mathbf{y}; \boldsymbol{\theta}^i) \right)_{i=1,\ldots,N}$$

and $A_{\mathcal{P}}$ contains the columns of A corresponding to the (groups of) tables in $\mathcal{P}$. We have now further restricted the p-value to $p(\mathbf{x}) \geq \alpha_1 \geq \max_{i=1,\ldots,N} b_i$. We can repeat this procedure again and again, narrowing down the range in which $p(\mathbf{x})$ could lie. At each iteration, we also reduce the pool of tables which we should consider to construct the next critical region. Indeed, if at the k-th iteration, we have $\mathbf{x} \in K^{\alpha_k}$, we can get rid of all tables in $K^{\alpha_k} \setminus K^{\alpha_{k-1}}$. If instead $\mathbf{x} \notin K^{\alpha_k}$, we can get rid of all tables in K^{α_k} and replace them by the vector $\mathbf{b}$ in (4.25). The new pool of tables is given by the solution of (4.25). By construction, we will eventually end up with just one table in the pool, which is necessarily $\mathbf{x}$. The (approximate) p-value can then be found again just as in (2.21):

$$p(\mathbf{x}) = \sup_{i=1,\ldots,N} \big\{\mathbf{b} + P(\mathbf{x}; \boldsymbol{\theta}^i)\big\},$$

since $\mathbf{b}$ is the sum of $P(\cdot; \boldsymbol{\theta}^i)$-functions of all tables which are more "extreme" than $\mathbf{x}$, in the sense that the LP-method would have preferred to use these tables over $\mathbf{x}$ to construct a critical region for a small enough significance level.

We should be cautious to keep track of the "base vector" $\mathbf{b}$. Once we have had to solve (4.25) once, our pool of tables no longer coincides with a critical region. Therefore, if after some iterations, it turns out that $\mathbf{x}$ is in the pool of tables, we should not solve (4.20), but instead (4.25) with $\mathcal{P}$ the current pool of tables and $\mathbf{b}$ the current base vector.

Note furthermore that each time we need to solve (4.25), we only consider a subset of the columns of A. The pool $\mathcal{P}$ of tables will contain only tables which have similar values for $\sup_{i=1,\ldots,N} P(\cdot; \boldsymbol{\theta}^i)$, as they either were in the same critical region, or they were all left out of the critical region at the same iteration because the significance level was too small. Consequently, dividing each entry in $A_{\mathcal{P}}$ by the constant $c = \sup_{\mathbf{y} \in \mathcal{P}} \sup_{i=1,\ldots,N} P(\mathbf{y}; \boldsymbol{\theta}^i)$, will get rid of any numerical issues related to working with small numbers. We found that performing this rescaling at each iteration greatly improved the quality of the solutions returned by the LP solver. The whole procedure is summarised in Algorithm 1.

The critical reader might have already spotted an issue with this approach. We could end up with a different p-value if we would have started with a different initial significance level α. Indeed, since the critical regions are not necessarily nested, there might exist values α and α' with $\alpha' < \alpha$ such that $\mathbf{x} \notin K^{\alpha}$ but $\mathbf{x} \in K^{\alpha'}$. Consequently, we could find $p(\mathbf{x}) > \alpha$ if we start our binary search at α, while starting at α' would yield $p(\mathbf{x}) \leq \alpha'$. However, we are still able to uniquely define a p-value via (2.6) with $T(\mathbf{x}) = \mathbf{x}$. We just need to find the smallest value of α for which the observed table $\mathbf{x}$ is still in K^{α}.

Notice that we also introduced an additional maximum $k_{\max}$ on the number of iterations that we want to do. This is essentially a specification of how precise we want our approximation to be. As long as we are not able to create two different critical regions which correspond to levels which are $2^{-k_{\max}}$ apart, which is of course the case for a set of sample sizes that grows with increasing $k_{\max}$, we will have $\mathcal{P} = \{\mathbf{x}\}$ before $k_{\max}$ will be reached. However, we did observe that even for relatively small sample sizes (say $(n_{1\cdot}, n_{2\cdot}) = (25, 25)$), $k_{\max}$ was, for certain tables, reached before $\mathcal{P} = \{\mathbf{x}\}$. This can happen if, for some level α_k we encounter in one of the iterations, the observed table $\mathbf{x}$ is in K^{α_k}, but, for all $l > k$, $\mathbf{x} \notin K^{\alpha_l}$. This can happen by coincidence (since the critical regions are not nested), or perhaps

Algorithm 1: `BinarySearch(`$\mathbf{x}, \alpha . k_{\max}$`)`

Input: an observed table $\mathbf{x}$, an initial significance level $\alpha_0 := \alpha$, and the maximum number $k_{\max}$ of iterations we are willing to do
Output: an approximate p-value $p(\mathbf{x})$
construct K^{α_0} by (4.20);
if $\mathbf{x} \notin K^{\alpha_0}$ **then**
  `BinarySearch(`$\mathbf{x}, 2\alpha \wedge 1$`)`;
end
else
  define the level $\alpha_1 = \alpha_0/2$;
  define the base vector $\mathbf{b} = \mathbf{0}$;
  construct the pool $\mathcal{P} = K^{\alpha_1}$ by (4.20) using only tables from K^{α_0};
  set $k = 2$;
  while $\mathcal{P} \neq \{\mathbf{x}\}$ *and* $k \leq k_{max} + 1$ **do**
    if $\mathbf{x} \in \mathcal{P}$ **then**
      set $\alpha_k = \alpha_{k-1} - \alpha_0 \cdot 2^{-k}$;
      construct the new $\mathcal{P}$ by (4.25) using only tables from the current $\mathcal{P}$ and $\mathbf{b}$ as base vector;
    end
    else
      set $\mathbf{b} = \mathbf{b} + \left(\sum_{\mathbf{y} \in \mathcal{P}} P(\mathbf{y}; \boldsymbol{\theta}^i)\right)_{i=1,\ldots,N}$;
      set $\alpha_k = \alpha_{k-1} + \alpha_0 \cdot 2^{-k}$;
      construct the new $\mathcal{P}$ by (4.25) using only tables from the current $\mathcal{P}$ and $\mathbf{b}$ as base vector;
    end
    set $k = k + 1$;
  end
end
define $p(\mathbf{x}) = \sup_{i=1,\ldots,N} \left\{\mathbf{b} + P(\mathbf{x}; \boldsymbol{\theta}^i)\right\}$;

because the p-value is actually exactly equal to α_k. What happens next will be the same either way; at each iteration we will execute the `else` statement in the `while` loop of Algorithm 1, meaning that the sequence $(\alpha_l)_{l>k}$ will monotonically converge to α_k, but only reach it in the limit. Introducing $k_{\max}$ is a lazy way to deal with this problem, avoiding the finding algorithm to go on for too long. This should be done with care however; at a certain sample size the chosen value of $k_{\max}$ will not be large enough, meaning that we would prematurely cut off the search while the scenario we just sketched did not at all occur. However, at the sample sizes we considered, setting $k_{\max}$ to 100 turned out to be sufficient.

4.3.2.3 Extending the binary search

The binary search method we just described will return some (arguably small) value of α for which $\mathbf{x}$ is still in K^{α}. It might not be the smallest such α, but it will serve as a good starting point for the iterative method we are about to describe. We can check very easily for a number of smaller significance levels $\alpha' < \alpha$, say between $\alpha/2$ and α, whether or not $\mathbf{x} \in K^{\alpha'}$ by solving (4.20). If,

for some α' we try out, it turns out that $\mathbf{x} \in K^{\alpha'}$, we set this α' as our new p-value, and we can repeat the searching procedure. We do this until the search no longer finds any smaller value for α', as indicated in Algorithm 2. This is of course not a guaranteed way to find the smallest significance level. We are trying a finite number of α'-values and might very well miss the smallest level. It is thus a trade-off between how much time we want to spend looking for a smaller α'-value and how satisfied we are with the upper bound for the actual p-value we already have.

Algorithm 2: `ExtendedBinarySearch(x,`α`,`m`)`

Input: an observed table $\mathbf{x}$, an initial significance level $\alpha_0 := \alpha$, and a number m of trial values to find a smaller significance level
Output: an approximate p-value $p(\mathbf{x})$
compute the initial guess for the p-value $p_0 =$ `BinarySearch(x,`α`)`;
set $k = 0$ and $p_{-1} = 1$;
while $p_k \neq p_{k-1}$ **do**
  set $p_{k-1} = p_k$;
  for $\alpha' \in \left\{ \frac{p_k}{2} + \frac{j}{m+1} \cdot \frac{p_k}{2} : j = 1, \ldots, m \right\}$ **do**
    if $\mathbf{x} \in K^{\alpha'}$ **then**
      set $p_k = \alpha'$;
      break;
    end
  end
  set $k = k + 1$;
end
define $p(\mathbf{x}) = p_k$;

4.3.2.4 Validity of the p-value

An important question to ask is whether or not the (approximate) p-value obtained with one of the LP tests will be valid. We will argue that because of the inevitable non-nestedness of the critical regions, the p-values of the LP tests defined as the smallest level α for which the observed table is still in K^α cannot be valid.

To this end, consider three tables $\mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3 \in \Omega$. In theory, one way to find the respective p-values $p(\mathbf{x}^1)$, $p(\mathbf{x}^2)$, and $p(\mathbf{x}^3)$ is to let the level α run from 0 up to 1, and record the first values of α for which $\mathbf{x}^1$, $\mathbf{x}^2$, and $\mathbf{x}^3$ get first included in the critical region K^α. We call these values α_1, α_2, and α_3 respectively, and furthermore assume without loss of generality that $\alpha_1 < \alpha_3$. By definition, we have that $p(\mathbf{x}^i) = \alpha_i$ for $i \in \{1, 2, 3\}$.

Since the critical regions need not be nested, it can occur that $K^{\alpha_3} = \{\mathbf{x}^2, \mathbf{x}^3\}$. This is enough to disprove the validity of the p-value. First note that by definition of the p-value, we have that $p(\mathbf{x}^2) = \alpha_2 \leq \alpha_3$. Indeed, since $\mathbf{x}^2$ is included in K^{α_3}, the smallest level for which $\mathbf{x}^2$ is rejected is at most α_3. But then, realise that

$$P_{H_0}(p(\mathbf{X}) \leq \alpha_3) = \sum_{p(\mathbf{x}) \leq \alpha_3} P(\mathbf{x}; \boldsymbol{\theta}) \geq P(\mathbf{x}^1; \boldsymbol{\theta}) + P(\mathbf{x}^2; \boldsymbol{\theta}) + P(\mathbf{x}^3; \boldsymbol{\theta}).$$

We claim that there must exist some $\boldsymbol{\theta} \in \Theta_0$ for which this probability is larger than α_3, which shows that $p(\mathbf{X})$ is not a valid p-value. Indeed, suppose instead

that for all $\theta \in \Theta_0$,

$$P(\mathbf{x}^1;\boldsymbol{\theta}) + P(\mathbf{x}^2;\boldsymbol{\theta}) + P(\mathbf{x}^3;\boldsymbol{\theta}) \leq \alpha_3.$$

That is, the set $\{\mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3\}$ satisfies the size constraint and is thus a feasible solution of the LP problem. No matter which of the four objective functions we have discussed in this Section, the objective value corresponding to $\{\mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3\}$ will always be larger than the one corresponding to $\{\mathbf{x}^2, \mathbf{x}^3\}$, meaning that actually we have $K^{\alpha_3} = \{\mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3\}$, contradicting our assumption that $\mathbf{x}^1 \notin K^{\alpha^3}$. Hence, indeed, we have $P_{H_0}(p(\mathbf{X}) \leq \alpha_3) > \alpha_3$ for some $\boldsymbol{\theta} \in \Theta_0$, showing that the p-value obtained with the LP test is not valid.

4.3.3 Another problem: Reduction to $r \times 2$ tables

When looking at the null hypothesis (4.3), realise that we are in fact performing a test for a multiple null hypothesis. Instead of tackling this large null hypothesis all at once, we could also try to test each of the constituent null hypotheses separately. This is the idea behind the concept of table reduction we will briefly mention here.

Table 4.1 can be rewritten to a set of $c-1$ tables with only two columns as shown in Table 4.5.

	1	Not 1	
1	x_{11}	$\sum_{j=2}^{c} x_{1j}$	$n_{1\cdot}$
2	x_{21}	$\sum_{j=2}^{c} x_{2j}$	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	x_{r1}	$\sum_{j=2}^{c} x_{rj}$	$n_{r\cdot}$
	$n_{\cdot 1}$	$\sum_{j=2}^{c} n_{\cdot j}$	$n_{\cdot\cdot}$

	2	Not 1,2	
1	x_{12}	$\sum_{j=3}^{c} x_{1j}$	$n_{1\cdot} - n_{\cdot 1}$
2	x_{22}	$\sum_{j=3}^{c} x_{2j}$	$n_{2\cdot} - n_{\cdot 1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	x_{r2}	$\sum_{j=3}^{c} x_{rj}$	$n_{r\cdot} - n_{\cdot 1}$
	$n_{\cdot 2}$	$\sum_{j=3}^{c} n_{\cdot j}$	$n_{\cdot\cdot} - n_{\cdot 1}$

$\vdots$

	$c-1$	c	
1	$x_{1,c-1}$	x_{1c}	$x_{1,c-1} + x_{1,c}$
2	$x_{2,c-1}$	x_{2c}	$x_{2,c-1} + x_{2,c}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	$x_{r,c-1}$	x_{rc}	$x_{r,c-1} + x_{rc}$
	$n_{\cdot,c-1}$	$n_{\cdot c}$	$n_{\cdot,c-1} + n_{\cdot c}$

Table 4.5: Reduction of Table 4.1 into $c-1$ $r \times 2$ tables.

We are now able to perform Barnard's CSM test on each of these tables separately. The advantage of that is that each CSM test now only involves maximisation over a one-dimensional simplex only, and we only need to look through a smaller set of tables. Indeed, instead of doing one test which goes through ω tables, where ω is given in (4.7), we perform $c-1$ tests that go through $\prod_{i=1}^{r}(n_{i\cdot}+1)$ tables. A quick computation shows that $(c-1)\prod_{i=1}^{r}(n_{i\cdot}+1)$ grows a lot less quicker than ω for increasing r, c or $(n_{i\cdot})_{i=1,\dots,r}$. Furthermore, as discussed in Section 4.2, we can still define a nice geometry-based convexity condition which is exactly the same as in the 2×2 case.

However, one difference is that because we now have a total of $c-1$ tests to perform, we also end up with $c-1$ p-values. How should we interpret such a result? If we test $c-1$ hypotheses separately, and require for each individual test the probability of a Type I error to be at most α, then the probability of making at least one Type I error over these $c-1$ tests is $1-(1-\alpha)^{c-1}$, assuming independence of the tests. For $\alpha = 0.05$ and $c = 5$, this is already a probability of 0.185. Different ways exist to handle this "multiple testing problem". For example, in order to control the *family-wise error rate* (FWER), i.e., the probability of making at least one Type I error, one could apply a Bonferroni correction. If we denote the, say, k null hypotheses we are testing for by $H^1,\dots,H^k$, and the respective valid p-values we obtain from those tests by $p^1,\dots,p^k$. Then the Bonferroni correction amounts to rejecting each hypothesis H^i for which $p^i \le \alpha/k$, where α is the desired upper bound for the FWER. By Boole's inequality, this procedure guarantees that α is actually upper bound for the FWER. Indeed, if we assume that k_T out of the k null hypotheses are actually true,

$$\text{FWER} = P_{H_0}\left(\bigcup_{i=1}^{k_T}\left\{p^i \le \frac{\alpha}{k}\right\}\right) \le \sum_{i=1}^{k_T} P_{H_0}\left(p^i \le \frac{\alpha}{k}\right) \le k_T\frac{\alpha}{k} \le \alpha.$$

In the penultimate inequality we used that the p-values are valid. One should realise that this approach is rather conservative. Especially for large k it will become very rare to actually reject one of the null hypotheses. Several other approaches can improve on this, and we recommend the paper by Goeman and Solari [74] to learn more on the topic, but whatever approach we choose, we will in one way or another always lose some test power by considering multiple hypotheses.

Furthermore, the reduction into $c-1$ $r\times 2$ tables essentially tries to answer a different question. In deciding whether or not to reject the general null hypothesis (4.3), we are not interested in finding out for which categorical outcome the association between the groups breaks down or not. Instead, the reduction will point us to which of the $c-1$ constituting null hypotheses in (4.3) we should actually reject. Because of this, and because of the complications of dealing with error rates in the case of multiple hypotheses, we will not be able to properly compare the reduction approach with the other ideas in this thesis. Although the author of this thesis would love to find out if this approach would be fruitful, it will not be researched in more depth in the remainder of this text.

CHAPTER

FIVE

RESULTS

Let us actually start trying out some of the methods described in the previous chapters. First of all, as the goal was to construct statistical tests on contingency tables that surpass a standard 2×2 size, let us see how large we can make the tables while still keeping the computation time "within an acceptable range", whatever that may mean. Afterwards, we want to see how "powerful" the different tests are. As we have discussed in Chapter 3, this might not seem a useful exercise for those in favour of conditional tests, but one might still interested in how the argument of unconditional tests being more powerful holds up when the table sizes start increasing. We will therefore keep in mind the trend observed by Mehta and Hilton [55], that the difference in power between the tests decreases for increasing table and sample sizes, when analysing the results from our power study.

In our comparisons, we will consider the following tests:

- *Fisher's Exact Test.* For larger tables, this will be the Fisher–Freeman–Hilton test. As mentioned in countless papers already [24], [25], we will expect this test to be rather conservative. We will make this more precise in Section 5.4. We will make use of the already existing `R` implementation `fisher.test` here.

- *Pearson's Chi-Square Test.* In particular, we will use the chi-square test with the continuity correction proposed by Yates [10], using (2.9) as test statistic. This will make this test comparable to Fisher's exact test. We will use the function `chisq.test`, which is again the existing `R` implementation.

- *Barnard's CSM-like Test.* With this, we refer to the CSM test as described in Section 4.2 (or rather the $\mathrm{CS_PM}$ test), as well as two variants, where we determine the symmetry classes using the S_χ or $\mathrm{S_V}$ conditions.

- *Suissa and Shuster's Unconditional Test.* Here we use the chi-square test statistic as an external test statistic to determine the ordering of the tables. We also define a variant of this test by using the mean value test statistic $\int_{\Theta_0} P(\mathbf{X};\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$ instead, and one by using the Fisher p-value as a test statistic. Recall from Proposition 2.4 that this is nothing else than Boschloo's test.

- *Linear Programming Tests.* Finally, we will consider the tests which convert our problem into a linear program, as described in Section 4.3.2. Numerous

variations are possible here too. We can construct a test based on each of the linear programs (4.20), (4.22), (4.23) and (4.24). For each of these versions, we can furthermore choose which of three definitions for symmetric tables to use; the $\mathrm{S_P}$, S_χ, or $\mathrm{S_V}$ conditions.

Recall that the power of a test is given by $\beta(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_r) = P_{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_r}(T(\mathbf{X}) \in K^\alpha)$. Alternatively, we can write this expression in terms of the p-value as $\beta(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_r) = P_{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_r}(p(\mathbf{X}) \leq \alpha)$. Because we are working in a discrete setting, we can actually exactly compute the power for any given $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_r) \in \Theta$ pretty easily by enumerating over all the tables $\mathbf{x}$ which have $p(\mathbf{x}) \leq \alpha$, i.e.,

$$\beta(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_r) = \sum_{\mathbf{x} \in K^\alpha} P_{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_r}(\mathbf{X} = \mathbf{x}) = \sum_{\mathbf{x} \in \Omega : p(\mathbf{x}) \leq \alpha} P_{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_r}(\mathbf{X} = \mathbf{x}). \tag{5.1}$$

One should be mindful of the notation here. We have replaced the subscript $\boldsymbol{\theta}$, representing the common value under the null hypothesis, by the "vector of vectors" $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_r) \in \Theta$. We thus have

$$P_{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_r}(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^{r} \prod_{j=1}^{c} \frac{n_{i\cdot}!}{x_{ij}!} \theta_j^{x_{ij}} = \frac{\prod_{i=1}^{r} n_{i\cdot}!}{\prod_{i=1}^{r} \prod_{j=1}^{c} x_{ij}!} \prod_{i=1}^{r} \prod_{j=1}^{c} \theta_{ij}^{x_{ij}}. \tag{5.2}$$

Given the critical regions or p-values obtained from two tests, we can easily compare their power for any given $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_r) \in \Theta$ by computing the probability of a table ending up in the critical region or having a p-value smaller than α via (5.1). Note that if $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_r) \in \Theta_0$, we refer to $\beta(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_r) = \beta(\boldsymbol{\theta})$ as the size of the test. In this case we will write $\boldsymbol{\theta}$ to save space (this comes with the abuse of notation that $\boldsymbol{\theta} \in \Theta_0$). In the Neyman–Pearson framework, this size is bounded by α. Furthermore, for $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_r) \in \Theta_1$, $\beta(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_r)$ represents the probability of rejecting the null hypothesis given that the null hypothesis is false. Of course, we want this probability to be as large as possible, minimising the probability of a Type II error. In the power comparisons that will follow, we will thus always be searching for the most powerful test. That is, we will be looking for the test with the largest power for $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_r) \in \Theta_1$, while keeping $\beta(\boldsymbol{\theta}) \leq \alpha$ for $\boldsymbol{\theta} \in \Theta_0$.

5.1 Grid Size

The first step in determining how large we can make tables in order for the computation time to be "within an acceptable range" is to find out how large of a grid for Θ_0 we actually need. The CSM and LP tests perform some kind of maximisation over values of $\boldsymbol{\theta}$ in Θ_0, and to limit the computation time we would like to keep this amount of $\boldsymbol{\theta}$-values to a minimum, without sacrificing too much on test power and validity. Indeed, if we would perform the maximisation over too few points, this also means that we enforce the test size to be smaller than α over few points, and it might thus occur that after we have selected the table in our critical region, we have $P_{\boldsymbol{\theta}}(\mathbf{X} \in K^\alpha) > \alpha$ for some $\boldsymbol{\theta} \in \Theta_0$ which was not included in the grid.

To investigate the effect of the grid size on the test performance, we will look at two unconditional tests: Barnard's CSM test where we utilise the $\mathrm{S}_t extP$ condition and the LP test based on (4.20) with the $\mathrm{S}_t extP$ condition. We will execute these tests on tables with different sizes and sample sizes, using different grid sizes. The

CSM test will return an ordering of the tables. Given a table size and sample sizes, the outcome space Ω is fully determined. We will then perform the test for a very large grid size ($N = 1000$), of which we assume that this grid size is large enough for the problem sizes we have been working with. This guess is based on experience, and will turn out to be large enough in a bit. The ordering we obtain from this will serve as our benchmark ordering. We will now decrease or increase the grid size in a similar fashion to how we altered ℓ in Algorithm 1 and perform the CSM test with this grid size, as long as it takes to find the smallest $N = N'$ that still results in the same ordering as the benchmark ordering. It is at N' that we can say with a certain degree of confidence that a finer grid is not necessary. For the LP test, we will perform the same procedure, but instead, given the significance level $\alpha = 0.05$, search for the smallest grid size for which the obtained critical region is the same as the benchmark critical region. This procedure is written out for the CSM test in Algorithm 3.

Algorithm 3: Find minimal grid size N.

Input: sample sizes $(n_{i\cdot})_{i=1,\dots,r}$, a number of columns c, and the large benchmark grid size $N_b = 1000$
Output: the minimal grid size N'
construct the outcome space Ω for the given values of $(n_{i\cdot})_{i=1,\dots,r}$ and c;
divide the sample space into symmetry classes according to the desired symmetry condition, in this case $\mathrm{S}_t extP$;
let o_b be the ordering of Ω using the CSM test with grid size N_b;
set $k = 2$, $N_{\text{old}} = N_b$ and $N_{\text{new}} = N_b/2$;
while $|N_{old} - N_{new}| > 1$ **do**
  set $N_{\text{old}} = N_{\text{new}}$;
  let o_{new} be the ordering of Ω using the CSM test with grid size N_{new};
  if $o_{new} = o_b$ **then**
    set $N_{\text{new}} = N_{\text{new}} - N_b \cdot 2^{-k}$;
  end
  else
    set $N_{\text{new}} = N_{\text{new}} + N_b \cdot 2^{-k}$;
  end
  set $k = k + 1$;
end
return $N_{\text{old}} \vee N_{\text{new}}$;

Notice that with this method we are guaranteed to find the smallest necessary grid size, as the QMC grids are nested for increasing grid sizes (see Appendix A). If for grid sizes $N_1 < N_2$, we find that the CSM test with grid size N_1 still yields the same ordering as the benchmark ordering, then so will the CSM test with grid size N_2.

We aim to find the minimal grid size for the following combinations of table

sizes and sample sizes:

$$\begin{aligned}
2 \times 2: &\quad (n_{1\cdot}, n_{2\cdot}) = (n, n) \text{ with } n \in \{5k : k = 1, \ldots, 6\}, \\
2 \times 3: &\quad (n_{1\cdot}, n_{2\cdot}) = (n, n) \text{ with } n \in \{5k : k = 1, \ldots, 6\}, \\
3 \times 2: &\quad (n_{1\cdot}, n_{2\cdot}, n_{3\cdot}) = (n, n, n) \text{ with } n \in \{5k : k = 1, \ldots, 6\},
\end{aligned}$$

For each of these tables we computed N' via Algorithm 3. The mentioned sample sizes, however, turned out to be a bit too optimistic in the case of the CSM tests. The results are shown in Table 5.1. A hyphen (-) indicates that the computation would have taken too long (in the order of magnitude of multiple hours). We will come back to this in Section 5.3.

n	2×2	3×2	2×3
5	1	385	96
10	58	62	-
15	26	-	-
20	33	-	-
25	67	-	-
30	226	-	-

n	2×2	3×2	2×3
5	1	1	67
10	1	53	127
15	29	63	1000
20	26	18	-
25	788	600	-
30	1000	-	-

Table 5.1: Minimal grid size N' yielding the same ordering as $N_b = 1000$ for the CSM test (left) and the LP test (right).

For the considered table and sample sizes, it seems that especially for smaller tables the minimal grid size is well below $N_b = 1000$. However, we also see that the increase in table and sample sizes also leads to an increase in N'. Eventually, in the right table, we even see a minimal grid size of $N' = N_b$, meaning that the ordering with $N = 999$ was already different than that with $N = 1000$. This might lead one to think that, when considering larger tables, $N_b = 1000$ grid points will no longer be sufficient to obtain the optimal ordering. One might however ask the question how good the test performs, even though we know that we might not have the optimal ordering. In that case, we can still make sure that the test is valid. Indeed, realise that when performing the CSM test, at each iteration, we make use of the grid $\{\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^N\}$-values to compute the maximal values of (4.15) for each candidate table $\mathbf{y}$. If the table $\mathbf{y}'$ yields the smallest maximum value for (4.15), we set $\mathbf{x}_{(k+1)} = \mathbf{y}'$ and set $p(\mathbf{y}')$ as this smallest maximum value. Note that we essentially kill two birds with one stone here. First, we decide on the next table in the ordering. Second, we immediately retrieve the p-value from the computations we have just performed, and can use the vector $\left(\sum_{i=1}^{k+1} P(\mathbf{x}_{(i)}; \boldsymbol{\theta}^j)\right)_{j=1,\ldots,N}$ as the base layer for the next iteration.

Although this is all performed using the same grid, it does not need to. We can use one grid with size N_o to choose the next table in the ordering, and another grid with size N_p to compute the p-value. It may seem at first that there is no advantage in doing this, but by using a fine grid only for one part of the computations, we can drastically cut down the computation time. By using a fine grid to compute the p-value, we can at the same time remain more confident that the outputted p-value is still valid, as it is less likely that we missed the actual maximum with our fine grid than with our coarse grid.

To illustrate this point, let us look at a specific example concerning the outcome space of 2×2 tables with sample sizes $(n_{1\cdot}, n_{2\cdot}) = (20, 20)$. We will perform two series of CSM tests; using alternative symmetry condition such S_χ or S_V yielded similar conclusions as the ones that will come. One series in which we always use a fine grid of $N_p = 1000$ points to determine the p-value, but vary the size N_o of a coarse grid which we use to determine at each iteration which table to choose next in the ordering. The coarse grid size will take the values $N_o \in \{1, 3, 5, 7, 9, 50, 1000\}$. In another series, we will set the size of both grids equal to $N_o = N_p \in \{1, 3, 5, 7, 9, 50, 1000\}$. For both series of tests, we will compare the size and power of the tests at the level $\alpha = 0.01$ in order to see how big an effect varying the grid sizes has on the power. The test sizes are shown in Figure 5.1 as a function of θ. In the left plot, we see $\beta(\theta, \theta)$ for $N_p = 1000$ and for N_o indicated in the legend. Note that we only see 3 lines. In the case of an overlap, only the colour corresponding to the highest N_o can be seen. Thus, $N_o \in \{3, 5\}$ yield the same size function, and so do $N_o \in \{7, 9, 50, 1000\}$. Because the grid used to compute the p-value is so fine, we never exceed the $\alpha = 0.01$ line and so the p-values we obtain are valid. Note that this is not the case in the right plot, where N_p also takes small values at first. For $N_f \in \{1, 3, 5\}$, we clearly see that the fine grid consists of too few points to accurately compute the p-value. As a consequence, some tables receive an severe underestimate for their p-value, and so are unjustly included into the critical region, yielding in the probability of a Type I error exceeding α. Also note that from $N_f = 7$ onwards, the size functions for both series of tests coincide.

It is worth noticing in the left plot that the size function corresponding to $N_o = 1$ is significantly smaller than the other ones. This can be explained as follows. Because we determine the ordering by comparing the values of (4.15) at very few points, the next table in the ordering may be a table which yields the smallest values of (4.15) at these few points, but yields large values of (4.15) everywhere else. Consequently, when we determine the p-value of this table with the fine grid, this table will receive a large p-value. This p-value can even be so large that it exceeds α, meaning that it will not be included in the critical region, along with all tables that come after it in the ordering. By increasing N_o, we decrease the likelihood of including these "wrong" tables prematurely, such that we can keep constructing a critical region which will yield a $\beta(\theta, \theta)$ closer to the $\alpha = 0.01$ line.

The power $\beta(\theta_1, \theta_2)$ is shown in Figure 5.2 for variable N_o and fixed N_p and in Figure 5.3 for variable N_o and N_p. In both figures, we indicate the line $\theta_1 = \theta_2$, i.e., Θ_0. Figure 5.1 is thus a close-up of the power function on that line. We show for the indicated values of N_o and N_f both the power function $\beta(\theta_1, \theta_2)$, and the difference $\beta(\theta_1, \theta_2) - \beta_{\text{benchmark}}(\theta_1, \theta_2)$ of the power function with the power function with the benchmark power function obtained when using $N_o = N_f = 1000$. Also observe that in both series of tests, we do not only get the same size function for $N_o \in \{3, 5\}$, but in fact the same power function on the whole of Θ. This should not come as a surprise; both N_o-values yield the same ordering. The same is true for $N_o \in \{7, 9, 50, 1000\}$, and therefore we only show the power functions for the N-values that actually yielded a different power function. From $N_o = 7$ onward, there is no difference with the benchmark anymore, which can also be seen from the plot in the bottom-right corners of Figures 5.2 and 5.3.

We can also perform the above exercise for the LP test. In contrast to the CSM

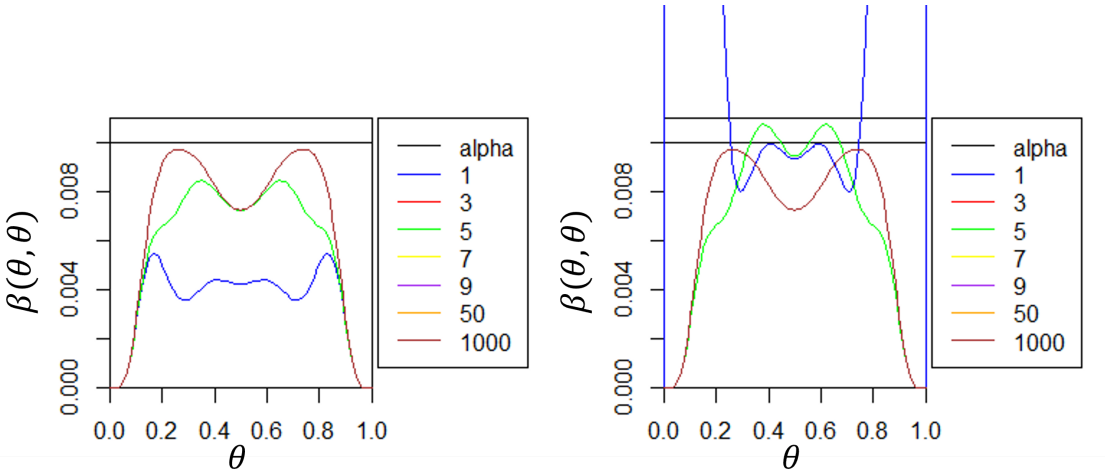


Figure 5.1: $\beta(\theta, \theta)$ for variable N_o and fixed N_p (left) and for variable N_o, N_p (right) with $\alpha = 0.01$.

tests we only use the grid once, to make sure that the constraint $\sum_{j=1}^{\omega} w_j P(\mathbf{x}^j; \boldsymbol{\theta}) \leq \alpha$ is satisfied for each grid point $\boldsymbol{\theta}$. We again consider the outcome space of 2×2 tables with sample size $(n_{1\cdot}, n_{2\cdot}) = (20, 20)$, but now only perform one series of tests with grid size $N \in \{1, 3, 5, 7, 9, 50\}$ and compare their power functions to that of the test with $N = 1000$, for the significance level of $\alpha = 0.01$. The respective size functions are given in Figure 5.4, and the power functions in Figure 5.15. We get similar results; for very small values of N (in fact $N \leq 7$), the constraint on the critical region is not enforced in enough values of $\boldsymbol{\theta}$, leading to critical regions that clearly entail a Type I error probability larger than α for many values of $\boldsymbol{\theta}$. However, from 9 grid points onward, the critical region is the same. This can again be seen from the overlapping size curves for $N \in \{9, 50, 1000\}$ in Figure 5.4, and by the zero difference plot in the bottom-right corner of Figure 5.15. The motivation for this long-winded example is essentially to say that even though the grid size might be too small for the actual problem dimensions to capture the actual CSM ordering or LP critical region, we still create an approximating ordering/region which performs relatively comparable to the actual ordering/region in terms of power. Table 5.1 does not tell the whole story as it only indicates the smallest grid size for which the obtained ordering/region is *exactly* the same as the benchmark ordering/region. The smaller grid size might yield the same ordering, up to one swap of two symmetry classes, or region, up to the inclusion of one symmetry class. This is already "punished" as a difference in Table 5.1, but it is definitely not necessarily the case that this slightly different ordering results in a less powerful test. This would only be the case if the swap in the ordering "unluckily" happens to occur around the chosen significance level, such that one set of tables would be part of the critical region in one ordering, but not anymore in the swapped ordering.

It is only when we go towards very small grid sizes (in the example below 7 grid points for the CSM test and below 9 grid points for the LP test), that we end up with dramatically different results. Based on the (empirical) observations made in this Section, we will from now on always set the grid sizes for the CSM test to $N_o = 10$ and $N_f = 100$, and for the LP test to $N = 100$, unless specified otherwise. Having run many simulations for different table and sample sizes, we

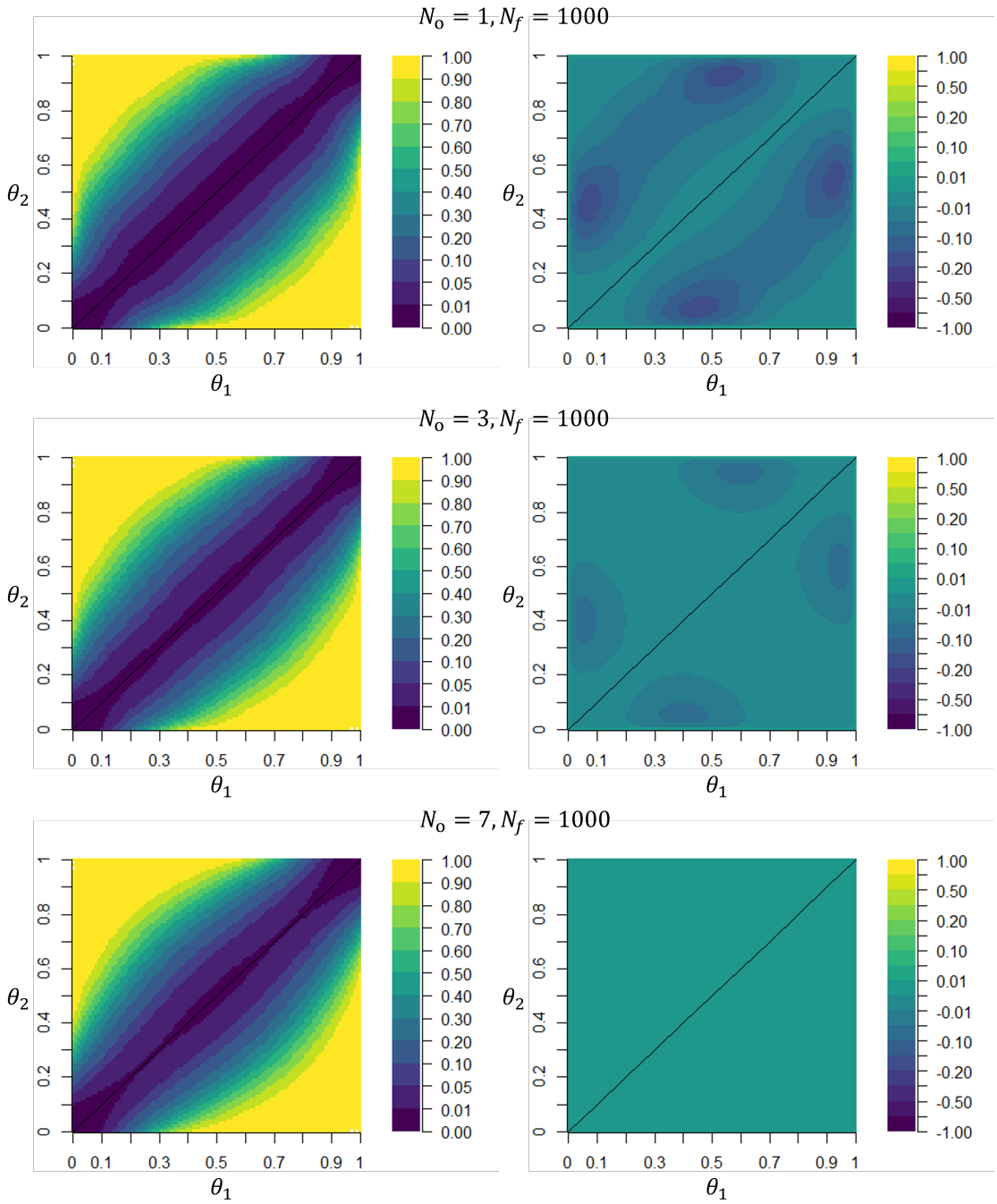


Figure 5.2: $\beta(\theta_1, \theta_2)$ for indicated values of N_o (left) and the difference $\beta(\theta_1, \theta_2) - \beta_{\text{benchmark}}(\theta_1, \theta_2)$ (right).

feel that this number of grid points is, until certain table dimensions and sample sizes, neither too large that the computations will take too long, nor too small that it allows any violations of the upper bound on the Type I error probability. We will however also encounter table and group sizes for which the aforementioned grid sizes will no longer be sufficient, as the test size constraint has been violated. We will indicate whenever this happened, and will also mention the (larger) grid size we used instead in order to again satisfy the test size constraint.

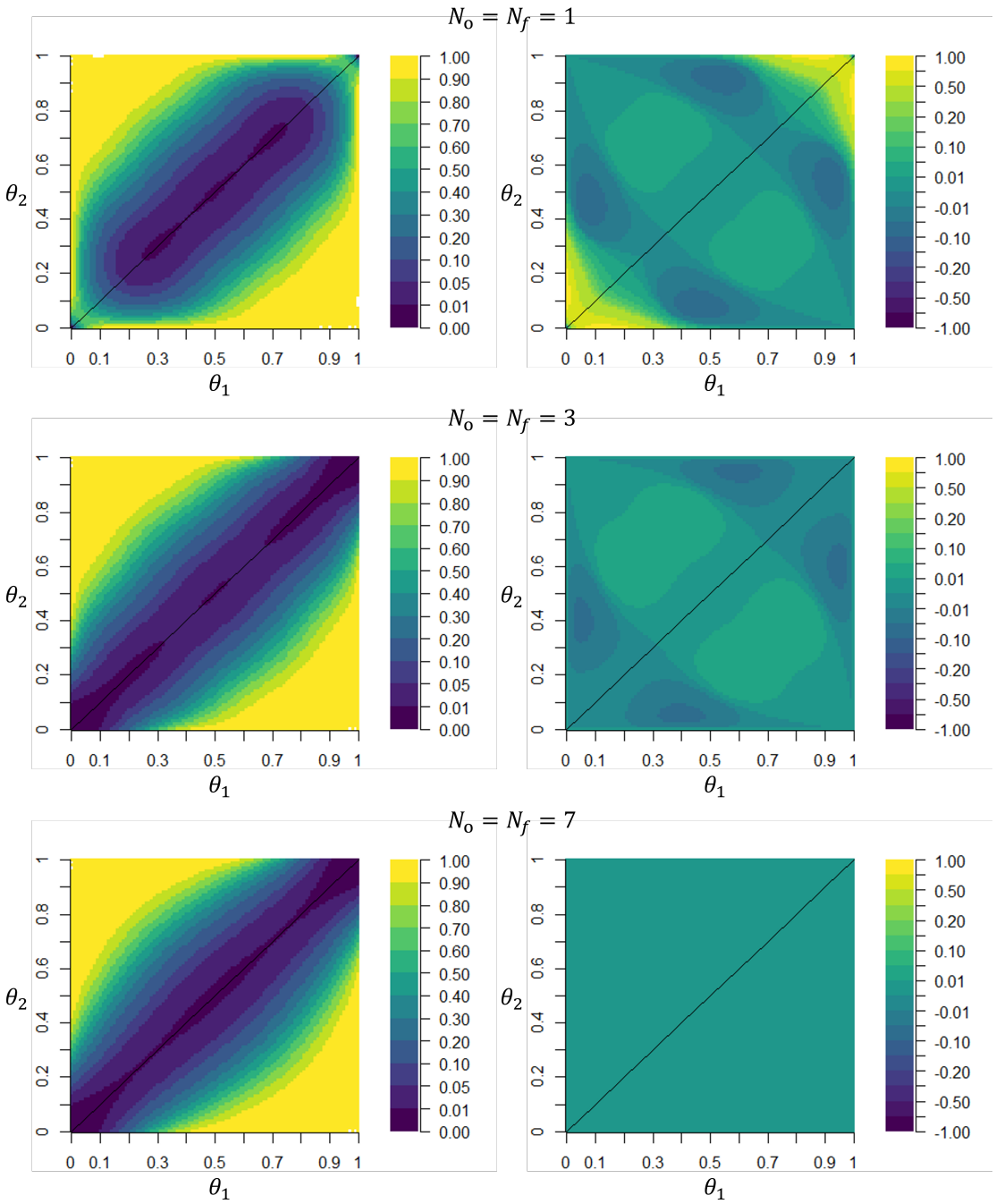


Figure 5.3: $\beta(\theta_1, \theta_2)$ for indicated values of N_o and N_f (left) and the difference $\beta(\theta_1, \theta_2) - \beta_{\text{benchmark}}(\theta_1, \theta_2)$ (column).

5.2 Cutting down on the number of LP tests

As one might have already noticed, the number of different LP tests one can perform is rather large. We mentioned four linear programming formulations, each of which can be performed with one of three symmetry conditions. It would not make sense to consider all twelve variations in the large speed and power comparisons that will follow in Sections 5.3 and 5.4. Therefore, let us preemptively eliminate a few of these tests from further consideration.

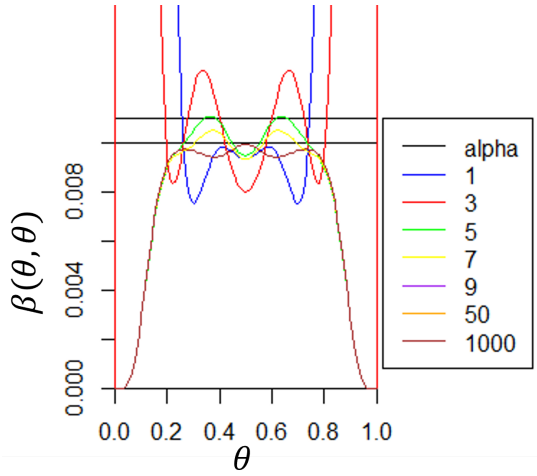


Figure 5.4: $\beta(\theta, \theta)$ for variable N with $\alpha = 0.01$.

5.2.1 The "maximin" test based on (4.22)

First of all, the tests based on (4.22) led to some odd behaviour. Consider the tables with $(n_{1\cdot}, n_{2\cdot}) = (20, 20)$. The critical region corresponding to $\alpha = 0.05$ we find for the test using (4.22) (which we will call test A) is shown in Figure 5.5 on the left. On the right, we find the critical region for the same level, but instead for the test using (4.20) (which we will call test B). Tables that are included in the critical region are marked in black.

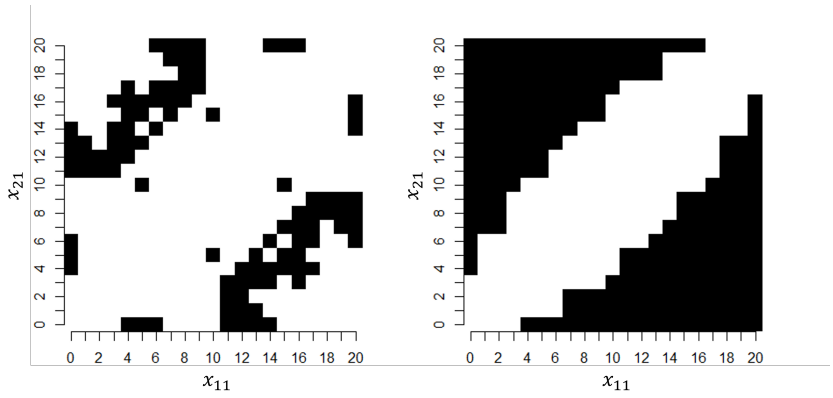


Figure 5.5: Critical regions for test A (left) and B (right).

Already, the critical region on the left does not look very intuitive; it does not include many of the most extreme tables in the top-left and bottom-right corners of the figure. This is a big difference with the critical region of test B, which in fact entirely includes the critical region in the left plot. The corresponding size functions $\beta(\theta, \theta)$ can be seen in Figure 5.6. The blue curve corresponds to test A, and the red curve to test B. Again, it seems that the blue curve could have easily been a lot closer to the black $\alpha = 0.05$ line (without violating the size constraint) by just adding some of the most extreme tables.

Finally, the power functions for the two tests are plotted in Figure 5.7. In the left plot, the power $\beta_A(\theta_1, \theta_2)$ in Θ_1 of test A can be seen. This shows that test A is really not that powerful when compared to the middle plot, which shows the power function of test B. Of course, this could have already been realised from

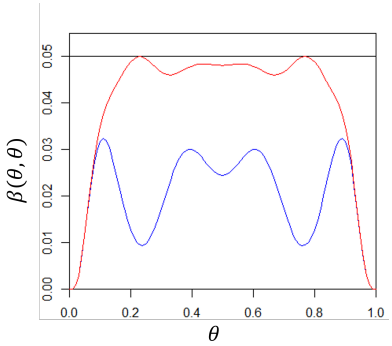


Figure 5.6: Size functions $\beta_A(\theta,\theta)$ (blue) and $\beta_B(\theta,\theta)$ (red).

the fact that the critical region for test A is a subset of the critical region for test B. The difference $\beta_A(\theta_1,\theta_2) - \beta_B(\theta_1,\theta_2)$ is shown in the right plot, and is negative everywhere.

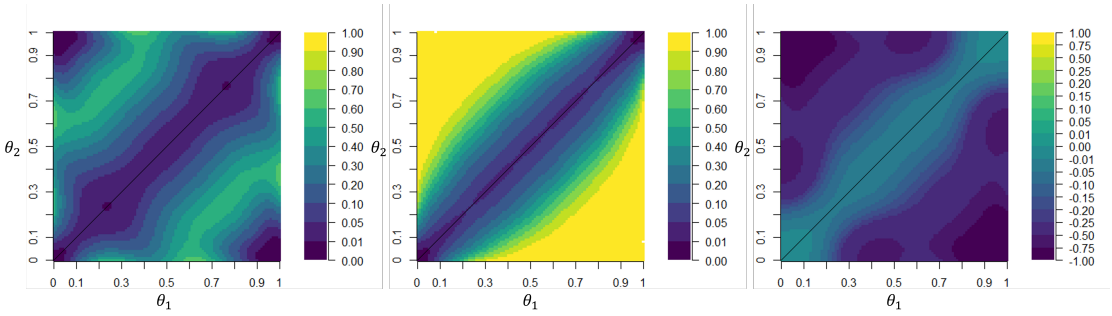


Figure 5.7: Power functions $\beta_A(\theta_1,\theta_2)$ (left), $\beta_B(\theta_1,\theta_2)$ (middle) and the difference $\beta_A(\theta_1,\theta_2) - \beta_B(\theta_1,\theta_2)$.

The obvious question that arises from looking at the critical regions and the corresponding power functions, is why the critical region does not include any of the "extreme" tables in the top-left and bottom-right corners of Figure 5.5. We suspect that this is due to numerical errors that arise when solving the linear programming problem (4.22). Again, the matrix A in (4.22) will contain a number of rows with a lot of small entries, corresponding to the $\boldsymbol{\theta}$-values close to the boundary of Θ_0. Although the entirety of the matrix A will be used to check the size constraint, only these few small-entry rows will be relevant to determine the value of the objective function. Round-off errors will therefore have a large effect on the final shape of the critical region. This is less of an issue in, say, the test based on (4.20). The same matrix A, with the same small entries, also appears there, but now these small-entry rows are not involved in determining the value of the objective function. This hypothesis is confirmed by instead solving, for example, the linear programming problem

$$\begin{aligned} &\text{maximise } z \\ &\text{subject to } z = \min\{v_i : i = 1,\dots,N\}, \\ &\qquad\qquad \mathbf{v} \leq M \cdot A\mathbf{w}, \\ &\qquad\qquad A\mathbf{w} \leq \alpha\mathbf{1}, \\ &\qquad\qquad \mathbf{w} \in \{0,1\}^{\omega},\ \mathbf{v} \in \mathbb{R}^N_{\geq 0}, \end{aligned}$$

where M is a large number. This will rescale the entries of $\mathbf{v}$ and thus of z too. If we execute test A again with this slightly altered LP formulation for increasing values of $M \in \{10^3, 10^4, 10^5\}$, we indeed see in the first column of Figure 5.8 that the critical region of test A takes a more "expected" shape. As M increases, the blue size function (second column) also seems to come closer to the $\alpha = 0.05$ line. Note however how the solver favours the inclusion of tables of which the $P(\mathbf{x}; \boldsymbol{\theta})$-function takes large values near $\boldsymbol{\theta} = (1, 0)$ and $\boldsymbol{\theta} = (0, 1)$, leaving a big gap in the size for $M = 10^3$ and $M = 10^4$. Even for $M = 10^5$ a small gap remains. However, the power $\beta_A(\theta_1, \theta_2)$ (third column) seems to get a lot closer to that of test B, as one can see from the shrinking difference $\beta_A(\theta_1, \theta_2) - \beta_B(\theta_1, \theta_2)$ (fourth column).

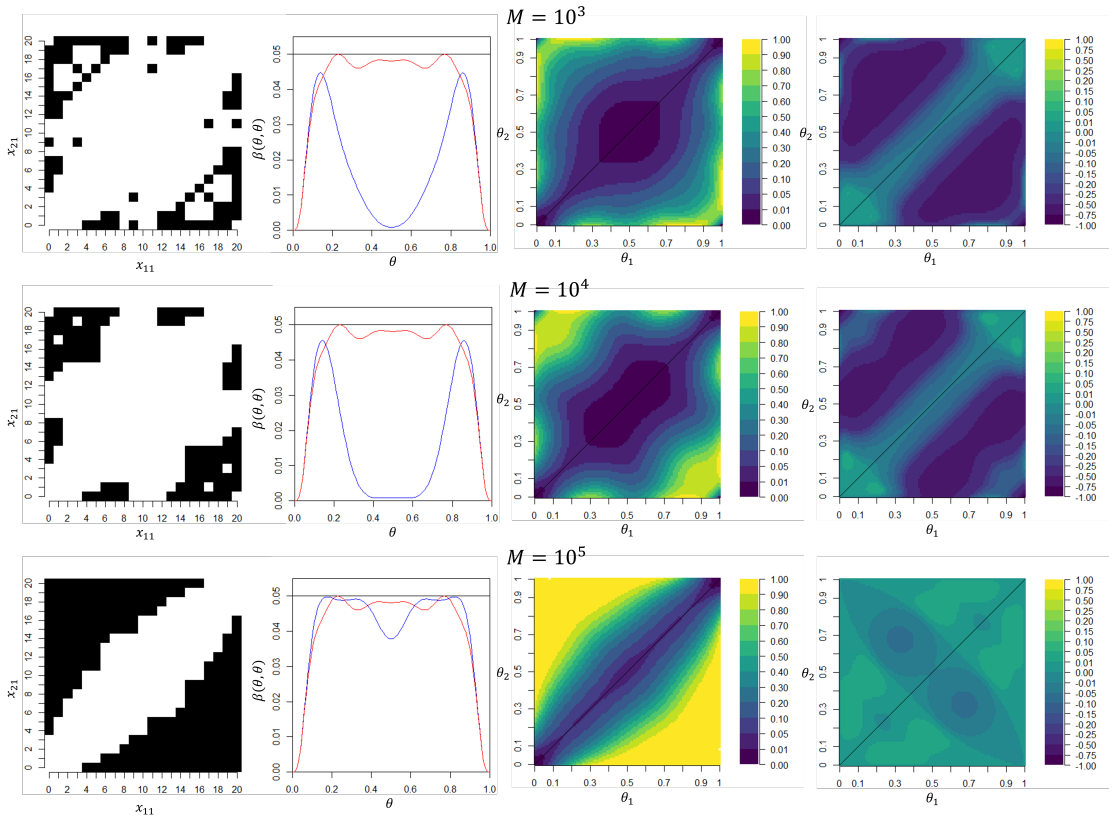


Figure 5.8: Critical regions (column 1), $\beta_A(\theta, \theta)$ (column 2), $\beta_A(\theta_1, \theta_2)$ (column 3), and $\beta_A(\theta_1, \theta_2) - \beta_B(\theta_1, \theta_2)$ (column 4) for the indicated values of M.

The smallest value of this constant M for which we obtain an "acceptable" critical region will dependent on the group sizes of the table. Indeed, the larger the group sizes, the smaller the values of $P(\cdot; \boldsymbol{\theta})$ (and hence the corresponding entries of A) will be for certain tables and values of $\boldsymbol{\theta}$. Just as we did with the question whether to consider the k-th smallest entry in $\mathbf{v}$ instead of the minimum, we will label the problem of determining the "smallest acceptable" value of M as future research. Because of the quirky behaviour of test A, i.e., the test based on (4.22), we will not take it into account for the large speed and power comparisons.

5.2.2 The maximal area test based on (4.23)

Below (4.23), we already foreshadowed that the test based on this LP formulation, which we will call test C in this Section, did not behave entirely as expected.

Our hope at first was that by maximising the area of the power function in Θ_0, we might also get a critical region that has high power in Θ_1 "for free". As the following example will show however, this is not the case.

Consider again the tables with $(n_{1\cdot}, n_{2\cdot}) = (20, 20)$. We will compare the performance of test C to that of the test where we instead use (4.24), i.e., maximise the integral of the power over Θ_1. This test will be referred to as test D. In Figure 5.9, the critical regions of tests C (left) and D (right) are compared. Note the odd shape of the critical region of test C: some tables which we would deem as "more extreme", such as $(x_{11}, x_{21}) = (14, 0)$, are not included, while other "less extreme" tables, such as $(x_{11}, x_{21}) = (4, 0)$, are included. When looking at the size

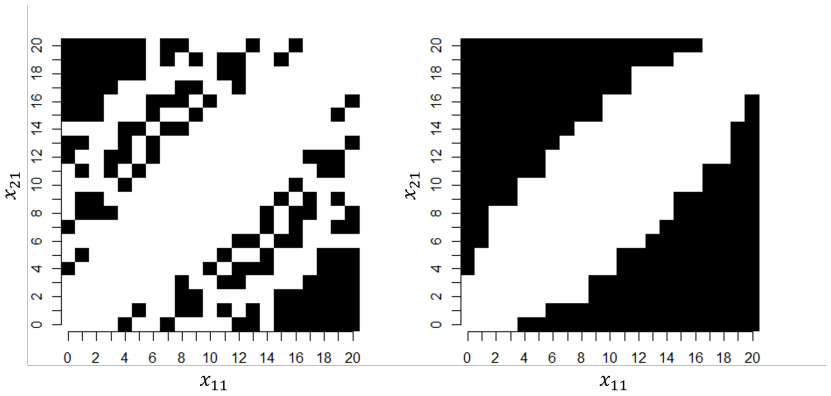


Figure 5.9: Critical regions for test C (left) and D (right).

$\beta(\theta, \theta)$ for both tests, we clearly see in Figure 5.10 that test C tried to make the area under the blue curve as large as possible. $\beta_C(\theta, \theta)$ manages to get a lot closer to α for many values of θ compared to $\beta_D(\theta, \theta)$.

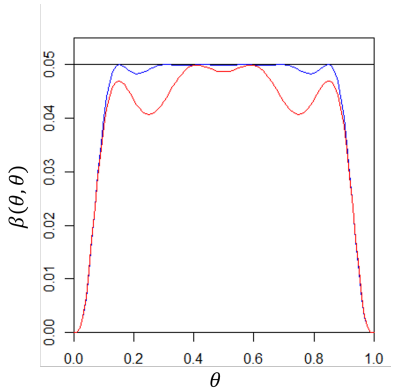


Figure 5.10: Size functions $\beta_C(\theta, \theta)$ (blue) and $\beta_D(\theta, \theta)$ (red).

However, this gain in size comes at a price. When comparing $\beta_C(\theta_1, \theta_2)$ to $\beta_D(\theta_1, \theta_2)$ on the whole of Θ, we observe in Figure 5.11 that test C is less powerful.

Test C looks for the set of tables that makes $\beta(\theta, \theta)$ as close to α as possible for as many θ as possible, while completely neglecting the power of the implied critical region in Θ_1. It is definitely worthwhile to give in a bit on the test size, if that means achieving a far greater power in Θ_1. We have observed this behaviour not only for this specific example, and have therefore opted to not consider test

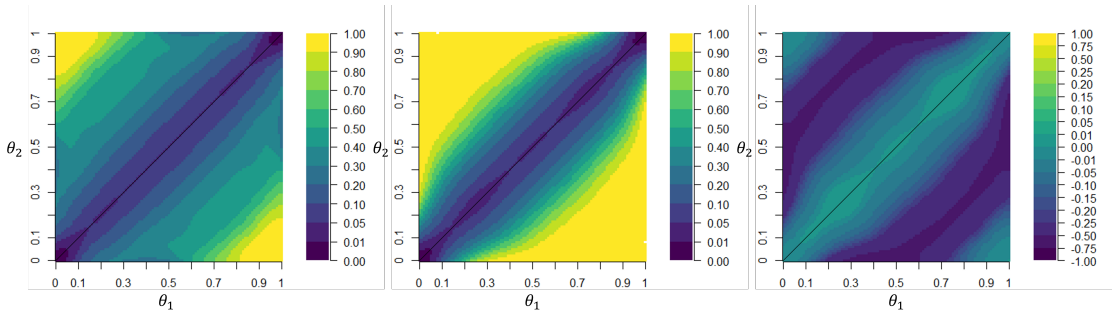


Figure 5.11: Power functions $\beta_C(\theta_1, \theta_2)$ (left), $\beta_D(\theta_1, \theta_2)$ (middle) and the difference $\beta_C(\theta_1, \theta_2) - \beta_D(\theta_1, \theta_2)$.

C, i.e., the test based on (4.23), in the rest of this chapter. For ease of notation, we will now always refer to the test based on (4.20) as test "LP1", and to the test based on (4.24) as test "LP2".

5.3 Speed

The larger the tables we are working with, the longer it will take to compute the different tests. In particular, the computation time for the unconditional tests will grow quickly with the table dimensions, as these tests need to consider the entire space of table outcomes. One might have already spotted the gaps in Table 5.1, indicating that for the corresponding table and sample sizes, the computations took longer than we were willing to let it run for. In this section, we will be particularly interested in how fast the different tests are compared to each other.

5.3.1 Preliminary computations for unconditional tests

For the unconditional tests, it is necessary to first construct a list of all possible table outcomes, and to find out how to divide these outcomes into equivalence classes according to the desired (symmetry) criterion. It is important to keep in mind that unconditional tests require this extra computation because they somehow need to determine, for an observed outcome, which tables are more extreme. In the case of CSM-like tests, this actually means that we compute the p-values for each of the more extreme outcomes before being able to compute the p-value of the table we are actually interested in. For tests using an external test statistic, we first need to compute the value of that test statistic for the more extreme outcomes. This extra work certainly slows down the unconditional tests when using them for their intended purpose; deciding on whether or not to reject the null hypothesis for an observed outcome. However, as we will see in Section 5.4, this does imply that we will only need to run a CSM-like test once in order to obtain the p-values for all table outcomes.

Recall that the total amount of table outcomes of a contingency table with c columns and sample sizes $(n_{i\cdot})_{i=1,\dots,r}$ is given by (4.7). The time it takes to construct the list of all possible outcomes should therefore scale as this function of the table and group sizes. A more interesting question is however to find out how long it will take to divide the set of outcomes into equivalence classes. In particular, how does this computation time depend on the table dimensions and group

sizes? This is a combinatorial exercise which depends on the type of symmetry condition we apply. However, with some justified simplifications, we can come up with pretty reasonable estimates for this computation time. First of all, realise that the procedure to construct the equivalence classes is the same regardless of which symmetry condition is applied. Given a possible table outcome, we will loop through the list of all possible outcomes and keep track of the ones which are a symmetric counterpart. After we have gone through the entire list, we will remove all these symmetric counterparts, as well as the given outcome, from the list of tables to consider and repeat the procedure, until there are no more tables left to consider.

To this end, let us consider the setting of 2×2 tables. For a table where $n_{1\cdot} = n_{2\cdot} =: n$, it is possible to obtain symmetry classes of size 1, 2 or 4 tables. If a table has no other symmetric counterparts, that means that it is invariant under switching rows or columns, i.e. all table entries should be the same. Thus, if n is even, the table $(n/2, n/2)$ forms its own symmetry class. If n is odd there are no such singleton symmetry classes. Moving on, the only way in which we can have a table with only one symmetric counterpart is if swapping the rows does not change the table, or if swapping the rows yields the same table as swapping the columns. That is, $x_{11} = x_{21}$ or $x_{11} = x_{22}$ respectively. As $x_{11} \in \{0, \dots, n\}$, there are $n + 1$ possible ways to achieve the former. However, if n is even, this also includes $(n/2, n/2)$, so to prevent double counting there are in fact n possible ways to have $x_{11} = x_{21}$. Similarly, if n is odd, we can see that there are n ways to choose $x_{11} = x_{22}$. Thus, if n is even, there are $2n$ tables which belong to a class of 2 tables, so there are n classes of size 2. If n is odd, there are $2(n + 1)$ such tables, so $n + 1$ classes of size 2. The remaining tables thus all belong to a class of size 4; switching rows and/or columns each time yields a different table. By (4.7) with $r = c = 2$, there are $(n + 1)^2$ tables in total, so if n is even there are $(n + 1)^2 - 1 - 2n = n^2$ tables which belong to a class of size 4. If n is odd there are $(n + 1)^2 - 2(n + 1) = n^2 - 1$ such tables. Thus, a total of $n^2/4$ and $(n^2 - 1)/4$ of such classes, respectively.

From this, we can easily see the total amount of symmetry classes G as a function of the common group size n:

$$G(n) = \begin{cases} \frac{1}{4}n^2 + n + 1, & n \text{ even}, \\ \frac{1}{4}n^2 + n + \frac{3}{4}, & n \text{ odd}, \end{cases} \tag{5.3}$$

and also compute the average size of a symmetry class $m(n)$:

$$m(n) = \frac{\omega}{G(n)} = \begin{cases} \frac{n^2+2n+1}{\frac{1}{4}n^2+n+1}, & n \text{ even}, \\ \frac{n^2+2n+1}{\frac{1}{4}n^2+n+\frac{3}{4}}, & n \text{ odd}. \end{cases}, \tag{5.4}$$

We can now come up with a rough estimate for the total amount of searches we need to perform in order to find all the symmetry classes. When we start, we have to go through all ω tables of the outcome space. We will then find a symmetry class, which on average has size m. Therefore, for the search of the next class, we will only need to go through $\omega - m$ tables. Afterwards, for the next one, only through $\omega - 2m$ and so on. Thus, if we let for simplicity ω be a multiple of m, i.e.

$\omega = km$, the total number of searches N will be approximately

$$\begin{aligned} N &= \omega + (\omega - m) + (\omega - 2m) + \ldots + 2m + m \\ &= m(k + (k-1) + \ldots + 2 + 1) \\ &= m\frac{k(k+1)}{2} \\ &= \frac{1}{2m}\omega^2 + \frac{1}{2}\omega. \end{aligned} \tag{5.5}$$

Note that the approximation here is made by setting all group sizes to m. If we would remain exact there would be different group sizes, and then the order in which we go through the groups will affect the sum in (5.5). For larger tables, the above computation becomes quite a lot more involved so we did not come up with an analytic expression there. However, we find it reasonable to believe that the total number of searches can still be approximated with some kind of arithmetic sequence as in (5.5), albeit with a different number value for m. The number of different group sizes will inevitably become larger, so the approximation will perform more and more poorly. Nevertheless, we can assess how well a quadratic function of the form $a\omega^2 + b\omega$ would fit to measurements of the time it takes to find all symmetry classes for given table and group sizes. This is shown in Figure 5.12. In all three graphs, we plot the computation time of finding all symmetry classes as a function of the number of tables. The left plot corresponds to 2×2 tables with $n_{1\cdot} = n_{2\cdot} = n \in \{5k \colon k = 1, \ldots, 30\}$, the middle one with 3×2 tables where $n_{1\cdot} = n_{2\cdot} = n_{3\cdot} = n \in \{5k \colon k = 1, \ldots, 10\}$, and the right one with 2×3 tables where $n_{1\cdot} = n_{2\cdot} = n \in \{5k \colon k = 1, \ldots, 6\}$.

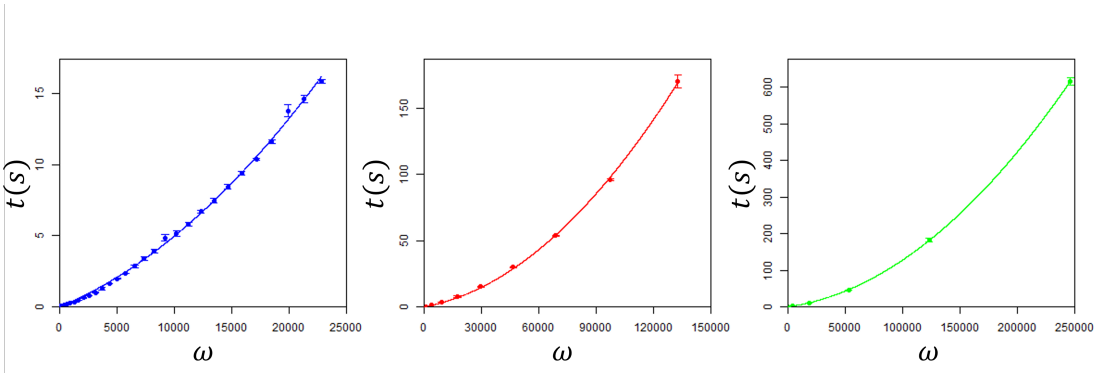


Figure 5.12: Runtime to find the symmetry classes as a function of the number of tables ω for indicated table dimensions.

The adjusted R^2 values of the fits are (from left to right) 0.9992, 0.9998, and 1. However, according to our rough estimate, the amount of searches (and so also roughly the computation time) should only depend on the amount of tables ω. That is, there should be no dependence on the table dimensions. If we were to put all the measurements into one plot, they should all lie approximately on one parabola. In Figure 5.13, we have done exactly that. We have also added a log-log plot for clarity. The colours correspond to those in Figure 5.12. Furthermore, we have added a quadratic regression line considering all measurements (with an adjusted R^2 of 0.9943).

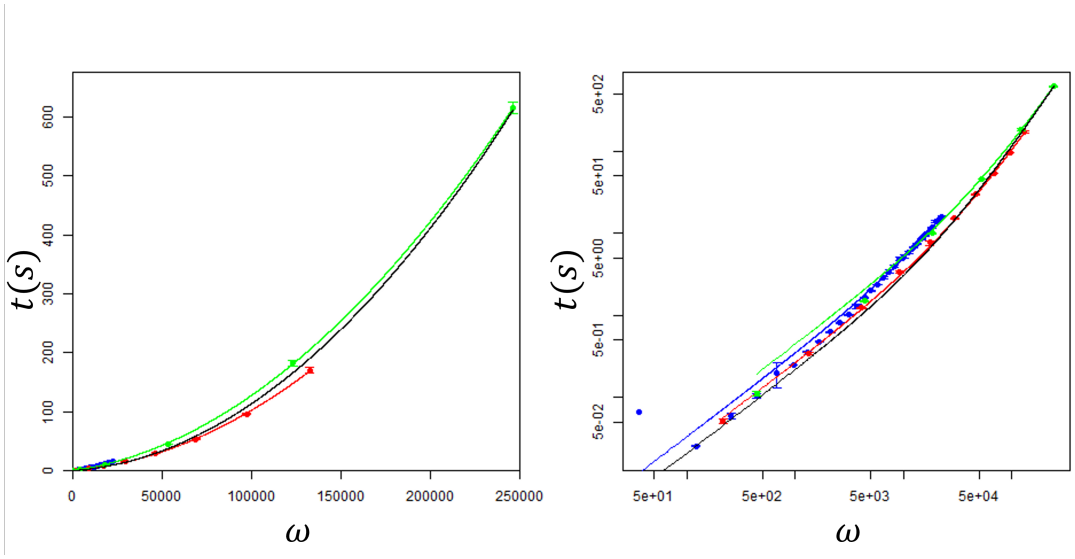


Figure 5.13: Computation time to find the symmetry classes as a function of the number of tables ω. The corresponding log-log plot can be found on the right.

The quadratic fit over all measurements is not as good as the individual fits of course (adjusted R^2 of 0.9932), however a visual inspection leads us to believe that our rough estimate of quadratic behaviour should serve as a good indication on how the time it takes to find all the equivalence classes should scale with the table and group sizes. The deviations that we observe might simply be a sign of measurement errors, or maybe some small effects of the table dimensions after all. For example, the time it takes to check whether a table is belongs to the same equivalence class as another table will depend on the table dimensions.

5.3.2 The effect of the grid size on the computation time

Before moving on to the main speed comparison, the computation time of the supremum and LP tests will inevitably depend on the chosen grid size N. Although we have fixed the grid size to $N = 100$ in the previous Section already, it is still worth briefly mentioning how the computation time grows with N. For example, for the CSM test, at each iterations, the amount of operations to perform is directly proportional to the amount of grid points we consider. Therefore, we expect the computation time to grow linearly in N. The tests using an external test statistic only need to perform this maximisation once to find the p-value of the table of interest. For the LP test, the grid size N comes back as the amount of rows of the matrix A in (4.20) and is thus the amount of inequalities we need to check for each proposed solution $\mathbf{w}$. This we also expect to scale linearly.

Let us consider the outcome space of tables with $(n_{1\cdot}, n_{2\cdot}) = (10, 10)$. We will measure for the CSM test the time it takes to compute the p-values of all tables in the outcome space, and for the LP test the time it takes to determine the critical region $K^{0.05}$. We will not include the time it takes to list all table outcomes and construct the equivalence classes, as both tests perform those steps anyway. Each time measurement will be repeated 10 times and then we will work with the average. One should not compare the computation times of the CSM and LP tests, as we end up with different amounts of information after performing both tests.

For now, we are only interested in the effect of N on the computation time. A fair comparison of the two tests will follow in Section 5.3.3. In Figure 5.14, we show the computation time as a function of the grid size $N \in \{10k : k = 1, \ldots, 10\}$. In the left plot, two sets of points can be seen. The blue points represent the CSM test durations where we vary $N_o = N$ and set $N_f = 100$ fixed set. The red points represent the CSM test durations where we vary both $N_o = N_f = N$. The dashed lines are the linear regression lines. The respective adjusted R^2-values of 0.9993 and 0.9997 for the blue and red lines confirm our earlier expectations of linear growth in N. Note that for $N = 100$ both lines correspond to the same CSM test with $N_o = N_f = 100$. The right plot shows the time durations of the LP test as a function of the grid size N. Again, we see a linear relation between N and the computation time ($R^2 = 0.9992$). The standard deviation over the 10 measurements is indicated with the error bars in both plots.

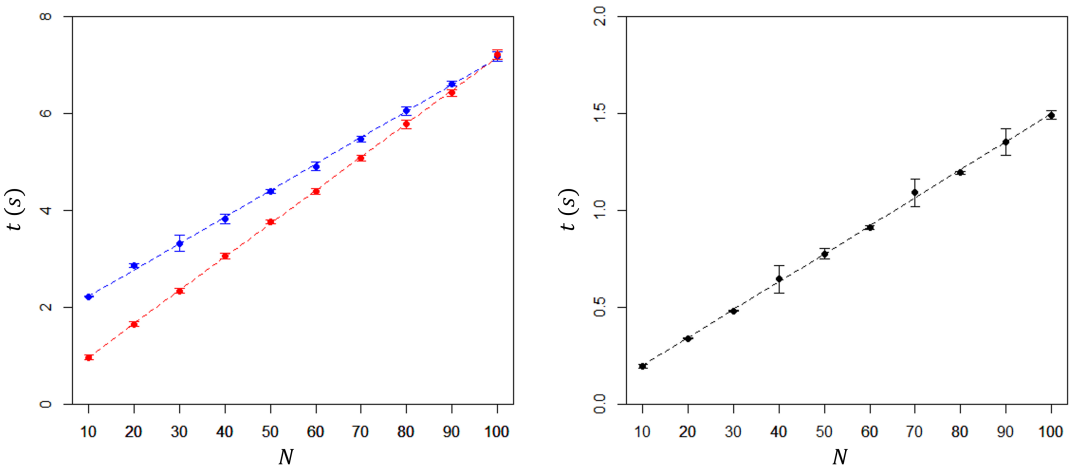


Figure 5.14: Computation time as a function of the grid size N for the CSM test with fixed N_f (left, blue), the CSM test without fixed N_f (left, red) and for the LP test (right).

5.3.3 Speed comparison for different table and group sizes

Finally, let us compare how quick the actual tests of significance are. In order to perform a fair comparison, there are a few things we should keep in mind. First of all, because the unconditional tests require the computation of the equivalence classes to reach a result, we should include that time into the time the actual test takes too. Second, our goal here will be to measure how long it takes for the tests to do what they are intended to do. That is, for an observed outcome, returning a p-value. To this end, we will proceed as follows. Given the group sizes $(n_{1\cdot}, \ldots, n_{r\cdot})$, we will perform the test for each table in a random selection of M tables from the outcome space Ω and measure how long it takes. This will return M time measurements. For CSM-like tests, we expect a wide range of test durations; tables which are more extreme will come earlier in the ordering, and we will be able to stop the computations whenever we have reached that table. Tables which

are not extreme will come late in the ordering, meaning that the CSM algorithm will have already worked through most of the outcome space before returning a result. For the other tests, we expect the test durations to be more or less the same for all outcomes. Also, keep in mind that we also want to end up with a p-value when performing the LP tests, so we will make use of Algorithm 2 to find an approximate p-value. We will thus need to solve the linear programming problem multiple times. Finally, all computations that involve some measurement of time should be performed on the same hardware for fair comparison. In this case, all time measurements made in the remainder of Section 5.3 were made on an Acer Travelmate P214-53-72EQ laptop with an 11^{th} generation Intel i7-1165G7 processor with a frequency of 2.80GHz, 16GB of RAM and running Windows 11 Pro version 22H2.

Even with all these precautions to ensure a comparison that is as fair as possible, one should keep in mind that the LP tests use the highly efficient software written by a private company to actually solve the linear programming problem, whereas the remaining tests (in particular the CSM tests) have to settle with the `R` implementation of the author. It remains an open question how fast and efficient one can make the implementation of the supremum tests. This makes the speed comparison of limited use, but it is the best we can do, as it is difficult to come up with any meaningful theoretical estimates in order to compare the computation times of the different tests. Of course, for the supremum tests, we can think of some rough estimates. For example, the tests using an external test statistic first need to compute the value of the test statistic for each table (ω computations of the test statistic). Next, these tests should sort the tables based on the values of the test statistic (this can be done in $\mathcal{O}(\omega)$). Finally, we should perform, based on the ordering, the maximisation procedure once to find the p-values for the table of interest. This amounts to computing and comparing N sums of $P(\cdot;\boldsymbol{\theta})$-functions, where N is the number of grid points. The number of terms in these sums will depend the position of the table of interest in the ordering. Hence the computation time for the tests using an external test statistic should scale linearly in ω.

For the CSM-like tests, we have a slightly different situation. There, we perform a set of maximisations at each iteration of the algorithm, in order to find out which candidate outcome has the smallest maximum value. Recall that we might use these smallest maximum values as the p-values, or alternatively we could, after determining the ordering, run the maximisation procedure again with a finer grid just as if the ordering was obtained by an external test statistic. However, how many maximisations are performed to determine the ordering? Note that if we would not include the C condition, we would perform ω maximisations at the first iteration, on average $\omega - m$ at the second one (where m is the average size of the symmetry classes we use), $\omega - 2m$ in the next one and so on, which yields a number of maximisations that scales with ω^2, as shown in (5.5). Including the C condition would lower this number substantially, but giving an estimate for general tables becomes a more involved task. Likewise, it is hard to give an indication of the speed of the LP tests. In the worst case, the branch-and-bound algorithm can be exponential in the number of integer constraints (in this case the number of grid points), if we would need to go through the entire tree of sub-problems. However, this is rarely the case and the actual complexity of dependent on the exact problem at hand. The best we can do for now is thus our speed comparison

described earlier.

With a fixed grid size, the computation times will only depend on the table dimensions and group sizes. We will consider 14 tests in total. These are Pearson's chi square test (denoted `PEARSON`), Fisher's exact test (denoted `FISHER`), 3 CSM-like tests (denoted `C <S> M` depending on the symmetry condition used), 3 supremum tests using an external test statistic (denoted `ET <T>` depending on the chosen test statistic), and 6 LP tests. The latter are denoted `LP<1,2> <S>`, indicating whether we are considering an LP1 test or an LP2 test, and specifying one of the three symmetry conditions. For each test we will perform two series of measurements. In the first series, we will fix the table dimensions to $r = c = 2$ and measure the runtimes for group sizes $(n_{1\cdot}, n_{2\cdot}) = (n, n)$ with n running from 5 to 25 in increments of 5. The second series will always consider groups of sizes equal to 5, but with varying table dimensions $(r, c) \in \{(2,2), (3,2), (4,2), (2,3)\}$.

In Figure 5.16, we show the results of the first series. For each test, and for each value of n, a boxplot of time measurements for $M = 30$ randomly selected tables in the corresponding outcome spaces. It seems that the asymptotic and conditional tests are not so much dependent on the group size. Of course, because the existing implementations of these tests are already so efficient, we did not expect to find any significant effect for the relatively small group sizes we are considering here. Moving on to the unconditional tests however, we clearly see a superlinear increase in runtime. This is to be expected, as the problem sizes increases at least at the same rate as the amount of symmetry classes, which we conjectured to grow quadratically in the amount of tables, which itself is a superlinear function of the group sizes, as can be seen in (4.7).

A perhaps more interesting result can be found by using the timings we just obtained to compare the speed of the tests relative to each other. In Figure 5.17, we have shown boxplots of the runtimes of each test, for fixed values of n. Pearson's asymptotic test and Fisher's exact test are clearly the quickest test, followed by the LP tests. The tests using external test statistics are slightly slower, and as expected, the three CSM tests can either be very quick, if the table we want to know the p-value of is extreme, or very slow, if we first need to order a large portion of the outcome space before we arrive at the table of interest. The runtime of the other tests is not that dependent on the table under consideration, hence the rather squeezed boxplots. Note that in this comparison, the 12 unconditional tests all include the time it took to do the required preliminary computations, i.e. constructing the relevant symmetry classes. Among the different test categories themselves, there does not seem to be much difference in runtime. Interestingly, even though the S_χ symmetry condition splits up the outcome space into fewer symmetry classes, this does not result in any considerable speedup.

Finally, the measurements for varying table dimensions and fixed group sizes are shown in Figure 5.18. A similar conclusion to the one from the 2×2 comparison can be drawn. The asymptotic and conditional tests take by far the least amount of time. Although the CSM tests seem to be quicker for a small number of tables, these are "lucky" observations in the sense that we executed the test on a relatively extreme table which arrived early in the ordering. Apart from these few exceptions, the LP tests again have the smallest runtime out of the unconditional tests, followed by the tests using an external test statistic. Of course, one can question the quality of the p-values obtained by the binary search in Algorithm 1. However, the

formulation in terms of a linear programming problem and the opportunity this brings to use efficient optimisation software, suddenly make unconditional tests a lot quicker (although nowhere near the speeds of the asymptotic and conditional tests yet). Again, the non-LP tests cannot use the fast LP solver software. A more efficient implementation of the supremum methods might be able to achieve comparable runtimes to those of the LP tests. A recommendation for further research would thus be to find good theoretical estimates on the runtimes of both the CSM and LP tests, try to optimise the CSM implementation as much as possible, and repeat the speed comparison in this more fair setting.

5.4 Size and power

Making use of (5.1), we are able to compute the size/power of the tests we discussed at any $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_r) \in \Theta$, for any given $\alpha \in [0, 1]$. In the case of tests on $r \times 2$ tables, Θ_0 is one-dimensional and so we are able to graphically visualise the size of the test for $\boldsymbol{\theta} = (\theta, 1 - \theta) \in \Theta_0$. Also, we would able to visualise the power on a colour plot for the 2×2 case as Θ_1 is two-dimensional, just as we did in Figure 5.2. The same holds true for the size of the tests on $r \times 3$ tables, as in this case $\Theta_0 = \{(\theta_1, \theta_2, 1 - \theta_1 - \theta_2) : \theta_1, \theta_2 \geq 0, \theta_1 + \theta_2 \leq 1\}$ is also two-dimensional. However, for larger tables we are unable to show plots covering the entirety of Θ. We should even ask ourselves whether comparing the power function based on images is such a good idea. Of course, we could just as in Figure 5.2 take the difference between two power functions to see which one is largest, and for which values in Θ. However, we will be looking at the same 14 tests as in the previous Section; Pearson's chi-square test, Fisher's exact test, 3 CSM-like tests, 3 tests using an external test statistic, and 6 LP tests, at different group sizes. Pairwise comparisons will yield a lot of – or rather, too much – plots and pairwise difference plots.

Hence, we should be selective about which results to include. As we are interested in the power function of the 14 aforementioned tests when applied to tables with different dimensions and group sizes, we will consider 2×2, 3×2, 2×3, 3×3, and 2×4 tables. For each of these table dimensions, we will indicate a number of group sizes for which we will then compute the power function on a grid over Θ using (5.1), for the significance levels $\alpha \in \{0.01, 0.05, 0.10\}$. We can then compare, for each choice of group sizes and significance level, the power functions of the different tests to each other in two ways. First, we will look at the power functions on Θ_0, i.e., at the size, as we can plot these as functions of a single parameter θ. Note that this we can only do for the 2×2 and 3×2 tables. Second, we will try to compare the power functions on Θ_1 via an approach used by Mehrorta, Chan and Berger [12]. This we will discuss in more detail once we get there.

Finally, one should realise that this power comparison is of limited significance. As noted by Upton [54], power comparisons are usually done for tests which share the same Type I error. However, because we are in a discrete setting, the significance level α we fixed beforehand will in general not be reached by all tests (and keep in mind that the size is in fact a function of the unknown nuisance parameter). However, in the words of Martín Andrés and Silva Mato [75], "the absence of a reasonable alternative makes the use of this procedure unavoidable".

5.4.1 2×2 tables

Let us start with the 2×2 table. We will consider all 14 tests we discussed at 3 significance levels $\alpha \in \{0.01, 0.05, 0.10\}$ for several sample sizes $(n_{1\cdot}, n_{2\cdot})$, given below

$$\begin{aligned} &(5,5),\ (10,5),\ (10,10),\\ &(20,5),\ (20,10),\ (20,20),\\ &(40,5),\ (40,10),\ (40,20),\ (40,40). \end{aligned} \tag{5.6}$$

Each of these 14 tests will either return a critical region, or an array of p-values from which we can deduce the critical region for the given level. With this, we can then easily use (5.1) to compute the power function on a grid of (θ_1, θ_2)-values. We will compute the power $\beta(\theta_1, \theta_2)$ for all $\theta_1, \theta_2 \in \{k\Delta : k = 0, \ldots, \Delta^{-1}\}$ with $\Delta = 0.01$. This will return a total of $14 \cdot 3 \cdot 10 = 420$ matrices with 101×101 entries representing the power function evaluated on this equidistant discretisation of Θ. Again, we will not show 420 colour plots. However, to get a first idea of how the different tests performed relatively to each other, let us look at the size functions for each combination of α and $(n_{1\cdot}, n_{2\cdot})$. For all combinations of α and $(n_{1\cdot}, n_{2\cdot})$, one can find the plots of the size functions in Figures C.1, C.2, and C.3. We use the same abbreviations for the tests as introduced in Section 5.3.3. The corresponding plots can be found for the LP tests in Figures C.4, C.5, and C.6.

First thing to note is that in general, all unconditional tests have a larger size for all $\theta \in [0, 1]$ than Pearson's and Fisher's tests. Since we use Yates' correction for Pearson's chi square test, we expected it already to behave like Fisher's exact test. The relatively small size of these two tests is in line with the often mentioned argument against conditional testing; that it is less powerful. Since the power function is continuous in $\boldsymbol{\theta}$ (this is easy to see from (5.1)), we would in general expect a test with lower power also to have lower size. For completeness, note that this observation does not always hold. There are a couple of cases in which the sizes of `ET chi` and `ET vol` dip below that of the size of Fisher's exact test. It is not really clear yet whether the gap between the size function of Fisher's test and that of the unconditional tests becomes smaller as the sample sizes increase.

Note of course that, by construction, Boschloo's test (`ET fisher`) is uniformly more powerful than Fisher's exact test and the size function of Boschloo's test always takes a larger value than that of Fisher's. Furthermore, we see for small group sizes (in particular the first rows of Figures C.1 and C.4), that also the unconditional tests have a size far below α, and sometimes even coincide with the size of Fisher's exact test. This is also to be expected, as the outcome space is simply not that large, the $P(\mathbf{x}; \boldsymbol{\theta})$-functions take relatively large values and therefore the supremum methods will not be able to "stack another table" on top of the already ordered tables without surpassing the level α.

Comparing the unconditional tests with each other (not including the LP tests for a moment), we see that especially for small sample sizes, there is a lot of overlap in the size functions. For certain tests, this overlap stays even for large sample sizes. For example, in every plot, the size function of `C S_P M` is exactly the same as that of `C S_V M`. `C S_chi M` sometimes also has the same size function. Overall, it does not seem as if one symmetry condition always yields superior size. We will see in a bit that this conclusion also follows for the power function on Θ_1.

Also when looking at the tests using an external test statistic, there does not seem to be one test statistic which yields superior size in all cases. However, it seems that unless the CSM-like tests and tests using external test statistics overlap, the CSM-like tests have a higher size for most values of θ.

Now only looking at the LP tests, there seems to be lot of overlap for small sample sizes (see Figure C.4). For larger sample sizes, the size functions of the different tests seem to differ more and more. However, also here there does not seem to be a test which clearly has the largest size for all $\theta \in [0, 1]$. In particular, there does not appear to be a big difference between the `LP1` tests and the `LP2` tests. We will see in a bit if this also holds for the power function on Θ_1. Comparing the LP tests to the other unconditional ones, the general observation seems to be that the LP tests always have a size comparable to that of the CSM-like tests.

As we are essentially only interested in the how powerful the tests are compared to each other, we decided to use a similar table representation as the one used by Mehrorta, Chan and Berger [12]. This is shown in Tables C.1, C.2, C.3, and C.4 for $\alpha = 0.01$. We have omitted the tables for other significance levels, as they did not provide much added value. How should one read these tables? All rows and columns are named after one of the tests we are investigating. Consider the cell in row R and column C (note that the table is rotated; we say that Table C.1, for example, has 13 rows and 4 columns). There are a bunch of numbers and signs in this cell. Each pair of one number or sign with one number between parentheses or an empty entry corresponds to one of the $(n_{1\cdot}, n_{2\cdot})$-values in (5.6) (in the same configuration). So for example, in Table C.1 in the cell in row `PEARSON` and column `FISHER`, the two entries "$<$" and "(-0.0590)" correspond to group size $(5, 5)$, and the two pairs of one "$=$" and an empty entry correspond to the group sizes $(10, 10)$ and $(40, 40)$.

What do these numbers and signs mean? For the given group size, we computed the difference in power functions $\beta_R(\theta_1, \theta_2) - \beta_C(\theta_1, \theta_2)$ for the grid $\theta_1, \theta_2 \in \{k\Delta : k = 0, \ldots, \Delta^{-1}\}$ with $\Delta = 0.01$. We can then count the number of (θ_1, θ_2)-values where the difference is positive, i.e. where test R has higher power than test C. If the difference is positive at all values of (θ_1, θ_2), we denote this in the table by a "$>$" sign as the upper entry, meaning that the row test has greater power than the column test at all (θ_1, θ_2)-values. If the row test has lesser power at all (θ_1, θ_2)-values, we place a "$<$" sign as the upper entry. If the power is equal at every grid point, we write a "$=$" sign. Finally, if one test is more powerful on some parts of Θ_1 and the other on other parts of Θ_1, we display the proportion of grid points where the row test has greater power than the column test. The lower entry (the number between parentheses, if present) is the numerical approximation of the volume between the two power functions, i.e. the number

$$\frac{1}{\Delta^2} \sum_{k=0}^{\Delta^{-1}} \sum_{\ell=0}^{\Delta^{-1}} \beta_R(k\Delta, \ell\Delta) - \beta_C(k\Delta, \ell\Delta). \tag{5.7}$$

In the case we have displayed an "$=$" sign, this sum is of course zero, so we left it out, explaining the empty entries. Thus, just as an example, if we want to compare the power functions of the tests `ET chi` and `LP1 S_V` for the group sizes $(n_{1\cdot}, n_{2\cdot}) = (20, 20)$, we should look at the cell the sixth row and the second column of Table C.3. Within that cell, the pair of entries on the second row and on the

third column is of interest to us. There we read the numbers 0.3641 and (0.0000). The first indicates that for 36.41% of the grid points in Θ_1, the power function of `ET chi` is higher than that of `LP1 S_V`. However, the volume between the two power surfaces is so small that it 0.0000 is displayed. Thus, at that 36.41% of Θ_1, the gain in power by using `ET chi` approximately compensates the loss in power by using `ET chi` in the remaining 63.59%.

There are a number of things to observe from these tables. First of all, just as we already expected a bit from the size functions, and from all the proponents of unconditional tests, we see a lot of "$<$" signs in the `PEARSON` and `FISHER` rows, indicating that the remaining tests (all unconditional) are more powerful. Also, in the `FISHER` row, one might argue that for increasing group sizes, the power function of Fisher's exact test seems to slightly "close the gap" with the other power functions. This observation is based on the numbers between parentheses to become slightly less negative when going going through the entries in each cell from top to bottom and left to right. For example, in the (`FISHER`, `C S_P M`) cell, the volume between the power surfaces is -0.0623 and -0.0863 for $(n_{1\cdot}, n_{2\cdot})$ equal to $(10, 5)$ and $(10, 10)$ respectively, whereas for $(n_{1\cdot}, n_{2\cdot}) = (40, 40)$ it is "just" -0.0403. This pattern does not hold every time however, in the (`FISHER`, `ET vol`) cell for example, we have a difference of -0.0303 at $(n_{1\cdot}, n_{2\cdot}) = (10, 5)$, but a difference of -0.0350 at $(n_{1\cdot}, n_{2\cdot}) = (40, 40)$.

Furthermore, looking at the three tables cells that compare the CSM-like tests, we see a lot equality signs. This we also saw coming by the many overlapping size functions of the CSM-like tests. Realise that if two size functions precisely overlap, one can say almost with certainty that the corresponding critical regions must consist of the same tables. But then from (5.1) it follows immediately that the power functions will also be identical on the whole of Θ. `C S_P M` and `C S_V M` perform identical, while one could argue that for the group sizes where `C S_chi M` has a different power functions than its two counterparts, it performs slightly worse most of the time. This can be seen in the (`C S_P M`, `C S_chi M`) cell, where the volumes between the two power surfaces are all non-negative (so in favour of `C S_P M`), and in the (`C S_chi M`, `C S_V M`) cell, where the volumes are all non-positive (thus in favour of `C S_V M`).

Moving on to the `ET` tests in Table C.2, we see many "$=$" signs. However, for the $(n_{1\cdot}, n_{2\cdot})$-values where the power functions are not identical, it seems that `ET chi` (recall, Suissa and Shuster's test) performs slightly worse than `ET fisher` (Boschloo's test). For $(n_{1\cdot}, n_{2\cdot}) \in \{(20, 5), (40, 5), (40, 10)\}$, Boschloo's test has a higher power function at all grid points. For $(n_{1\cdot}, n_{2\cdot}) \in \{(20, 10), (40, 20)\}$, there is no superior test for all values of (θ_1, θ_2), and `ET chi` has larger power on a slightly larger proportion of Θ_1. However, the volume between the two power surfaces is in both cases close to zero. Comparing `ET chi` with `ET vol` gives a more mixed impression, with `ET chi` being less powerful everywhere for certain $(n_{1\cdot}, n_{2\cdot})$-values, but also more powerful on a large portion of Θ_1 for other $(n_{1\cdot}, n_{2\cdot})$-values. A similar observation can be made comparing `ET fisher` with `ET vol`. However, it occurs less often that the power function of `ET fisher` is worse on the whole of Θ_1 than that of `ET vol`. If we would have to pick one test out of these three external statistic tests, we would suggest to use `ET fisher`, as it seems to generally have the highest power. However, this is not intended to be a rigorous statement.

Comparing the LP tests in Tables C.3 and C.4, we again see a lot of identical power functions. The `LP1 S_P` and `LP1 S_V` tests behave almost identically, just as the `LP2 S_P` and `LP2 S_V` tests. Furthermore, it again seems that the S_χ-based tests perform slightly worse. If we want to compare the `LP1` tests with the `LP2` tests, we should look at the cells in the intersection of the `LP1 S_P` and `LP1 S_V` rows and `LP2 S_P` and `LP2 S_V` columns. We temporarily leave out the S_χ-based tests because of their slightly lower power. In these cells, we see many "=" signs, but it seems that the remaining non-"=" entries are overall either smaller than 0.5, or not much bigger than 0.5, where the volume between the power functions is often either 0.0000, or something close to that. One might be tempted to say that because the `LP2` approach tries to maximise the average power on Θ_1, the `LP2` tests will achieve higher power for a greater portion of grid points than the `LP1` tests, explaining many of the entries smaller or around 0.5. On the other hand, as the volume between the power functions is always pretty close to zero, this indicates that on the smaller part of Θ_1 where `LP1` tests achieve greater power than `LP2` tests, this gain in power is greater than the loss of power on the remainder of Θ_1. There does not seem a clearly more powerful approach here either.

There are three more comparisons we would like to make. First of all, let us look at the `ET` tests versus the CSM-like tests in Table C.2. If not equal, the power functions of the CSM tests are almost always larger than that of the `ET` on large portions of Θ_1, and the occurrences where this is not true are most often observed when using the `C S_chi M` test, of which we had already noted it seemed to be less powerful. Focusing on `C S_P M` (`C S_V M` performs identical), out of the 19 non-"=" entries in the 3 cells comparing `C S_P M` with every `ET` test, the proportion of Θ_1 where the `C S_P M` power is just around 0.5 three times (once 0.4579 and twice 0.5458, all with positive volumes in favour of `C S_P M`). Apart from these three $(n_{1\cdot}, n_{2\cdot})$-values, the proportion is convincingly in favour of `C S_P M`, even coming out above 94% for 9 $(n_{1\cdot}, n_{2\cdot})$-values. The volumes between the power surfaces are in fact also always positive (in favour of the CSM tests), except for the `C S_chi M` test. We are therefore inclined to say that the `C S_P M` (and `C S_V M`) tests are generally more powerful than the `ET` methods. Again, no strong, uniform statement can be made however.

Comparing the `ET` tests with the LP tests in Tables C.2, C.3, and C.4, if equal power functions are observed, they are only found for the small $(n_{1\cdot}, n_{2\cdot})$-values of $(5, 5)$ and $(10, 5)$. For the rest, we either have a "$<$" sign (note that most of them are in the `ET chi` row), indicating that the `ET` power function takes smaller values everywhere on Θ_1, or a proportion of grid points where `ET` has larger power that comes above 0.5 only in 6 out of 126 non-"$<$" and non-"=" entries, and can get as small as 0.0345. The volumes between power functions is 0.0000 or higher (but never above 0.0005) in 38 out of the earlier 126 entries. These observations also give reason to believe that the LP tests might be more powerful in general than the `ET` tests.

Finally, looking at the CSM tests versus the LP tests in Tables C.2, C.3, and C.4, we see that the methods all have the same power for $(n_{1\cdot}, n_{2\cdot})$ taking values in $\{(5, 5), (10, 5), (10, 10)\}$. Some individual pairs of tests have identical power also for other $(n_{1\cdot}, n_{2\cdot})$-values. For 18 out of the 97 entries, we see that the power functions of the LP tests are larger than that of the CSM tests on a larger part of Θ_1. However, the corresponding volumes are all very close to zero. Interesting

to note is that `LP2 S_P` (and `LP2 S_V`, which performed identically), is the only LP test never has smaller power on a majority of Θ_1 than a CSM test for all $(n_{1\cdot}, n_{2\cdot})$-values. Summarising, it seems that the LP tests generally have higher power on a bigger part of Θ_1, but the gain in power relative to CSM tests on this majority of Θ_1 is smaller than the loss in power on the remainder of Θ_1, as the volumes between the power surfaces are generally close to zero.

5.4.2 3×2 tables

For 3×2 tables, we are still able to plot the size function. Let us consider the outcome spaces of tables with group sizes $(n_{1\cdot}, n_{2\cdot}, n_{3\cdot})$ equal to

$$\begin{aligned}(5,5,5),\ (10,5,5),\ (20,5,5),\ (20,10,10),\\ (10,5,5),\ (20,10,5),\ (20,20,10),\\ (10,10,10),\ (20,20,5),\ (20,20,20).\end{aligned} \tag{5.8}$$

We can again plot the size functions for different values of $(n_{1\cdot}, n_{2\cdot}, n_{3\cdot})$. This time, we limited ourselves to $\alpha = 0.01$; the other plots would convey a similar message. In Figures C.7, C.8, and C.9, the reader can find the size functions for the indicated tests and groups sizes. In the right column of plots, one can find all the LP tests, the remaining ones are always in the left column. For the small group size of $(5, 5, 5)$, we again see that all tests all have a overlapping size which is quite far away from α. Only Fisher's exact test has a smaller size there. Apart from that, glossing over all the plots there are three important observations to make. First is that we the blue line cross over the α level a couple of times. This should not come as a surprise, as there is no reason why the asymptotic test should actually be valid. It is shown in Aanes [5] that in fact the Pearson's chi squared test is only asymptotically valid, and these plots form a clear counterexample that it need not be valid in general. The second observation is that the LP tests manage to have size very close to α for many θ-values, achieving greater size than many of their non-LP counterparts. Due to the low resolution of the `R` plots, it seems that the LP tests seem to surpass the α level. We checked however that this is not the case. Finally, observe that the `ET chi` and `ET vol` test often have a size lower than that of Fisher's test for many values of θ. Boschloo's test of course always has a larger size than Fisher's exact test. In general, the gap between the size function of `FISHER` and that of the unconditional tests is already a lot smaller than in the 2×2 case, especially for the larger group sizes. This is in agreement with the observations from Mehta and Hilton [55].

Let us move on to comparing the power functions of the tests on Θ_1. Realise that the power function is now a function of three parameters; θ_1, θ_2, and θ_3. Consequently, keep in mind that the power comparisons in Tables C.5, C.6, C.7 and C.8 are now made over a three-dimensional grid of $(\theta_1, \theta_2, \theta_3)$-values. We chose the spacing between grid points a bit larger now, such that $\theta_1, \theta_2, \theta_3 \in \{k\Delta : k = 0, \ldots, \Delta^{-1}\}$ with $\Delta = 0.1$. The numbers between parentheses should be interpreted as in (5.7), but now with a triple sum as β is a function of three arguments, and a division by Δ^3 instead of Δ^2 (the exponent should in general be equal to $r(c-1)$).

A first observation we can instantly make is that all the tests have identical power for $(n_{1\cdot}, n_{2\cdot}, n_{3\cdot}) = (5, 5, 5)$, except Fisher's test, which has lower power.

Another thing that stands out is that the PEARSON and FISHER rows no longer contain as many "<" signs as they did in the 2×2 case. Still, the PEARSON and FISHER power functions are smaller than those of the unconditional tests on a majority of Θ_1, but this again points towards Fisher's exact test becoming less conservative for larger tables. Moreover, going through the entries from left to right in each cell, we observe the numbers in parentheses to become once again less negative; Fisher's power function moves closer to its unconditional counterparts.

Recall that for several group sizes, the chi square test actually had a size larger than α, which gives it a "head start" to achieve greater power on some parts of Θ_1 too. Also, the numbers between parentheses, i.e., the numerical integrals of the differences of two power functions on Θ_1, are most of the time negative. It seems that the head start the chi squared test got did not last that long. As a side note, observe in Table C.6 that Boschloo's test is the only test which has uniformly greater power than Fisher's exact test for all $(n_{1\cdot}, n_{2\cdot}, n_{3\cdot})$-values.

Moving down to the three CSM rows, we are again inclined to say that C S_chi M is less powerful than C S_P M and C S_V M, while there does not seem to be a clear winner in terms of power between C S_P M and C S_V M. Comparing with the ET tests, we again see many proportions larger than 0.5 and positive numbers between the parentheses, similar to the 2×2 tables. Looking even further to the LP columns to compare CSM and LP tests however, notice that the LP1 S_P and LP1 S_V columns only have proportions smaller than 0.5 and negative numbers between the parentheses for all CSM tests. Contrary to the 2×2 tables, it now seems that at least the LP1 S_P and LP1 S_V tests are in general more powerful than the CSM tests. For the remaining LP tests, there are also some proportions larger than 0.5 and positive integral values, so we cannot draw as strong a conclusion as with LP1 S_P and LP1 S_V.

Amongst the ET tests, it seems that Boschloo's test performs better than the other two, as can be seen in the mostly negative integral values and less-than-half proportions in the (ET chi, ET fisher)-cell, and the solely positive integral values and mostly more-than-half proportions in the (ET fisher, ET vol)-cell. Furthermore, just as in the 2×2 setting, comparing the ET tests with the LP tests yields cells with almost exclusively entries in favour of LP tests being more the powerful ones.

Finally, comparing the LP tests amongst each other, we see multiple "=" signs, and the non-identical entries are again of such a mixed nature that it is difficult to say anything meaningful about which of the 6 tests is the most powerful. However, looking at the cells involving either LP1 S_chi or LP2 S_chi, it again seems that the chi-square approaches are less powerful. Also note the greater level of similarity between the power functions of the (LP1 S_P, LP1 S_V)- and (LP2 S_P, LP2 S_V)-pairs.

5.4.3 2×3, 3×3, and 2×4 tables

Considering tables which all have more than two columns, we will only look at the power comparison tables. We will consider the following combinations of c and

$(n_{1\cdot}, n_{2\cdot})$ (or $(n_{1\cdot}, n_{2\cdot}, n_{3\cdot})$ in the 3×3 case):

$$\begin{aligned} &(5,5),\ c=3;\ (10,5),\ c=3;\ (10,10),\ c=3; \\ &(20,5),\ c=3;\ (5,5,5),\ c=3;\ (5,5),\ c=4. \end{aligned} \tag{5.9}$$

As discussed in Section 4.2.3, we will not include the `C S_chi M` test in this study. The power comparison tables can be found in Tables C.9 and C.10, and are based on grids with a spacing of $\Delta = 0.2$ now.

First thing to note is that we once again see many "$<$" signs in the `PEARSON` and `FISHER` rows (especially in the former). Apart from the `LP S_P`, `LP S_V` and `C S_P M` tests, who generally have higher power than `FISHER`, the power of Fisher's exact test is comparable to that of the other tests. For these table dimensions the results seem to be less in agreement with Mehta and Hilton's conclusions, but keep in mind that the sample sizes have been kept relatively small in order to limit the computation time, leading to a smaller outcome space and less opportunities for the power of Fisher's exact test to "catch up" with that of the unconditional methods.

Moving on to the CSM tests, the `C S_P M` test has higher power than `C S_V M` on a majority of Θ_1. `C S_P M` also has generally higher power than the `ET` tests and the LP tests using the S_χ condition. It also has a power comparable to that of the remaining LP tests. `C S_chi M` does not fare so well though; it has a power function lower than that of the `ET fisher` and `ET vol` tests, more comparable to that of the two `LP S_chi` tests. Out of the `ET` tests, Boschloo's test again seems to be the most powerful, and Suissa and Shuster's test the least powerful. Note however that Suissa and Shuster's test has to suffer from the odd behaviour of the chi-square test statistic in tables with $c > 2$, as mentioned in Section 4.2.3. All `ET` tests have generally lower power than the LP tests, except when comparing to the two `LP S_chi` tests, to which the power functions of `ET fisher` and `ET vol` seem more similar. Finally, comparing the LP tests to each other, we already mentioned the relatively low power of the two `LP S_chi` tests. Apart from that, the differences are smaller, but `LP1 S_P` and `LP2 S_P` seems to have a comparable power functions, that are greater than the power functions of the `LP S_V` tests on a majority of Θ_1.

5.4.4 Main takeaways from the power study

The observations we just made are necessarily hand-waving, containing a lot of words like "generally", "similar", "almost", and "often". There is no single test which we (theoretically) expect to always have higher power than another test (other than Boschloo's test having higher power than Fisher's exact test, by construction). Consequently, there will always be certain table dimensions and $(n_{1\cdot}, n_{2\cdot})$-values where one test performs better than the other and other $(n_{1\cdot}, n_{2\cdot})$-values were the reverse is true. We can only make some cautious statements about some of the tests based on the few $(n_{1\cdot}, n_{2\cdot})$-values we have considered. These should serve as the main takeaways from Sections 5.4.1, 5.4.2, and 5.4.3, but it cannot be stressed enough that these should be interpreted with a certain degree of caution. The key points are:

- The asymptotic test is often either too conservative (having low power), or

too liberal (having high power but also a size higher than the given significance level).

- Just as the critics of Fisher's exact test point out, this test is often conservative compared to the unconditional methods. However, in agreement with Mehta and Hilton [55], this conservativeness becomes less visible for increasing table dimensions and sample sizes.
- Boschloo's (unconditional) test has uniformly higher power than Fisher's exact test, and turned out overall to be the most powerful test using an external test statistic.
- The tests using the chi square test statistic (either as an external test statistic or via the symmetry condition S_χ) generally have lower power than their respective counterparts.
- The `ET` tests are less powerful than the CSM-like and LP tests.
- Between the LP1 and LP2 tests, we did not manage to label one of the two approaches as more powerful.
- Comparing the CSM and LP tests, there is no single clearly more powerful approach.

5.5 Long-term power

Martín Andrés and Silva Mato [75] found the results from a power comparison to be "unsatisfactory", as no test has been proven to be uniformly more powerful than another test (again, Boschloo's test being the exception here). It is hard to compare power functions which take different values at different values of θ_1 and θ_2 (in the 2×2 case; more generally at $\boldsymbol{\theta}_{ij}$ for $i = 1, \ldots, r$ and $j = 1, \ldots, c$). To this end, they introduce a metric called the long-term power. They argue that in order to make a global statement about the power function over the whole parameter space Θ, we should assign an "a priori" distribution to the $\boldsymbol{\theta}_i$ vectors. Martín Andrés and Silva Mato assume that, in the long term, every time a researcher will encounter a contingency table like this, the θ_i vectors will follow a uniform distribution [1]. They consider this distribution to reflect our ignorance about the θ_i-values. Given a level α, the long-term power β' is then the overall probability of rejecting the null hypothesis and can be found by the law of total probability, as done below.

$$
\begin{aligned}
\beta' &:= \int_\Theta \beta(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_r)\, \mathrm{d}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_r) \\
&\overset{(5.1)}{=} \int_\Theta \sum_{\mathbf{x}\in K^\alpha} P_{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_r}(\mathbf{X} = \mathbf{x})\, \mathrm{d}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_r) \\
&\overset{(5.2)}{=} \int_\Theta \sum_{\mathbf{x}\in K^\alpha} \frac{\prod_{i=1}^r n_{i\cdot}!}{\prod_{i=1}^r \prod_{j=1}^c x_{ij}!} \prod_{i=1}^r \prod_{j=1}^c \theta_{ij}^{x_{ij}}\, \mathrm{d}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_r) \\
&= \sum_{\mathbf{x}\in K^\alpha} \frac{\prod_{i=1}^r n_{i\cdot}!}{\prod_{i=1}^r \prod_{j=1}^c x_{ij}!} \int_\Theta \prod_{i=1}^r \prod_{j=1}^c \theta_{ij}^{x_{ij}}\, \mathrm{d}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_r) \\
&\overset{(i)}{=} \sum_{\mathbf{x}\in K^\alpha} \frac{\prod_{i=1}^r n_{i\cdot}!}{\prod_{i=1}^r \prod_{j=1}^c x_{ij}!} \prod_{i=1}^r \left(\int_{\Theta^i} \prod_{j=1}^c \theta_{ij}^{x_{ij}} \mathrm{d}\boldsymbol{\theta}_i \right) \\
&\overset{(ii)}{=} \sum_{\mathbf{x}\in K^\alpha} \frac{\prod_{i=1}^r n_{i\cdot}!}{\prod_{i=1}^r \prod_{j=1}^c x_{ij}!} \prod_{i=1}^r \frac{\prod_{j=1}^c x_{ij}!}{(n_{i\cdot} + c - 1)!} \\
&= \sum_{\mathbf{x}\in K^\alpha} \frac{\prod_{i=1}^r n_{i\cdot}!}{\prod_{i=1}^r (n_{i\cdot} + c - 1)!} \\
&= \sum_{\mathbf{x}\in K^\alpha} \frac{1}{\prod_{i=1}^r \binom{n_{i\cdot}+c-1}{n_{i\cdot}} (c-1)!} \\
&= \frac{1}{((c-1)!)^r} \sum_{\mathbf{x}\in K^\alpha} \frac{1}{\omega},
\end{aligned}
$$

where in the last equality we recognise the expression for the number of tables with r rows and c columns with group sizes $(n_{i\cdot})_{i=1,\ldots,r}$. Thus, the long-term power β' is nothing else than the ratio of tables in the critical region to the total number of tables in the outcome space (up to a constant, which equals one in the 2×2 case that Martín Andrés and Silva Mato considered, and in general as long as

[1] That is, each θ_i will follow a uniform distribution on the $c-1$-dimensional simplex. Martín Andrés and Silva Mato argued only for the 2×2 case, where the uniform distribution is on the unit interval.

$c = 2$). At equality (i) we wrote $\Theta = \Theta^1 \times \ldots \times \Theta^r$, where the Θ^i are just the $c-1$-dimensional simplexes in which the respective $\boldsymbol{\theta}_i$ vectors reside. This makes it straightforward to pull the big integral over Θ apart into a product of r integrals over the simplexes Θ_i. But then, realise that one such an integral is nothing else than the integral in (4.13) , where the $n_{\cdot j}$ are replaced by the x_{ij}. Equality (ii) then directly follows from (4.13).

From this, we can immediately give an alternative justification for the LP formulation (4.20). As this linear programming problem tries to maximise the number of tables in the critical region, it will, by construction, have the largest possible long-term power. Thus, we can say that `LP1` (with no symmetry condition) is the most powerful test in the long term. It is up to the reader to decide how much value to attach to this property. As soon as we add a symmetry condition to the LP test, it is not necessarily the most powerful on the long-run. Indeed, because we can now only add groups of tables to the critical region instead of individual tables, we are less flexible to make the critical region as large as possible. The larger the equivalence classes, the smaller β' will be.

For all the table dimensions and group sizes we have considered for the power comparison in the previous section, it is now straightforward to list the corresponding long-term powers. This is done in Table C.11. For each row, we indicate the group sizes, as well as the number of columns. The maximum value in each row is displayed in bold. Notice how `LP1 S_P` always has the largest long-term power, even though it is (in theory) not the optimal test in terms of long-term power. For 2×2 tables we actually see a lot of equal values, indicating that the critical regions of many tests consist of the same number of tables (but they are generally not identical, as the tables from the power comparison tell us). However, as the table dimensions increase, we see how only `LP1 S_P` manages to keep the highest value every time, while the `C S_P M` and `LP2 S_P` tests often have a long-term power not much smaller than that of `LP1 S_P`, in particular for the outcome spaces of the tables with more than 3 columns.

If we only consider the CSM-like tests, both `C S_P M` and `C S_V M` perform identical in the 2×2 case (here the critical regions are in fact the same since $\mathrm{S_P}$ yields the same symmetry classes as $\mathrm{S_V}$), with `C S_chi M` test sometimes lacking behind. For 3×2 tables, `C S_P M` has a slightly higher long-term power than `C S_V M` in most cases, and both have a higher long-term power than `C S_chi M` (except when $(n_{1\cdot}, n_{2\cdot}, n_{3\cdot}) = (10, 10, 10)$). This should again indicate that the S_χ ordering generally yields fewer, larger symmetry classes, resulting in smaller critical regions (and generally lower power as seen in the previous Section). For the tables with more than two columns, `C S_P M` always had a higher long-term power than `C S_V M`.

Out of the three `ET` tests, Boschloo's test has the highest long-term power every time, except for the 2×2 table with group sizes $(40, 20)$ and the 3×2 table with group sizes $(10, 5, 5)$. Compared to the CSM and LP methods however, the `ET` tests generally have a lower long-term power. Finally, between the LP tests, `LP1 S_P` always has the highest long-term power. In the 2×2 case, all LP tests not using S_χ have an identical long-term power. For 3×2 tables, the `LP1 S_V` still has identical long-term power to `LP1 S_P`, while `LP2 S_P` and `LP1 S_V` do not lack far behind. For tables with more than 3 columns, `LP2 S_P` has the highest long-term power of all the LP tests not including `LP1 S_P`.

Keep in mind that Table C.11 essentially only shows the number of outcomes in the critical region, and does certainly not tell a complete story of the power of the different tests. Just as the power comparison of Section 5.4, this is not a perfect metric. However, in the interpretation described by Martín Andrés and Silva Mato, we would conclude that under the assumption of uniformity of the probabilities $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_r$, in the long run, for $\alpha = 0.01$, `LP1 S_P` is be the most powerful test, `C S_P M` generally the most powerful CSM test, and Boschloo's test generally the most powerful `ET` test.

5.6 The choice of a mathematics programme among male and female students

In Table 1.2, we showed the number of male and female students in each of the three mathematics programmes offered by NTNU in 2007. We moved on with the rest of this thesis leaving an important question unanswered. Is there a significant difference in the choice of mathematics study programme between the male and female students? Let us conclude this Chapter by picking a significance level of $\alpha = 0.05$, and performing some of the tests that we have encountered throughout this text, to decide whether or not to reject the null hypothesis

$$H_0 \colon \theta_{\mathrm{M1}} = \theta_{\mathrm{F1}},\ \theta_{\mathrm{M2}} = \theta_{\mathrm{F2}},$$

where we defined $\theta_{\mathrm{M1}}, \theta_{\mathrm{F1}}$ as the proportion of male/female students that opted for a 3-year bachelor and 2-year master programme, and $\theta_{\mathrm{M2}}, \theta_{\mathrm{F2}}$ as the proportion of male/female students that chose a teacher education.

First of all, to answer this question, we may opt to use Pearson's chi square test, although Cochran's rule does not recommend it. Using Yates' correction, we find a chi-square value of 8.8791, corresponding in this case to a p-value of 0.0118. In case one is hesitant to use the asymptotic test here, we might also use the Fisher–Freeman–Halton exact test, which gives a p-value of 0.0583.

Apart from these two popular tests, which return a result almost instantly, we might opt for an exact unconditional test. As we have seen in Section 3.7, because this is a purely descriptive experiment, the use of an unconditional test does not lead to logical incoherencies. Some might still deem the use of this test inappropriate, others will argue that it is the only correct way to do inference in this case. Either way, executing a $\mathrm{CS_PM}$ test yields a p-value of 0.0502. Alternatively, we can use a test which orders the table based on an external test statistic, such as Boschloo's test, which yields a p-value of 0.0524 and, if we perform an LP2 $\mathrm{S_P}$ test combined with a binary search for the p-value, we find that Table 1.2 has a p-value of 0.0290.

As we would expect, the unconditional methods return a – in this case slightly – smaller p-value than Fisher's exact test. However, for all tests but the LP test, the obtained p-value is larger than the significance level of 0.05, meaning that we would not reject the null hypothesis for all but the LP test. The fact that the LP p-value (which really is just an upper bound on the p-value) is so much smaller than the other ones, again illustrates the flaw of the LP method. Because the LP test tries to find the smallest level for which Table 1.2 is still included in the critical region, the p-value it will return is artificially low, corresponding to a most

probably non-intuitive critical region which happens to include the observed table. Not requiring the nestedness of the critical regions allows to label an observation as significant evidence against the null hypothesis, even though other tests would not reject H_0 based on that observation. Clearly, striving for the highest power should not be the only aim when devising hypothesis tests.

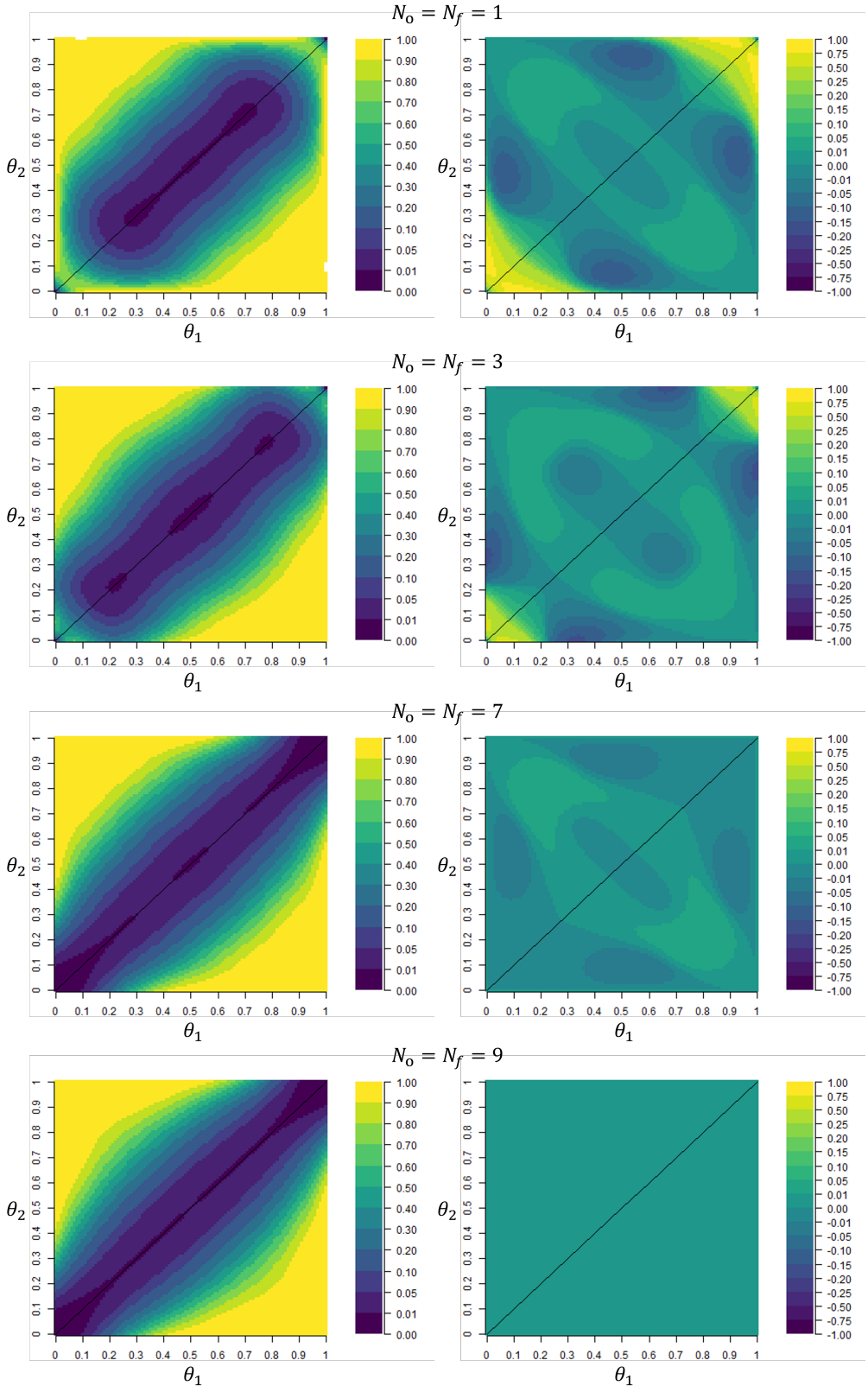


Figure 5.15: $\beta(\theta_1, \theta_2)$ for indicated values of N (left) and the difference $\beta(\theta_1, \theta_2) - \beta_{\text{benchmark}}(\theta_1, \theta_2)$ (column).

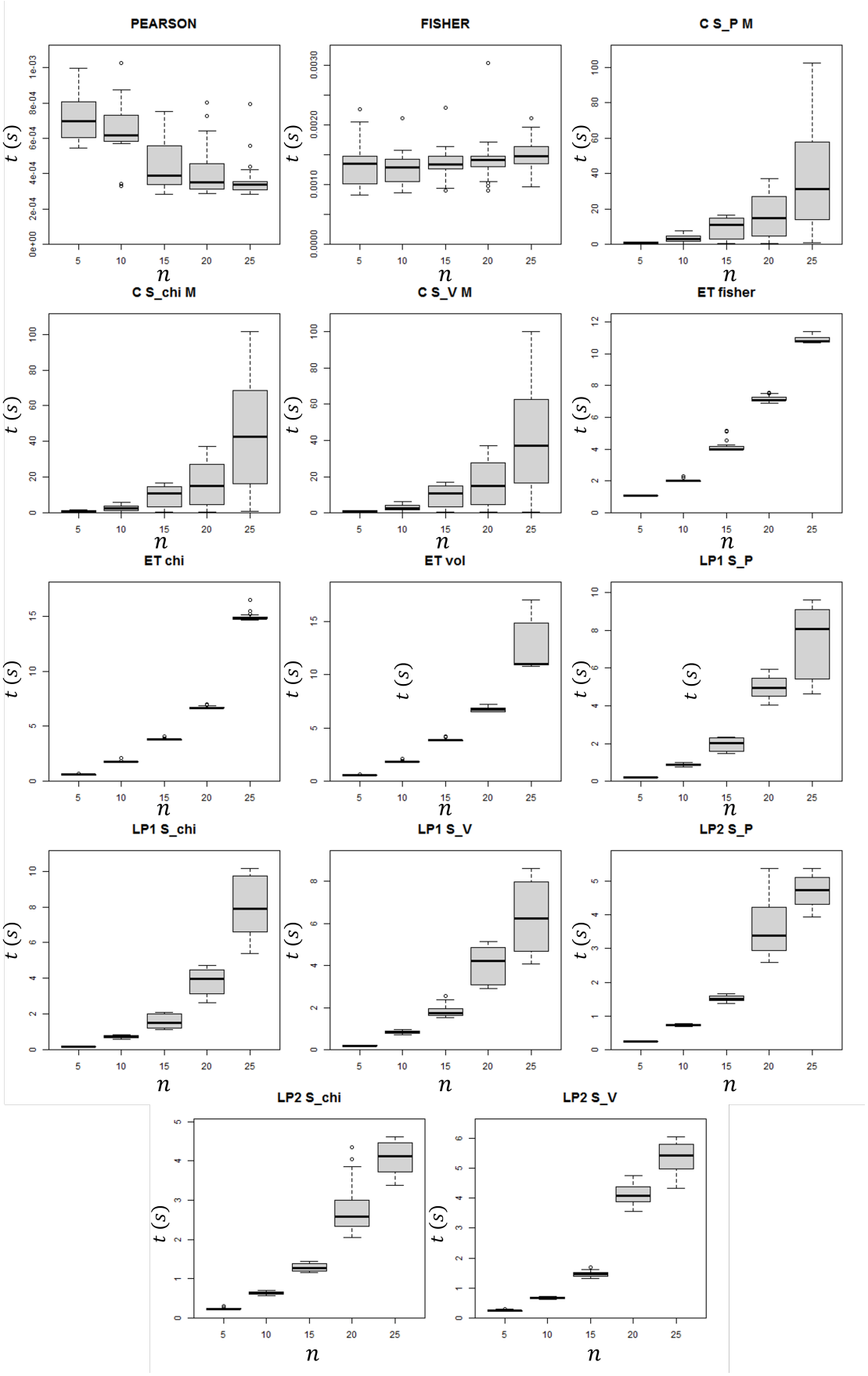


Figure 5.16: Runtime as a function of the common group size n for the 14 tests.

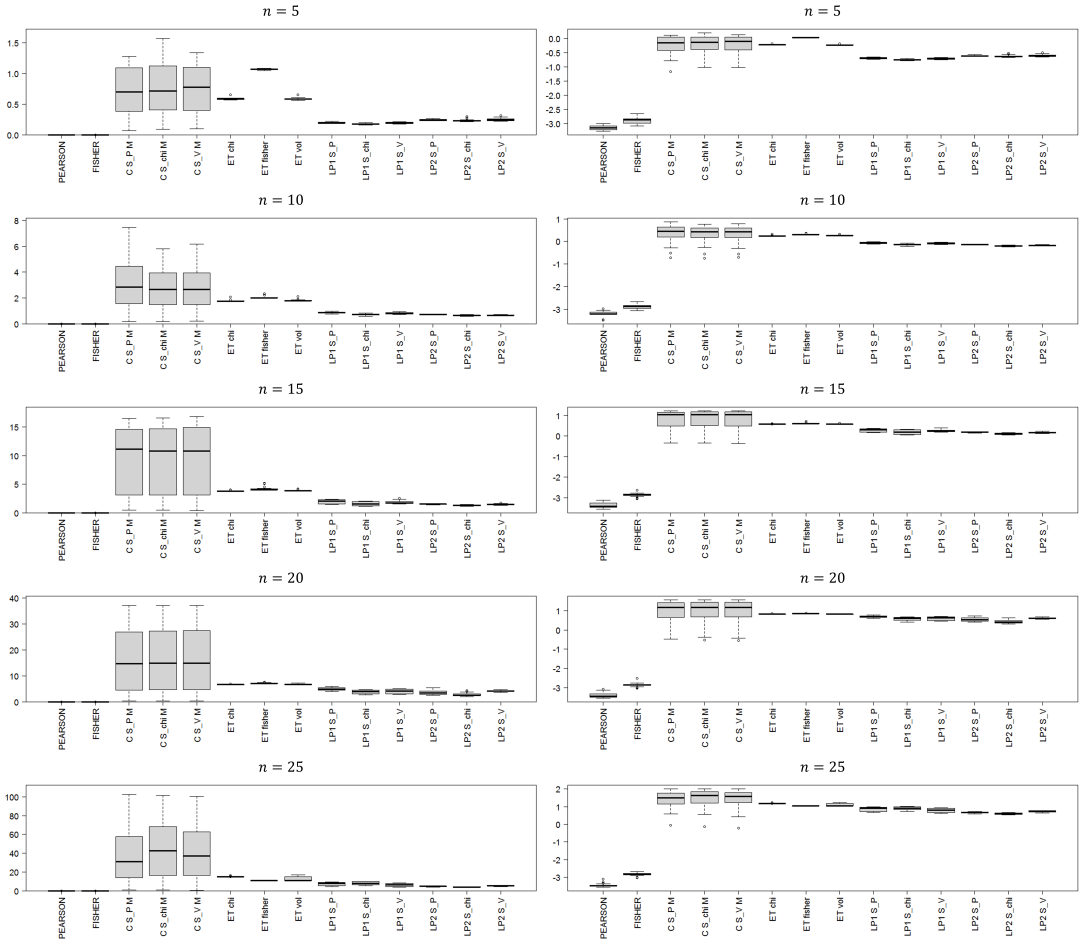


Figure 5.17: Runtime and log-runtime comparison of the 14 tests on 2×2 tables with indicated values of $n_{1\cdot} = n_{2\cdot} = n$.

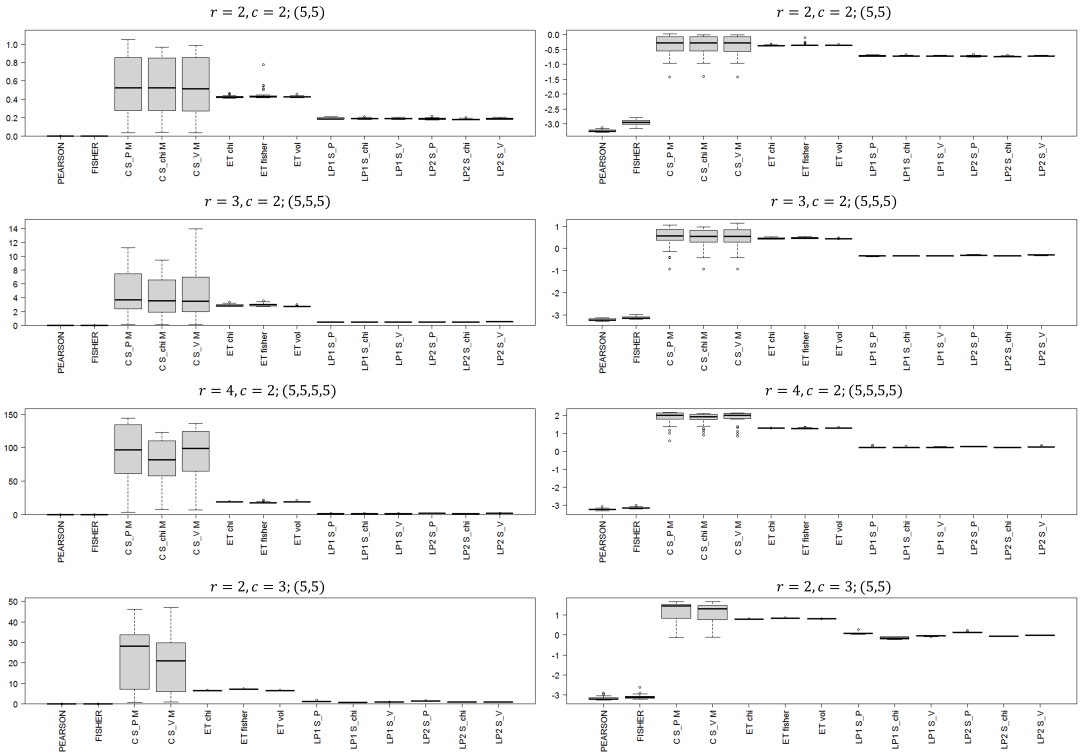


Figure 5.18: Runtime and log-runtime comparison of the 14 tests (13 in the case of 3 columns by removal of the $CS_\chi M$ test) with indicated group sizes and table dimensions.

CHAPTER

SIX

DISCUSSION

In this Chapter, we will discuss some of the issues, limitations, and design choices of the main additions of this thesis to the field of exact unconditional hypothesis tests on general $r \times c$ contingency tables. As to whether there is any justification, other than pure mathematical interest, to research this field, we gladly refer to Section 3.7. In Chapter 4, we essentially introduced two approaches to testing the hypothesis (4.3). The first was a generalisation of the already existing unconditional methods for 2×2 tables, in particular Barnard's CSM test. The second was an approach entirely inspired by the Neyman–Pearson school of thought on hypothesis testing. This approach uses a linear programming (LP) formulation to construct an – in some sense – optimal critical region. In total, we ended up with 14 different tests, which we tried to compare in Chapter 5 in terms of speed and power.

One should be cautious with the conclusions drawn from these two comparisons. Of course, all other things being equal, the speed of the different tests still depends on the efficiency of the underlying implementation. Therefore, our speed comparison does not say anything about the efficiency of the methods themselves, only about the efficiency of our implementations. As mentioned in Section 5.3.3, we did not manage to find any meaningful theoretical results regarding the expected runtime or number of operations of the different methods (especially for the LP methods). This would be a great first topic for further study.

Talking about the efficiency of the presented implementations, one particular optimisation immediately comes to mind. Right now, when performing a CSM-like test, all the symmetry classes are computed beforehand. If we are not interested in a complete ordering of the outcome space, some of these equivalence classes may not ever be considered by the algorithm. A speedup could therefore be reached if we were to only compute a symmetry class the moment a table in that class is a possible candidate to be the next table in the ordering.

Moving on to the power comparison, we remarked multiple times throughout Section 5.4 that also the power comparison is of limited use. On one hand, we are comparing tests with different sizes, as the precise level α will in general not be reached in this discrete setting. On the other hand, there are no results grounded in theory that would lead us to expect one particular test to be more powerful than the other. Once again, the only exception is Boschloo's test, which is uniformly more powerful than Fisher's exact test, because Boschloo's test is a supremum

test which uses the Fisher p-value as a test statistic. Since the test power is a function of the level α (of which we have just looked at one value: $\alpha = 0.01$), and of the table probabilities $(\theta_{ij})_{i=1,\dots,r;\ j=1,\dots,c}$ described under (4.2) , different tests will often alternate being the one having the higher power of the two for different levels and probabilities. To this end, we briefly looked at the long-term power, but there are a number of alternatives which are worth looking into for further research. For the 2×2 setting, Haber [76] suggested for example to compute, for a given level α, the minimum of the power function a distance d from Θ_0, i.e.

$$\beta_{\min}(d) := \min_{|\theta_1 - \theta_2| = d} \beta(\theta_1, \theta_2).$$

Just as the size function, this univariate function can easily be visualised and compared for different tests. It should be possible to generalise this quantity to general $r \times c$ tables as well by defining an appropriate metric to quantify the distance that $\boldsymbol{\theta} = (\theta_{ij})_{i=1,\dots,r;\ j=1,\dots,c}$ lies from Θ_0. Other alternatives, which remove for example the dependence on the chosen level α can be found in the paper by Martín Andrés and Silva Mato [75], from which we also got the idea to consider the long-term power.

All unconditional tests we proposed in this text at some point required a discretisation of Θ_0 (or even the whole of Θ in the case of LP formulation (4.24)). To construct this discretisation, we have used a low-discrepancy sequence transformed from $[0,1]^c$ onto the $c-1$-dimensional simplex Θ_0. There are some technical objections in doing this, for which we refer to Appendix A. Also, there are many types of low-discrepancy sequences one could choose from, each of them showing different behaviour depending, amongst others, on the number of columns c. We discuss the choice we have made for this text as well in Appendix A. The main takeaway is that for larger tables than the ones considered here, it would be interesting to look into the quality of the tests when executed using different types of low-discrepancy sequences.

Apart from these considerations specific to the use of low-discrepancy sequence, there is a more general comment to make. Realise that, using the Quasi-Monte Carlo approach, the grid has been constructed to be as evenly distributed over Θ_0 as possible. One might actually wonder whether this is the best approach. For example, as mentioned by Lydersen, Langaas and Bakke [4], the sum of $P(\cdot;\boldsymbol{\theta})$-functions can have sharp peaks near the boundaries of Θ_0. It may be wise to have a higher concentration of grid points around the edges, in order not to miss any potential maxima. Related to this is a completely different approach that might have been worth implementing and investigating. For a given table, the sharp peaks may have been far away from the "interesting" possible values for $\boldsymbol{\theta}$ (for example, in the 2×2 case, near the MLE $\hat{\theta} = (x_{11} + x_{21})/n_{..}$). Therefore, we might not want to include the peaks when determining a supremum p-value $p(\mathbf{X}) = \sup_{\boldsymbol{\theta} \in \Theta_0} P_{H_0}(T(\mathbf{X}) \geq T(\mathbf{x}))$ for some test statistic $T(\mathbf{X})$. Berger and Boos [77] showed that

$$\sup_{\boldsymbol{\theta} \in C_\gamma} P_{H_0}(T(\mathbf{X}) \geq T(\mathbf{x})) + \gamma,$$

where C_γ is a $1 - \gamma$-confidence interval for θ, is also a valid p-value. This would allow us to restrict the maximisation to a domain $C_\gamma \subset \Theta_0$, most likely excluding the region closest to the boundaries.

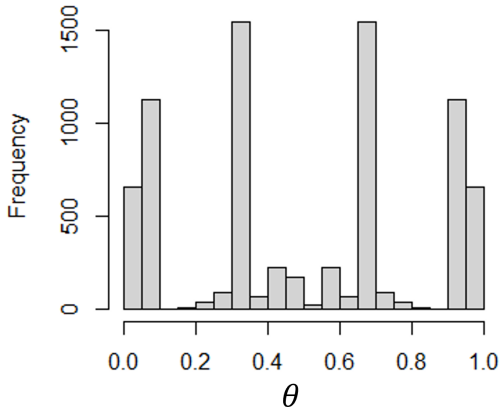


Figure 6.1: Frequency of θ-values at which a minimal maximum has been recorded in the CSM procedure.

Apart from near the edges, there exist perhaps other areas of Θ_0 where we are more likely to encounter, say, a minimal maximum of the current sum of $P(\cdot; \boldsymbol{\theta})$-functions when doing a CSM test. For example, in Figure 6.1, we display a histogram indicating the frequency of θ-values at which minimal maxima occurred when ordering the space of 2×2 tables with $(n_{1\cdot}, n_{2\cdot}) = (75, 50)$. That is, at each iteration of the CSM method, we recorded the θ-value at which the sum of $P(\cdot; \theta)$-functions of all tables that have already been ordered reached its maximum. There seems to be higher concentration of maxima near the edges, and around (but not including) $\theta = 0.5$. It would be an interesting research topic to find out whether one can, a priori, locate these areas of higher concentration given only the table and group sizes. A non-uniform (and probably smaller) grid can then be constructed accordingly, without giving in on the quality of the table ordering we end up with. This is also relevant for the LP methods. When constructing the constraint matrix A, we can now decrease the number of rows to only include fewer, "more relevant" values of $\boldsymbol{\theta} \in \Theta_0$ (or Θ), leading to a smaller LP problem that needs solving.

Finally, as we mentioned the LP methods, we would like to briefly discuss whether this approach has a future in hypothesis testing. Although very quick and powerful compared to many of the other unconditional methods, the non-nestedness of the critical regions is quite a big flaw of this testing approach. We wonder if this is an inherent flaw of using a general optimisation method, which will simply find the most optimal solution given the significance level α, not interested in any underlying structure we find desirable from a statistical point of view. If so, then we fear that the LP methods might be of limited use, other than perhaps giving a rough outline of the shape of a critical region in certain cases. If not, however, a recommendation for further research would be to look for ways in which we can add additional requirements, constraints or rules to the LP approach that would make the obtained critical regions nested. Solving this issue would mean that there exists an exact unconditional test that is computationally feasible within reasonable time, whenever the corresponding linear programming problem is. The LP tests would thus benefit from all the effort being put in pushing the

boundaries of what optimisation software can solve. Another added flexibility of the LP method is that the branch-and-bound method used to solve mixed integer linear programmings also keeps track of feasible solutions along the way. If the contingency table does become too big at some point, it is still possible to find non-optimal solutions of a certain quality. It would be very interesting to see if these less-than-optimal solutions can be used, perhaps in combination with a more conventional unconditional method to come more quickly to a p-value for an observed table, or even an ordering of the entire outcome space.

CHAPTER

SEVEN

CONCLUSIONS

Whether it is the testing of the effectiveness of a new medicine, or finding out whether male and female students behave differently when deciding on a university education in mathematics, one is looking for a statement about the significance of a potential effect or difference. Inevitably, a statistical test of significance should be performed on a contingency table. We have discussed what types of test can be performed, first for the simple 2×2 table, and later for general $r \times c$ tables. Along the way, we have also explored the different opinions on which tests *should* be performed, in which situations.

For $r \times c$ tables, the literature on asymptotic tests and exact conditional tests, which consider both marginal totals of the contingency table as fixed, is already quite extensive. Meanwhile, the exact unconditional tests, which only fix the margin concerning the group sizes, have not received the same attention. This is most probably due to philosophical objections, and due to their computational complexity, especially for larger tables. Having considered both the philosophical arguments in favour of and against the use of each type of contingency table test, it is the opinion of the author that the exact unconditional tests certainly have a right to exist. Whether or not a certain test is appropriate or not, should depend on the experimental design that led to the contingency table of interest. With that being said, there are definitely scenarios in which one can deem the unconditional approach as appropriate, or at least where no logical case can be made against this approach.

The computational burden an unconditional test brings cannot be denied, nor can it completely be removed. After all, conditioning on only one set of marginal totals leads to a far greater set of possible, alternate table outcomes that one needs to work through. However, in order to partially remove that workload, we have introduced two approaches. The first one generalises the already existing methods for the 2×2 table to larger tables, applying a couple of speedups along the way. The most important of those were to perform all maximisations using a Quasi-Monte Carlo method, and the definition of new symmetry conditions that would split up the space of possible table outcomes into equivalence classes.

The second approach aims to construct a critical region that is in some sense optimal, given a significance level α. This is achieved by formulating a binary linear programming problem, which can then be solved with existing optimisation software. This makes the second approach much faster than the first one, while

maintaining comparable power. This approach is fully in line with the Neyman–Pearson school of hypothesis testing, which might be a downside for many already. Furthermore as the optimal solution is dependent on the level α, there is no guarantee that obtained critical regions are nested when α increases. We proposed ideas to define a p-value based on the linear programming approach. However, because of the non-nestedness of the critical regions, this p-value need not be valid. For the applicability of this approach, it is very important to find out whether or not this non-nestedness issue can be resolved.

Finally, we performed a comparison study to see how the power functions of asymptotic, conditional and unconditional methods would fare in both the 2×2 case, as for larger tables. From that, we drew a number of conclusions. First of all, although Fisher's exact test is often criticised for being too conservative, this critique becomes less and less applicable for increasing table dimensions. This is in agreement with Mehta and Hilton [55]. Furthermore, the tests inspired on Barnard's original CSM test, and the LP tests seem to have comparable power, and higher power than the asymptotic test, Fisher's exact test, and the tests using external test statistics. Most of the time there was however no single test which had the highest power on the whole of Θ. Only in a few specific cases was it impossible for someone to find some $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_2)$-value where the power of any test of choice would be higher than the power of another test. To this end, we also compared the long-term power as introduced by Martín Andrés and Silva Mato [75]. Evidently, the LP test which maximised the amount of tables in the critical region (using the S_P condition) was the most powerful in a long-term sense. The CS_PM was generally the most powerful CSM-like test, and Boschloo's test the most powerful test using an external test statistic.

Apart from a formal disclaimer that one can always speed up the `R` implementations of tests made by the author, there are a number of interesting other directions for further research. First of all, one might aim to develop a smarter discretisation of Θ_0 over which to perform the maximisations required for the LP tests to construct the constraint matrix, and for the supremum tests in order to determine the ordering of the table outcomes and/or to compute the p-values of the table outcomes. A more general recommendation would be to research different metrics in order to compare power functions in higher dimensions, as will occur in the case of contingency tables with a large amount of rows or columns. Finally, if the LP approach can be forced to return nested critical regions, many interesting research questions come up, such as looking for alternative LP formulations, or combining LP tests with classical unconditional tests.

REFERENCES

[1] M. Fagerland, S. Lydersen and P. Laake, *Statistical Analysis of Contingency Tables*. Chapman and Hall/CRC, 2017, ISBN: 9780367495268. [Online]. Available: https://contingencytables.com/.

[2] K. Pearson, '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*, vol. 50, no. 302, pp. 157–175, 1900. DOI: 10.1080/14786440009463897. [Online]. Available: https://doi.org/10.1080/14786440009463897 (visited on 19/01/2023).

[3] F. Yates, 'Tests of significance for 2 × 2 contingency tables,' *Journal of the Royal Statistical Society: Series A (General)*, vol. 147, no. 3, pp. 426–449, 1984. DOI: https://doi.org/10.2307/2981577. [Online]. Available: https://rss.onlinelibrary.wiley.com/doi/abs/10.2307/2981577.

[4] S. Lydersen, M. Langaas and Ø. Bakke, 'The exact unconditional z-pooled test for equality of two binomial probabilities: Optimal choice of the berger and boos confidence coefficient,' *Journal of Statistical Computation and Simulation*, vol. 82, no. 9, pp. 1311–1316, 2012. DOI: 10.1080/00949655.2011.579969. [Online]. Available: https://doi.org/10.1080/00949655.2011.579969.

[5] F. L. Aanes, 'Testing equality of the success probabilities in two independent binomial distributions,' M.S. thesis, Norwegian University of Science and Technology, 2016.

[6] F. van der Meulen, *Statistical inference*, 2021. [Online]. Available: https://github.com/fmeulen/fmeulen.github.io/blob/main/ln-wi4455.pdf (visited on 26/02/2023).

[7] G. Casella and R. Berger, *Statistical Inference* (Duxbury advanced series in statistics and decision sciences). Thomson Learning, 2002, ISBN: 9780534243128. [Online]. Available: https://books.google.no/books?id=0x%5C_vAAAAMAAJ.

[8] E. H. Brataas, 'Barnards csm test,' BA thesis, Norwegian University of Science and Technology, 2020.

[9] R. Fisher, *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd, 1934.

[10] F. Yates, 'Contingency tables involving small numbers and the χ^2 test,' *Supplement to the Journal of the Royal Statistical Society*, vol. 1, no. 2, pp. 217–235, 1934.

[11] A. Agresti, 'A survey of exact inference for contingency tables,' *Statistical science*, vol. 7, no. 1, pp. 131–153, 1992.

[12] D. V. Mehrotra, I. S. F. Chan and R. L. Berger, 'A cautionary note on exact unconditional inference for a difference between two independent binomial proportions,' *Biometrics*, vol. 59, no. 2, pp. 441–450, 2003, ISSN: 0006341X, 15410420. [Online]. Available: http://www.jstor.org/stable/3695522 (visited on 14/02/2023).

[13] G. A. Barnard, 'Significance tests for 2 × 2 tables,' *Biometrika*, vol. 34, no. 1/2, pp. 123–138, 1947, ISSN: 00063444. [Online]. Available: http://www.jstor.org/stable/2332517 (visited on 15/01/2023).

[14] G. Barnard, 'In contradiction to j. berkson's dispraise: Conditional tests can be more efficient,' *Journal of Statistical Planning and Inference*, vol. 3, no. 3, pp. 181–187, 1979, ISSN: 0378-3758. DOI: https://doi.org/10.1016/0378-3758(79)90009-0. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0378375879900090.

[15] S. Suissa and J. J. Shuster, 'Exact unconditional sample sizes for the 2 × 2 binomial trial,' *Journal of the Royal Statistical Society. Series A (General)*, vol. 148, no. 4, pp. 317–327, 1985, ISSN: 00359238. [Online]. Available: http://www.jstor.org/stable/2981892 (visited on 20/03/2023).

[16] R. D. Boschloo, 'Raised conditional level of significance for the 2 × 2-table when testing the equality of two probabilities,' *Statistica Neerlandica*, vol. 24, no. 1, pp. 1–9, 1970. DOI: https://doi.org/10.1111/j.1467-9574.1970.tb00104.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9574.1970.tb00104.x.

[17] L. L. MacDonald, B. M. Davis and G. A. Milliken, 'A nonrandomized unconditional test for comparing two proportions in 2×2 contingency tables,' *Technometrics*, vol. 19, no. 2, pp. 145–157, 1977. DOI: 10.1080/00401706.1977.10489522. [Online]. Available: https://doi.org/10.1080/00401706.1977.10489522.

[18] G. G. Crans and J. J. Shuster, 'How conservative is fisher's exact test? a quantitative evaluation of the two-sample comparative binomial trial,' *Statistics in Medicine*, vol. 27, no. 18, pp. 3598–3611, 2008. DOI: https://doi.org/10.1002/sim.3221. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.3221.

[19] D. Basu, 'On the elimination of nuisance parameters,' *Journal of the American Statistical Association*, vol. 72, no. 358, pp. 355–366, 1977, ISSN: 01621459. [Online]. Available: http://www.jstor.org/stable/2286800 (visited on 31/01/2023).

[20] G. A. Barnard, 'A new test for 2× 2 tables,' *Nature*, vol. 156, no. 3954, p. 177, 1945.

[21] R. A. Fisher, 'A new test for 2× 2 tables,' *Nature*, vol. 156, no. 3961, p. 388, 1945.

[22] G. A. Barnard, 'A new test for 2× 2 tables,' *Nature*, vol. 156, no. 3974, p. 784, 1945.

[23] E. S. Pearson, 'The choice of statistical tests illustrated on the interpretation of data classed in a 2 x 2 table.,' *Biometrika*, vol. 34 1-2, pp. 139–169, 1947.

[24] M. Gail and J. J. Gart, 'The determination of sample sizes for use with the exact conditional test in 2 x 2 comparative trials,' *Biometrics*, vol. 29, no. 3, pp. 441–448, 1973, ISSN: 0006341X, 15410420. [Online]. Available: http://www.jstor.org/stable/2529167 (visited on 19/02/2023).

[25] R. B. D'Agostino, W. Chase and A. Belanger, 'The appropriateness of some common procedures for testing the equality of two independent binomial populations,' *The American Statistician*, vol. 42, no. 3, pp. 198–202, 1988, ISSN: 00031305. [Online]. Available: http://www.jstor.org/stable/2685002 (visited on 19/02/2023).

[26] G. A. Barnard, 'The meaning of a significance level,' *Biometrika*, vol. 34, no. 1/2, pp. 179–182, 1947.

[27] G. A. Barnard, 'Statistical inference,' *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 11, no. 2, pp. 115–149, 1949.

[28] R. A. Fisher, 'Theory of statistical estimation,' *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 22, no. 5, pp. 700–725, 1925. DOI: 10.1017/S0305004100009580.

[29] M. Ghosh, N. Reid and D. A. S. Fraser, 'Ancillary statistics: A review,' *Statistica Sinica*, vol. 20, no. 4, pp. 1309–1332, 2010, ISSN: 10170405, 19968507. [Online]. Available: http://www.jstor.org/stable/24309506 (visited on 19/05/2023).

[30] N. Reid, 'Ancillary statistics,' in *Encyclopedia of Biostatistics*. John Wiley & Sons, Ltd, 2005, ISBN: 9780470011812. DOI: https://doi.org/10.1002/0470011815.b2a15002. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/0470011815.b2a15002.

[31] R. J. A. Little, 'Testing the equality of two independent binomial proportions,' *The American Statistician*, vol. 43, no. 4, pp. 283–288, 1989.

[32] J. D. Kalbfleisch and D. A. Sprott, 'Marginal and conditional likelihoods,' *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, vol. 35, no. 3, pp. 311–328, 1973, ISSN: 0581572X. [Online]. Available: http://www.jstor.org/stable/25049882 (visited on 05/04/2023).

[33] H. I. Chinn, W. K. Noell and P. K. Smith, 'Prophylaxis of motion sickness: Evaluation of some drugs in seasickness,' *AMA Archives of Internal Medicine*, vol. 86, no. 6, pp. 810–822, 1950.

[34] J. Berkson, 'In dispraise of the exact test: Do the marginal totals of the 2x2 table contain relevant information respecting the table proportions?' *Journal of Statistical Planning and Inference*, vol. 2, no. 1, pp. 27–42, 1978, ISSN: 0378-3758. DOI: https://doi.org/10.1016/0378-3758(78)90019-8. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0378375878900198.

[35] R. A. Fisher, 'The logic of inductive inference,' *Journal of the Royal Statistical Society*, vol. 98, no. 1, pp. 39–82, 1935, ISSN: 09528385. [Online]. Available: http://www.jstor.org/stable/2342435 (visited on 27/02/2023).

[36] J. Berkson, 'Do the marginal totals of the 2x2 table contain relevant information respecting the table proportions?' *Journal of Statistical Planning and Inference*, vol. 2, no. 1, pp. 43–44, 1978, ISSN: 0378-3758. DOI: https://doi.org/10.1016/0378-3758(78)90020-4. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0378375878900204.

[37] R. L. Plackett, 'The marginal totals of a 2 × 2 table,' *Biometrika*, vol. 64, no. 1, pp. 37–42, 1977, ISSN: 00063444. [Online]. Available: http://www.jstor.org/stable/2335767 (visited on 28/02/2023).

[38] K. D. Tocher, 'Extension of the neyman-pearson theory of tests to discontinuous variates,' *Biometrika*, vol. 37, no. 1/2, pp. 130–144, 1950, ISSN: 00063444. [Online]. Available: http://www.jstor.org/stable/2332156 (visited on 28/02/2023).

[39] E. Lehmann and J. Romano, *Testing Statistical Hypotheses* (Springer Texts in Statistics), 3rd ed. New York: Springer, 2005, ISBN: 0387988645.

[40] S. Suissa and J. J. Shuster, 'Are uniformly most powerful unbiased tests really best?' *The American Statistician*, vol. 38, no. 3, pp. 204–206, 1984.

[41] A. Birnbaum, 'On the foundations of statistical inference,' *Journal of the American Statistical Association*, vol. 57, no. 298, pp. 269–306, 1962. DOI: 10.1080/01621459.1962.10480660. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/01621459.1962.10480660.

[42] J. Durbin, 'On birnbaum's theorem on the relation between sufficiency, conditionality and likelihood,' *Journal of the American Statistical Association*, vol. 65, no. 329, pp. 395–398, 1970, ISSN: 01621459. [Online]. Available: http://www.jstor.org/stable/2283601 (visited on 01/05/2023).

[43] J. D. Kalbfleisch, 'Sufficiency and conditionality,' *Biometrika*, vol. 62, no. 2, pp. 251–259, 1975.

[44] S. Wechsler, C. A. de B. Pereira and P. C. Marques F, 'Birnbaum's theorem redux,' in *AIP Conference Proceedings*, American Institute of Physics, vol. 1073, 2008, pp. 96–100.

[45] J. Grossman, 'The likelihood principle,' in *Philosophy of Statistics*, ser. Handbook of the Philosophy of Science, P. S. Bandyopadhyay and M. R. Forster, Eds., vol. 7, Amsterdam: North-Holland, 2011, pp. 553–580. DOI: https://doi.org/10.1016/B978-0-444-51862-0.50017-4. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780444518620500174.

[46] M. Evans, 'What does the proof of Birnbaum's theorem prove?' *Electronic Journal of Statistics*, vol. 7, pp. 2645–2655, 2013. DOI: 10.1214/13-EJS857. [Online]. Available: https://doi.org/10.1214/13-EJS857.

[47] D. G. Mayo, 'On the Birnbaum Argument for the Strong Likelihood Principle,' *Statistical Science*, vol. 29, no. 2, pp. 227–239, 2014. DOI: 10.1214/13-STS457. [Online]. Available: https://doi.org/10.1214/13-STS457.

[48] V. Peña and J. O. Berger, *A note on recent criticisms to birnbaum's theorem*, 2017. arXiv: 1711.08093 [math.ST].

[49] I. S. Helland, 'Simple counterexamples against the conditionality principle,' *The American Statistician*, vol. 49, no. 4, pp. 351–356, 1995.

[50] M. Evans, D. A. S. Fraser and G. Monette, 'On regularity for statistical models,' *Canadian Journal of Statistics*, vol. 13, no. 2, pp. 137–144, 1985. DOI: https://doi.org/10.2307/3314876. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.2307/3314876.

[51] M. Evans, D. A. S. Fraser and G. Monette, 'On principles and arguments to likelihood,' *Canadian Journal of Statistics*, vol. 14, no. 3, pp. 181–194, 1986. DOI: https://doi.org/10.2307/3314794. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.2307/3314794.

[52] S. Greenland, 'On the logical justification of conditional tests for two-by-two contingency tables,' *The American Statistician*, vol. 45, no. 3, pp. 248–251, 1991.

[53] O. Kempthorne, 'In dispraise of the exact test: Reactions,' *Journal of Statistical Planning and Inference*, vol. 3, no. 3, pp. 199–213, 1979, ISSN: 0378-3758. DOI: https://doi.org/10.1016/0378-3758(79)90012-0. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0378375879900120.

[54] G. J. G. Upton, 'A comparison of alternative tests for the 2 times 2 comparative trial,' *Journal of the Royal Statistical Society: Series A (General)*, vol. 145, no. 1, pp. 86–105, 1982. DOI: https://doi.org/10.2307/2981423. [Online]. Available: https://rss.onlinelibrary.wiley.com/doi/abs/10.2307/2981423.

[55] C. R. Mehta and N. R. Patel, 'A network algorithm for performing fisher's exact test in r × c contingency tables,' *Journal of the American Statistical Association*, vol. 78, no. 382, pp. 427–434, 1983, ISSN: 01621459. [Online]. Available: http://www.jstor.org/stable/2288652 (visited on 22/03/2023).

[56] G. Barnard, 'Discussion on yates' tests of significance for 2 × 2 contingency tables,' *Journal of the Royal Statistical Society: Series A (General)*, vol. 147, no. 3, pp. 449–450, 1984.

[57] D. Cox, 'Discussion on yates' tests of significance for 2 × 2 contingency tables,' *Journal of the Royal Statistical Society: Series A (General)*, vol. 147, no. 3, p. 451, 1984.

[58] G. A. Barnard, 'On alleged gains in power from lower p-values,' *Statistics in Medicine*, vol. 8, no. 12, pp. 1469–1477, 1989. DOI: https://doi.org/10.1002/sim.4780081206. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780081206.

[59] G. J. G. Upton, 'Fisher's exact test,' *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 155, no. 3, pp. 395–402, 1992.

[60] G. Camilli, 'The test of homogeneity for 2× 2 contingency tables: A review of and some personal opinions on the controversy.,' *Psychological Bulletin*, vol. 108, no. 1, p. 135, 1990.

[61] W. R. Rice, 'A new probability model for determining exact p-values for 2 x 2 contingency tables when comparing binomial proportions,' *Biometrics*, vol. 44, no. 1, pp. 1–22, 1988, ISSN: 0006341X, 15410420. [Online]. Available: http://www.jstor.org/stable/2531892 (visited on 06/04/2023).

[62] R. D. Routledge, 'Resolving the conflict over fisher's exact test,' *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, vol. 20, no. 2, pp. 201–209, 1992, ISSN: 03195724. [Online]. Available: http://www.jstor.org/stable/3315468 (visited on 18/01/2023).

[63] R. Fisher, *Statistical Methods and Scientific Inference*. Hafner Press, 1973, ISBN: 9780028447407. [Online]. Available: https://books.google.no/books?id=IHdqAAAAMAAJ.

[64] J. Ludbrook, 'Is there still a place for pearson's chi-squared test and fisher's exact test in surgical research?' *ANZ journal of surgery*, vol. 81, no. 12, pp. 923–926, 2011.

[65] R. R. Macdonald, 'Conditional and unconditional tests of association in 2 × 2 tables,' *British Journal of Mathematical and Statistical Psychology*, vol. 51, no. 2, pp. 191–204, 1998. DOI: https://doi.org/10.1111/j.2044-8317.1998.tb00676.x. [Online]. Available: https://bpspsychub.onlinelibrary.wiley.com/doi/abs/10.1111/j.2044-8317.1998.tb00676.x.

[66] J. J. Shuster, S. Greenland, D. P. Nichols, A. Cohen and J. R. Lackritz, 'Letters to the editor,' *The American Statistician*, vol. 46, no. 2, pp. 163–165, 1992, ISSN: 00031305. [Online]. Available: http://www.jstor.org/stable/2684190 (visited on 22/05/2023).

[67] E. Ripamonti, C. Lloyd and P. Quatto, 'Contemporary frequentist views of the 2× 2 binomial trial,' *Statistical science*, pp. 600–615, 2017.

[68] D. Basu, 'Discussion of joseph berkson's paper "in dispraise of the exact test",' *Journal of Statistical Planning and Inference*, vol. 3, no. 3, pp. 189–192, 1979, ISSN: 0378-3758. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0378375879900107.

[69] E. Benhamou and V. Melot, 'Seven proofs of the pearson chi-squared independence test and its graphical interpretation,' *arXiv preprint arXiv:1808.09171*, 2018.

[70] W. G. Cochran, 'Some methods for strengthening the common χ^2 tests,' *Biometrics*, vol. 10, no. 4, pp. 417–451, 1954, ISSN: 0006341X, 15410420. [Online]. Available: http://www.jstor.org/stable/3001616 (visited on 22/03/2023).

[71] G. H. Freeman and J. H. Halton, 'Note on an exact treatment of contingency, goodness of fit and other problems of significance,' *Biometrika*, vol. 38, no. 1/2, pp. 141–149, 1951, ISSN: 00063444. [Online]. Available: http://www.jstor.org/stable/2332323 (visited on 22/03/2023).

[72] K. Berry, J. Johnston and P. Mielke, *A Chronicle of Permutation Statistical Methods: 1920–2000, and Beyond*. Springer International Publishing, 2014, pp. 287–297, ISBN: 9783319027449. [Online]. Available: https://books.google.no/books?id=BX3BBAAAQBAJ.

[73] L. Devroye, *Non-Uniform Random Variate Generation*. Springer New York, 2013, ISBN: 9781461386438. [Online]. Available: https://books.google.no/books?id=9tLcBwAAQBAJ.

[74] J. J. Goeman and A. Solari, 'Multiple hypothesis testing in genomics,' *Statistics in medicine*, vol. 33, no. 11, pp. 1946–1978, 2014.

[75] A. M. Andrés and A. S. Mato, 'Choosing the optimal unconditioned test for comparing two independent proportions,' *Computational Statistics & Data Analysis*, vol. 17, no. 5, pp. 555–574, 1994, ISSN: 0167-9473. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0167947394901481.

[76] M. Haber, 'A comparison of some conditional and unconditional exact tests for 2x2 contingency tables,' *Communications in Statistics - Simulation and Computation*, vol. 16, no. 4, pp. 999–1013, 1987. DOI: 10.1080/03610918708812633. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/03610918708812633.

[77] R. L. Berger and D. D. Boos, 'P values maximized over a confidence set for the nuisance parameter,' *Journal of the American Statistical Association*, vol. 89, no. 427, pp. 1012–1016, 1994, ISSN: 01621459. [Online]. Available: http://www.jstor.org/stable/2290928 (visited on 31/05/2023).

[78] H. Niederreiter, *Random Number Generation and Quasi-Monte Carlo Methods*. Society for Industrial and Applied Mathematics, 1992. DOI: 10.1137/1.9781611970081. [Online]. Available: https://epubs.siam.org/doi/abs/10.1137/1.9781611970081.

[79] E. Hlawka, 'Funktionen von beschränkter variatiou in der theorie der gleichverteilung,' *Annali di Matematica Pura ed Applicata*, vol. 54, no. 1, pp. 325–333, 1961.

[80] W. Schmidt, 'Irregularities of distribution, vii,' *Acta Arithmetica*, vol. 21, no. 1, pp. 45–50, 1972.

[81] I. M. Sobol', 'On the distribution of points in a cube and the approximate evaluation of integrals,' *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, vol. 7, no. 4, pp. 784–802, 1967.

[82] A. Savine, *Modern Computational Finance*. John Wiley & Sons, Ltd, 2018, ISBN: 9781119539452. [Online]. Available: https://www.wiley.com/en-us/Modern+Computational+Finance:+AAD+and+Parallel+Simulations-p-9781119539452.

[83] P. Glasserman, *Monte Carlo methods in financial engineering*. Springer, 2004, vol. 53, ISBN: 9780387216171. [Online]. Available: https://link.springer.com/book/10.1007/978-0-387-21617-1.

[84] H. Niederreiter, 'Methods for estimating discrepancy,' in *Applications of Number Theory to Numerical Analysis*, S. Zaremba, Ed., Academic Press, 1972, pp. 203–236, ISBN: 978-0-12-775950-0. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B978012775950050011X.

APPENDICES

APPENDIX

A

LOW-DISCREPANCY SEQUENCES

In Sections 4.3.1 and 4.3.2, we made use of low-discrepancy sequences in order to create a deterministic grid over Θ_0 that would not necessarily scale with the number of columns of the contingency table being considered. We could then use this grid to compute and compare the maximal values of sums of $P(\mathbf{x};\boldsymbol{\theta})$-functions in the case of CSM-like tests, or use it to construct the matrix A for the LP tests.

Probably the most intuitive way in which low-discrepancy sequences are introduced is via numerical integration. We follow the treatment in the book of Niederreiter [78] here. When evaluating the integral $\int_0^1 f(s)\mathrm{d}s$, a classical approach is to construct an equidistant grid $\{i/N : i = 0, \ldots, N\}$ of $N+1$ points over the interval $[0, 1]$ and use a method like the midpoint rule to obtain the approximation

$$\int_0^1 f(s)\mathrm{d}s \approx \sum_{i=1}^{N} \frac{1}{N} f\left(\frac{i-\frac{1}{2}}{N}\right).$$

It is easy to show that this approximation has an error of order $\mathcal{O}(N^{-2})$. When generalising this rule to d dimensions, say to approximate the integral $\int_{[0,1]^d} f(\mathbf{s})\mathrm{d}\mathbf{s}$, the equidistant grid will consist of $(N+1)^d$ grid points instead, and one can show that the corresponding approximation will be of order $\mathcal{O}(N^{-2/d})$. For large d, this quickly becomes a rather useless error bound. This inherent weakness of the classical integration methods is often referred to as the *curse of dimensionality*.

To avoid this issue, one can use the Monte Carlo method of integration instead. The idea here is to regard the integral $\int_{[0,1]^d} f(\mathbf{s})\mathrm{d}\mathbf{s}$ as an expected value. Indeed, for $\mathbf{U} = (U_1, \ldots, U_d) \in [0, 1]^d$ a random vector of uniform random variables,

$$\mathbb{E}[f(\mathbf{U})] = \int_{[0,1]^d} f(\mathbf{u})\mathrm{d}\mathbf{u}.$$

However, we can approximate the expected value $\mathbb{E}[f(\mathbf{U})]$ by taking N independent uniform random samples $\mathbf{u}^1, \ldots, \mathbf{u}^N$ and computing their average $1/N \sum_{i=1}^{N} f(\mathbf{u}^i)$. This approximation is bound to converge to the actual expected value, as by the Strong Law of Large Numbers,

$$\frac{1}{N}\sum_{i=1}^{N} f(\mathbf{u}^i) \xrightarrow{\text{a.s.}} \mathbb{E}[f(\mathbf{U})],$$

as $N \to \infty$. Based on this, we can approximate our integral by

$$\int_{[0,1]^d} f(\mathbf{s})\mathrm{d}\mathbf{s} \approx \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{u}^i). \tag{A.1}$$

One can prove using the Central Limit Theorem that the average error of this approximation is $\mathcal{O}(N^{-1/2})$ [78]. The big advantage of the Monte Carlo approach is thus that the error does not longer depend on the number of dimensions d. Thus, if one is able to generate independent random samples, the Monte Carlo integration method is definitely an attractive alternative to the classical integration rules for $d \geq 5$. Keep in mind however that because we are taking random samples, we do not have a guaranteed error bound as with the classical (deterministic) methods. Some samples might yield a worse approximation, and some might yield a better approximation.

A.1 Quasi-Monte Carlo integration

What if we would be able to pick samples which yield a better approximation? This is where the Quasi Monte Carlo method of integration comes in. Instead of sampling the points at which we evaluate f in (A.1) at random, we can construct a specific *deterministic* sequence $(\mathbf{v}^1, \ldots, \mathbf{v}^N)$ of points for which the approximation error of (A.1) is smaller than this average $\mathcal{O}(N^{-1/2})$. The resulting Quasi Monte Carlo approximation is given by

$$\int_{[0,1]^d} f(\mathbf{s})\mathrm{d}\mathbf{s} \approx \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{v}^i). \tag{A.2}$$

Thus, if we are able to construct a "good" sequence of points, we have a method of numerical integration that just as the Monte Carlo approach does not suffer from the curse of dimensionality, but has a smaller approximation error. Furthermore, since the sequence is completely deterministic, we can come up with an actual error bound, and not some average statement on the approximation error.

The question we thus want to answer is how to construct a "good" sequence of points. Of course, the larger we make N, the better we expect the approximation (A.2) to be. That is, we would like the sequence $(\mathbf{v}^i)_{i\in\mathbb{N}}$ to be *uniformly distributed* on $[0,1]^d$, in the sense that

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{v}^i) = \int_{[0,1]^d} f(\mathbf{s})\mathrm{d}\mathbf{s},$$

for a class of functions f. Niederreiter requires f to be continuous on $[0,1]^d$. Equivalently, one can say that $(\mathbf{v}^i)_{i\in\mathbb{N}}$ is uniformly distributed if for all subintervals $I \subset [0,1]^d$,

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{\{\mathbf{v}^i \in I\}} = \lambda_d(I),$$

where λ_d is the d-dimensional Lebesgue measure. To quantify how evenly the point set $\left\{\mathbf{v}^1, \ldots, \mathbf{v}^N\right\}$ is distributed, we can define its discrepancy.

Definition A.1. For a point set $\{\mathbf{v}^1, \ldots, \mathbf{v}^N\} \subset [0,1]^d$ and for a collection of Lebesgue-measurable subsets $\mathcal{A} \in \mathcal{P}([0,1]^d)$, we define its *discrepancy* relative to $\mathcal{A}$ by

$$\mathcal{D}_N(\mathbf{v}^1, \ldots, \mathbf{v}^N; \mathcal{A}) := \sup_{A \in \mathcal{A}} \left| \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{\mathbf{v}^i \in A\}} - \lambda_d(A) \right|. \tag{A.3}$$

The index N indicates the number of points in the point set. If we have a sequence $\mathcal{V} = (\mathbf{v}^i)_{i \in \mathbb{N}}$, we denote by $\mathcal{D}_N(\mathcal{V}; \mathcal{A})$ the discrepancy relative to $\mathcal{A}$ of the first N elements of $\mathcal{V}$. There exist two common choices for the collection $\mathcal{A}$. First is the ordinary discrepancy $\mathcal{D}_N(\mathbf{v}^1, \ldots, \mathbf{v}^N) := \mathcal{D}_N(\mathbf{v}^1, \ldots, \mathbf{v}^N; \mathcal{A})$, with $\mathcal{A}$ the collection of all rectangles in $[0,1)^d$, i.e., all subintervals of $[0,1)^d$ of the form $\prod_{j=1}^d [s_j, t_j)$. The second one is the star discrepancy $\mathcal{D}_N^*(\mathbf{v}^1, \ldots, \mathbf{v}^N) := \mathcal{D}_N(\mathbf{v}^1, \ldots, \mathbf{v}^N; \mathcal{A})$, where $\mathcal{A}$ is the collection of all rectangles with the origin as their bottom-left corner, i.e., all subintervals of $[0,1)^d$ of the form $\prod_{j=1}^d [0, s_j)$. One can show that the ordinary discrepancy and star discrepancy have the same order of magnitude (see Proposition 2.4 in Niederreiter [78]).

The smaller the discrepancy of a point set, the more evenly distributed it is over $[0,1]^d$, and the better the approximation (A.2) will be. This idea is formalised via the following result.

Theorem A.2 (Koksma-Hlawka inequality)**.** Suppose that f has bounded Hardy-Krause variation $V(f)$ on $[0,1)^d$. Then, for any $\mathbf{v}^1, \ldots, \mathbf{v}^N \in [0,1)^d$,

$$\left| \frac{1}{N} \sum_{i=1}^{N} f(\mathbf{v}^i) - \int_{[0,1]^d} f(\mathbf{s}) \mathrm{d}\mathbf{s} \right| \leq V(f) \mathcal{D}_N^*(\mathbf{v}^1, \ldots, \mathbf{v}^N). \tag{A.4}$$

Proof of Theorem A.2. See the proof by Hlawka [79] (in German). □

The Hardy-Krause variation $V(f)$ is a rather involved generalisation of the total variation to functions of multiple variables. We refer for its definition to Niederreiter [78]. This complicated definition makes the actual error bound hard to compute, but nevertheless (A.4) should indicate that the Quasi Monte Carlo approach can certainly be effective if we manage to find sequences with a small star discrepancy. We call a sequence $\mathcal{V}$ a *low-discrepancy sequence* if it has for all $N \in \mathbb{N}$ a "small" value of $\mathcal{D}_N(\mathcal{V})$ (and thus of $\mathcal{D}_N^*(\mathcal{V})$ too).

A.2 Examples of low-discrepancy sequences

Let us first give a very important example of a one-dimensional low-discrepancy sequence: the van der Corput sequence. Although its applicability is limited in one dimension, since classical integration methods outperform (Quasi) Monte Carlo methods for $d = 1$, it is fundamental for the construction of many higher-dimensional low-discrepancy sequences.

Consider an integer $b \leq 2$. Then every integer $k \geq 0$ has a b-ary expansion $\sum_{i=0}^{\infty} a_i(k) b^i$, where $a_i(k) \in \{0, \ldots, b-1\}$ for all $i \geq 0$ and eventually $a_i(k) = 0$, such that this expansion is actually a finite sum. The van der Corput sequence is constructed by "flipping this expansion around the decimal point". This is made more precise in the following definition.

Definition A.3. For an integer $b \geq 2$, we define the *van der Corput sequence in base b*, denoted by $\mathcal{C}_b$, as the sequence $(c_k)_{k\geq 0}$ where $c_k = \phi_b(k)$. The so-called *radical inverse function ϕ_b in base b* is defined as

$$\phi_b(k) = \sum_{i=0}^{\infty} a_i(k) b^{-i-1}.$$

It turns out that for all $N \geq 2$, $\mathcal{D}_N^*(\mathcal{C}_b) = \mathcal{O}(N^{-1}\log N)$ [78]. This is the best we can do in one dimension, as Schmidt [80] has shown that there exists a constant $c > 0$ such that, for any sequence $\mathcal{V}$ in $[0,1]$, $D_N(\mathcal{V}) \geq cN^{-1}\log N$ for infinitely many N.

As an example, the first 9 terms of $\mathcal{C}_3$ are given in Table A.1, as well as an illustration of how to compute them. Intuitively, it seems that the first N terms of $\mathcal{C}_3$ try to fill up $[0,1]$ as evenly as possible, no matter at which value of N we decide to stop the sequence.

k	k (base 3)	$\phi_3(k)$ (base 3)	$c_k = \phi_3(k)$
0	0	0	0
1	1	0.1	1/3
2	2	0.2	2/3
3	10	0.01	1/9
4	11	0.11	4/9
5	12	0.21	7/9
6	20	0.02	2/9
7	21	0.12	5/9
8	22	0.22	8/9

Table A.1: Construction of the first 9 terms of $\mathcal{C}_3$.

As we mentioned earlier, we can use the van der Corput sequence to construct higher-dimensional low-discrepancy sequences. The most straightforward example of this is the so-called Halton family of sequences. Given $d \geq 1$, we choose integers $b_1, \ldots, b_d \geq 2$ and define the *Halton sequence in the bases $b_1, \ldots, b_d$*, denoted by $\mathcal{H}_{b_1,\ldots,b_d}$, as the sequence $(h_k)_{k\geq 0}$ where, for all $k \geq 0$,

$$h_k = (\phi_{b_1}(k), \ldots, \phi_{b_d}(k)).$$

Niederreiter [78] showed that when choosing the $b_1, \ldots, b_d \geq 2$ relatively prime,

$$D_N^*(\mathcal{H}_{b_1,\ldots,b_d}) \leq K_1(b_1,\ldots,b_d)N^{-1}(\log N)^d + \mathcal{O}(N^{-1}(\log N)^{d-1}), \tag{A.5}$$

for all $N \geq 2$, and for some constant $K_1(b_1,\ldots,b_d)$ only dependent on the bases. It is believed that $D_N^*(\mathcal{H}_{b_1,\ldots,b_d}) \geq K_2(d)N^{-1}(\log N)^d$ for infinitely many N and a constant[78], which would also mean that a Halton sequence with relatively prime bases is the best we can do regarding the order of magnitude of the star discrepancy. However, the constant $K_1(b_1,\ldots,b_d)$ can be shown to grow superexponentially as d increases. This makes the bound (A.5) not so informative for larger d. Indeed, it also turns out that for large dimensions, we will be forced to use van der Corput sequences with high bases, which have long monotone subsequences, leading to a large discrepancy.

As an alternative to the Halton sequence, Sobol' [81] introduced the concept of (t,m,d)-nets and (t,d)-sequences.

Definition A.4. Let $0 \leq t \leq m$ be integers and fix a base $b \geq 0$. A (t, m, d)-*net in base* b is a set of b^m points in $[0,1)^d$ such that every b-ary box, i.e., each interval of the form

$$B = \prod_{j=1}^{d} \left[\frac{a_j}{b^{k_j}}, \frac{a_j + 1}{b^{k_j}} \right),$$

with $k_j \geq 0$ and $a_j \in \left\{0, \ldots, b^{k_j} - 1\right\}$ integer for all $j \in \{1, \ldots, d\}$ which has a volume $\lambda_d(B) = b^{t-m}$ contains exactly b^t points.

Definition A.5. Let $t \geq 0$ be an integer and fix a base $b \geq 0$. A sequence $(\mathbf{v}^i)_{i \geq 0} \subset [0,1)^d$ is a (t, d)-*sequence in base* b, denoted $\mathcal{S}_{t,d}$, if for all integers $k \geq 0$ and $m > t$, the point set $\{\mathbf{v}_i : kb^m < i \leq (k+1)b^m\}$ is a (t, m, d)-net in base b.

The intuition behind these definitions is that for any point set taken from a (t, d)-sequence, each b-ary box of a certain volume contain an "expected" number of points. The smaller t, the smaller boxes this will hold for and so the more evenly distributed the sequence points will be. Moreover, if we feel that a point set of b^m points from a (t, d)-sequence still leaves a number of gaps in $[0,1)^d$, the next b^m points of the sequence will tend to fill those gaps. Niederreiter shows that (t, d)-sequences are low-discrepancy sequences with star discrepancy

$$D_N^*(\mathcal{S}_{t,d}) \leq K_3(b, d) b^t N^{-1} (\log N)^d + \mathcal{O}(b^t N^{-1} (\log N)^{d-1}),$$

where the constant can be shown to grow far less quickly than $K_1(b_1, \ldots, b_d)$. One important example of a (t, d)-sequence (in base 2) is the Sobol' sequence. Its construction goes far beyond the scope of this thesis, and is described in detail in Chapter 4 of Niederreiter [78]. Other nice explanations can be found in Section 5.4 of Savine [82] and Section 5.2.3 of Glasserman [83]. Because one can use bit-level operations to build up the sequence, its construction knows very efficient implementations. Although Niederreiter [78] mentions a couple of ways to obtain theoretically better performing sequences, the Sobol' sequence seems to be go-to low-discrepancy sequence nowadays when d gets large, outperforming most other low-discrepancy sequences in many numerical experiments [83].

Despite the attractiveness of the Sobol' sequence, we still opted in this text to exclusively use another low-discrepancy sequence; the Torus sequence. This has been mentioned briefly in Chapter 6. Essentially, although the Sobol' sequence can be implemented efficiently, elements from the Torus sequence can still be generated more quickly. Furthermore, even though this thesis was focused on larger tables with multiple columns and thus required the generation of low-discrepancy sequences in multiple dimensions, we never needed to consider the high-dimensional regime in which the performance of the Torus sequence degrades very quickly, and in which the Sobol' sequence's superiority really pays off. In the dimensions we considered, i.e. $c \in \{2, 3, 4\}$, tests using the Torus sequence behaved very similarly to those using the Sobol' sequence – when used with a reasonable amount of grid points – and we thus opted to choose the sequence with the fastest generation. If one would like to use low-discrepancy sequences to test contingency tables with a higher number of columns, it would be worthwhile using a low-discrepancy sequence that is known to perform well in high dimensions.

The Torus sequence is a special case of the so-called Kronecker family of sequences, which are obtained by taking the fractional parts of multiples of certain

irrational numbers. That is, for an irrational number z, the corresponding Kronecker sequence $\mathcal{K}_z = (\kappa^i)_{i\geq 0} \subset [0,1)$ is defined by

$$\kappa^i = iz - \lfloor iz \rfloor,$$

for all $i \geq 0$. In higher dimensions, this easily generalises to the sequence $(\boldsymbol{\kappa}^i)_{i\geq 0} \subset [0,1)^d$, with for all $i \geq 0$,

$$\boldsymbol{\kappa}^i = (iz_1 - \lfloor iz_1 \rfloor, \ldots, iz_d - \lfloor iz_d \rfloor).$$

One should however additionally require that $1, z_1, \ldots, z_d$ are linearly independent over $\mathbb{Q}$ [78]. The Torus sequence can be constructed by choosing $(z_1, \ldots, z_d) = (\sqrt{p_1}, \ldots, \sqrt{p_d})$, where $p_1, \ldots, p_d$ are prime numbers. Note that the $\sqrt{p_1}, \ldots, \sqrt{p_d}$ are algebraic numbers: $\sqrt{p_j}$ is the root of the rational-coefficient polynomial $x^2 - p_j$. Niederreiter has shown that therefore, $\mathcal{D}_N(\mathcal{K}_{\sqrt{p_1},\ldots,\sqrt{p_d}}) = \mathcal{O}(N^{-1+\varepsilon})$, for all $N \geq 0$ and all $\varepsilon > 0$ [84].

A.3 Quasi Monte Carlo optimisation

So far, we have been talking a lot about how low-discrepancy sequences appear to be a useful tool for numerical integration. However, recall that we are not using them for that purpose at all. Instead, we use them in the CSM tests for finding the maxima of the sums of $P(\mathbf{x};\theta)$-functions, and in the LP tests we look at these sequence points whether or not the probability of a Type I error is still under the level α. In both cases, we can view this as a problem of finding the maximum of the power function $\beta(\cdot)$

$$m(\beta) := \sup_{\boldsymbol{\theta}\in\Theta_0} \beta(\boldsymbol{\theta}) \tag{A.6}$$

for a suited level α, by choosing a set of points $\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^N$ in Θ_0, evaluating $\beta(\cdot)$ at those points and then approximating the maximum $m(\beta)$ by

$$m(\beta) \approx m_N(\beta; \boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^N) := \max_{1\leq i\leq N} \beta(\boldsymbol{\theta}^i).$$

This is inspired on the so-called random search; sampling random points and evaluating the function at these points to approximate the maximum. This Monte Carlo approach works as one can show that for continuous functions f defined on S and a sequence $(\mathbf{u}^i)_{i\geq 1} \subset S$ of independent random samples,

$$m_N(f; \mathbf{u}^1, \ldots, \mathbf{u}^N) \xrightarrow{\text{a.s.}} m(f).$$

Just as with numerical integration, the Quasi Monte Carlo alternative to this random search is to use a deterministic (reproducible) sequence $(\mathbf{v}^i)_{i\geq 1}$ instead of a random one to approximate $m(f)$. For continuous f and a sequence $(\mathbf{v}^i)_{i\geq 1}$ that is dense in S, $m_N(\beta; \mathbf{v}^1, \ldots, \mathbf{v}^N)$ will also converge to $m(f)$ as $N \to \infty$.

Where we needed sequences with small discrepancy, i.e., sequences that were evenly distributed, to make sure that (A.2) was a good approximation, we now need sequences that are dense in S. To determine how "dense" a sequence is, we can define a quantity playing a similar role to the discrepancy.

Definition A.6. Suppose (S, δ) is a bounded metric space. Then the *dispersion* of a point set $\left\{\mathbf{v}^1, \ldots, \mathbf{v}^N\right\} \subset S$ in S is defined by

$$\delta_N(\mathbf{v}^1, \ldots, \mathbf{v}^N; S) := \sup_{\mathbf{s} \in S} \min_{1 \leq i \leq N} \delta(\mathbf{s}, \mathbf{v}^i). \tag{A.7}$$

Niederreiter [78] showed that for a point set $\left\{\mathbf{v}^1, \ldots, \mathbf{v}^N\right\} \subset S$,

$$m(f) - m_N(f; \mathbf{v}^1, \ldots, \mathbf{v}^N) \leq \sup_{\substack{\mathbf{s},\mathbf{t} \in S \\ \delta(\mathbf{s},\mathbf{t}) \leq \delta_N(\mathbf{v}^1, \ldots, \mathbf{v}^N; S)}} |f(\mathbf{s}) - f(\mathbf{t})|. \tag{A.8}$$

This formalises the idea that low-dispersion sequences will yield good approximations to $m(f)$. Indeed, if f is uniformly continuous on S, the right-hand side of (A.8) will go to zero for sequences $(\mathbf{v}^i)_{i \geq 1}$ that have $\lim_{N \to \infty} \delta_N(\mathbf{v}^1, \ldots, \mathbf{v}^N; S) = 0$. Luckily for us, it turns out that on $[0, 1)^d$, a low-discrepancy sequence is automatically a low-dispersion sequence, since for any point set $\left\{\mathbf{v}^1, \ldots, \mathbf{v}^N\right\} \subset [0, 1]^d$,

$$\delta'_N(\mathbf{v}^1, \ldots, \mathbf{v}^N) \leq \left(\mathcal{D}_N(\mathbf{v}^1, \ldots, \mathbf{v}^N)\right)^{1/d}$$

[78]. Here $\delta'_N(\mathbf{v}^1, \ldots, \mathbf{v}^N) := \delta_N(\mathbf{v}^1, \ldots, \mathbf{v}^N; [0, 1]^d)$ is the dispersion defined on the metric space $([0, 1]^d, \delta')$, with δ' the maximum metric defined by

$$\delta\prime(\mathbf{s}, \mathbf{t}) := \max_{1 \leq j \leq d} |s_j - t_j|$$

for all $\mathbf{s}, \mathbf{t} \in [0, 1]^d$. Because this metric is equivalent with the Euclidean metric, this dispersion can be shown to be of the same order of magnitude as the dispersion defined with the Euclidean metric. In particular, we can use the Torus sequence discussed earlier to approximate $\sup_{\mathbf{s} \in [0,1]^d} f(\mathbf{s})$.

However, the attentive reader might notice that in (A.6) we are not maximising over $[0, 1]^d$, but over Θ_0, which is a $d-1$-dimensional simplex (with d the number of table columns in this case). We can call upon the work of Niederreiter [78] one last time, who shows that for a map $g : (S_1, \delta^1) \to (S_2, \delta^2)$ that is Lipschitz continuous, i.e., for which there exists a constant $L \geq 0$ such that for all $\mathbf{s}, \mathbf{t} \in S_1$,

$$\delta^2(g(\mathbf{s}), g(\mathbf{t})) \leq L\delta^1(\mathbf{s}, \mathbf{t}),$$

we have for any point set $\left\{\mathbf{v}^1, \ldots, \mathbf{v}^N\right\} \subset S_1$ that

$$\delta_N(g(\mathbf{v}^1), \ldots, g(\mathbf{v}^N); S_2) \leq L\delta_N(\mathbf{v}^1, \ldots, \mathbf{v}^N; S_1).$$

In the setting of (A.6), we have $S_1 = [0, 1]^c$ and $S_2 = \Theta_0$, both equipped with the Euclidean metric. The function $g : [0, 1]^c \to \Theta_0$ follows immediately from (4.16) and is given by:

$$g(s_1, \ldots, s_c) = \left(\frac{\log s_1}{\sum_{j=1}^c \log s_j}, \ldots, \frac{\log s_c}{\sum_{j=1}^c \log s_j}\right).$$

However, this function is not differentiable on the whole of $[0, 1]^c$. In particular, an easy computation shows that the partial derivatives at the boundary points can be infinite. Hence, we do not have Lipschitz continuity. However, we

still argue that for any $N \geq 1$, the transformation via g of any Torus point set $\left\{\mathbf{v}^1, \ldots, \mathbf{v}^N\right\} \subset [0,1]^c$ is also a low-dispersion point set on the simplex. This is because by construction, we in fact have $\left\{\mathbf{v}^1, \ldots, \mathbf{v}^N\right\} \subset (0,1)^c$. We can thus find some $\epsilon > 0$ for which $\left\{\mathbf{v}^1, \ldots, \mathbf{v}^N\right\} \subset [\epsilon, 1-\epsilon]^c$. On this reduced domain, g can easily be seen to be differentiable and thus Lipschitz continuous. Therefore, the point set $\left\{g(\mathbf{v}^1), \ldots, g(\mathbf{v}^N)\right\}$ is a low-discrepancy point set on $g([\epsilon, 1-\epsilon]^c)$. For small ϵ, $g([\epsilon, 1-\epsilon]^c)$ will cover a large enough part of Θ_0 such that we do not miss out on any potential maxima in the case of the supremum tests, and do not violate the size constraint in the case of the LP tests.

APPENDIX

B

GITHUB REPOSITORY

All methods implemented for this project can be found in the Github repository

`https://github.com/PimKeer/contingency_tables`.

The most important file of the project is `function.R`, containing all implementations of the different tests discussed in this thesis, as well as helper functions needed for these implementations. There are also a number of functions used for computation of the power function, and for visualisation of the critical region, size functions and power functions. The function `supremum_ordering()` can be called to run any of the supremum tests: the CSM tests, with one of the three symmetry conditions (or none at all), and with or without the convexity (C) condition, or the tests using external test statistics (chi-square, mean value of $P(\cdot;\boldsymbol{\theta})$ or the Fisher p-value). It returns an array containing the ordering of the tables, an array with the p-values, and a list of arrays containing the values of $\sum_{i=1}^{k} P(\mathbf{x}_{(k)};\boldsymbol{\theta})$ at each $\boldsymbol{\theta}$-value on the grid, at each iteration k. Therefore the function does in fact not take as input an observed table, but just the table dimensions and group sizes that uniquely define the outcome space Ω. Based on that it can then return an ordering of tables. By specifying to run `supremum_ordering()` until a single table of interest has been reached, we can turn this function into a test. Some of the helper functions that make up `supremum_ordering()` are

- `gen_tables()`, to list all the possible tables in the desired outcome space,
- `group_reduce()`, to reduce the outcome space into all its equivalence classes, according to the desired symmetry condition,
- `find_extreme()`, to find the symmetry classes that contain the most extreme tables, as we described in the last paragraph of 4.2.2,
- `next_indices()`, to find which tables to consider as candidates for the next step in the ordering (for the CSM methods), given the tables which were candidates in the previous step, and the tables which have already been ordered,
- `make_grid_qmc()`, to construct the Quasi-Monte Carlo grid used as a discretisation of Θ_0.

There are many small optimisations that happen "under the hood" which we shall not discuss at length here. For example, instead of always carrying around the actual tables to go computations with, we can represent each table by a unique integer for the bulk of computations (amongst others, to keep track of which tables we still have to order for example), and only use the table representation when needed. Another example would be that whenever we are constructing the symmetry classes, we do not keep track of a list of varying length containing the tables we still need to sort, but instead update a fixed-length vector of "status" values, indicating whether or not the corresponding table has already been sorted.

In the `functions.R` file, one can also find the implementations for all of the LP tests. Given table dimensions, group sizes, and a significance level, the function `lp_K()` returns the critical region found by one of the linear programming formulation discussed in Section 4.3.2, for a specified symmetry condition. There is also a function `lp_test()`, which takes as input a single table and returns an approximate p-value for that table using the binary search described in Algorithm 1. `lp_test_explorer()` does the same but with the extended binary search Algorithm 2.

Apart from this main file, there are a couple of other `R` scripts in the repository. These were mostly to work out specific examples treated in the text or to generate certain figures or tables. We list them below in order of appearance.

- `kalbfleisch.R` served to reproduce the results from Kalbfleisch and Sprott [32] as discussed in Section 3.2.
- `S_equivalences.R` helped us establishing the relationships between the different symmetry conditions implied each other in Section 4.2.4.
- `stack_example.R` created Figure 4.2 to explain the LP formulations as a packing problem.
- `optimal_grid.R` was used to build up Table 5.1.
- `grid_size_effects.R` has been used to investigate the effect of the chosen grid size on the test size and power. In particular, to produce Figure 5.1, 5.2, 5.3, 5.4, and 5.15.
- `lp_comparisons.R` compared the power of the critical regions, size functions and power functions of the original four LP formulations we had, as discussed in Section 5.2.
- `speed_plots.R` determined how long the preliminary computations took, as described in Section 5.3.1, and created Figures 5.12 and 5.13.
- `grid_time.R` we used to look at the effect of the grid size on the computation time, which led to Figure 5.14.
- `time_comparison.R` looked at the runtimes of the different test relative to each other, and as a function of the table and group sizes and outputted the timings in Excel spreadsheets (stored in the folder `time_comparison_files`). These were used by `time_boxplots.R` to generate Figures 5.16, 5.17, and 5.18 (raw images can be found in the folder `time_plots`).

- `gen_power_mat.R` generated, for the 2×2 case, a matrix of values of the power function on a grid of $\boldsymbol{\theta}$-values on Θ, for the desired tests, significance levels, and group sizes. This matrix was outputted as an Excel spreadsheet for later use (stored in the folder `power_mats`). For example, the script `plot_power_mat.R` used these spreadsheets to construct some of the plots of the size functions one can see in Appendix C (raw images in the folder `size_plots`). It is also able to plot the power function in the 2×2 case, a feature we did not end up using in the thesis. However, a number of power functions have been plotted and can be seen in the folder `power_plots`. `gen_power_mat.R` could also output an Excel spreadsheet with only the p-values (stored in the folder `p_arrs`). This was useful when we were not interested in a power matrix for a whole grid of $\boldsymbol{\theta}$-values, or when we moved on to larger tables for which a simple matrix would not suffice to represent the power function. Instead of `plot_power_mat.R`, we could then use `plot_size.R`, which would use the p-value spreadsheet to only compute the power for a grid on Θ_0 and output a plot. These p-value spreadsheets were also helpful to compute the power on coarser grids than the one for which we executed `plot_power_mat.R`. This was helpful when constructing the power comparison tables in Appendix C, which was done by the `gen_power_comp.R` script (which outputted files to the folder `power_comps`).

- `gen_long_term_power.R` outputted Excel spreadsheets (stored in the folder `long_term_powers`) of the long-term power, i.e., the proportion of tables in the critical region (up to a factor).

- `ntnu_example.R` and `discussion_example.R` were used to come up with the p-values given in Section 5.6 and generate Figure 6.1, respectively.

APPENDIX

C

LARGE FIGURES AND TABLES

The size and power study in Chapter 5 came with a number of very large figures and tables. In order to keep the main body of this thesis as organised as possible, we have kept the bulk of these plots and tables for this Appendix.

C.1 Size functions on 2×2 tables

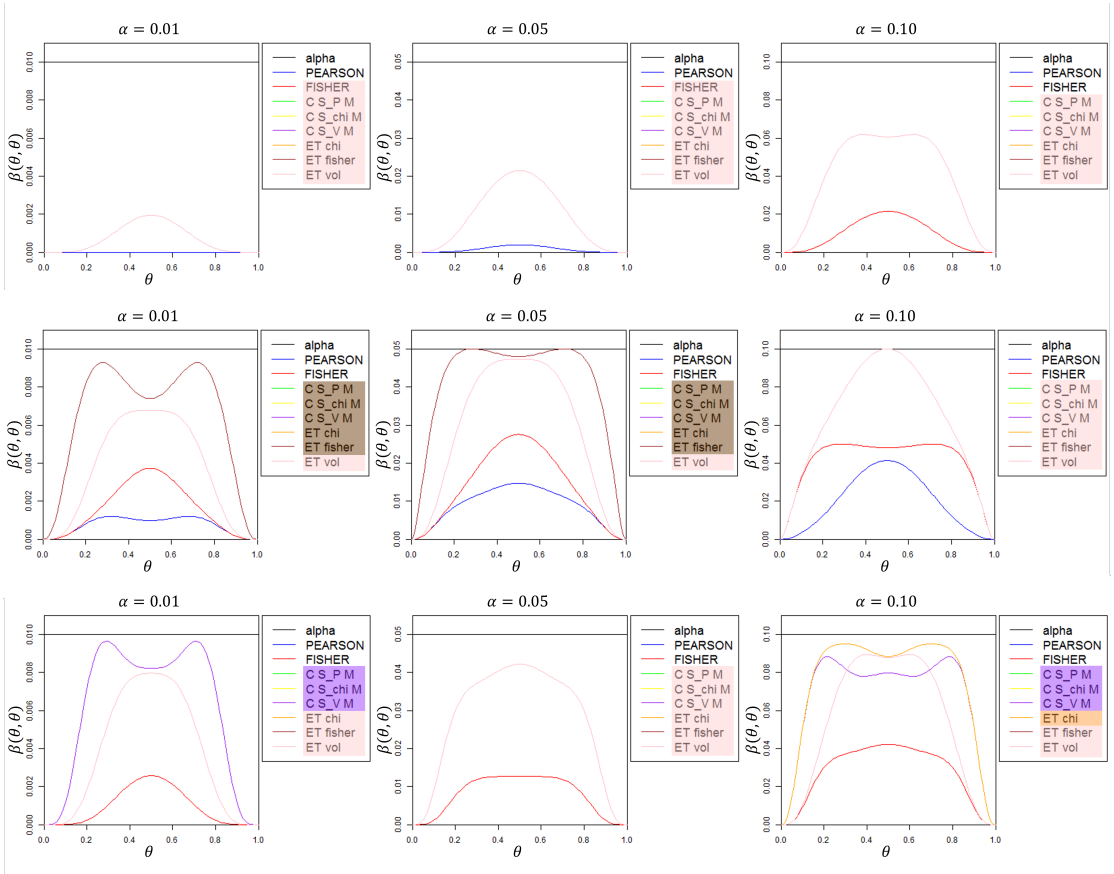


Figure C.1: $\beta(\theta, \theta)$ for indicated tests and α-values. Group sizes (from top to bottom) are $(5, 5)$, $(10, 5)$, and $(10, 10)$. Ambiguous overlaps are indicated in the legend.

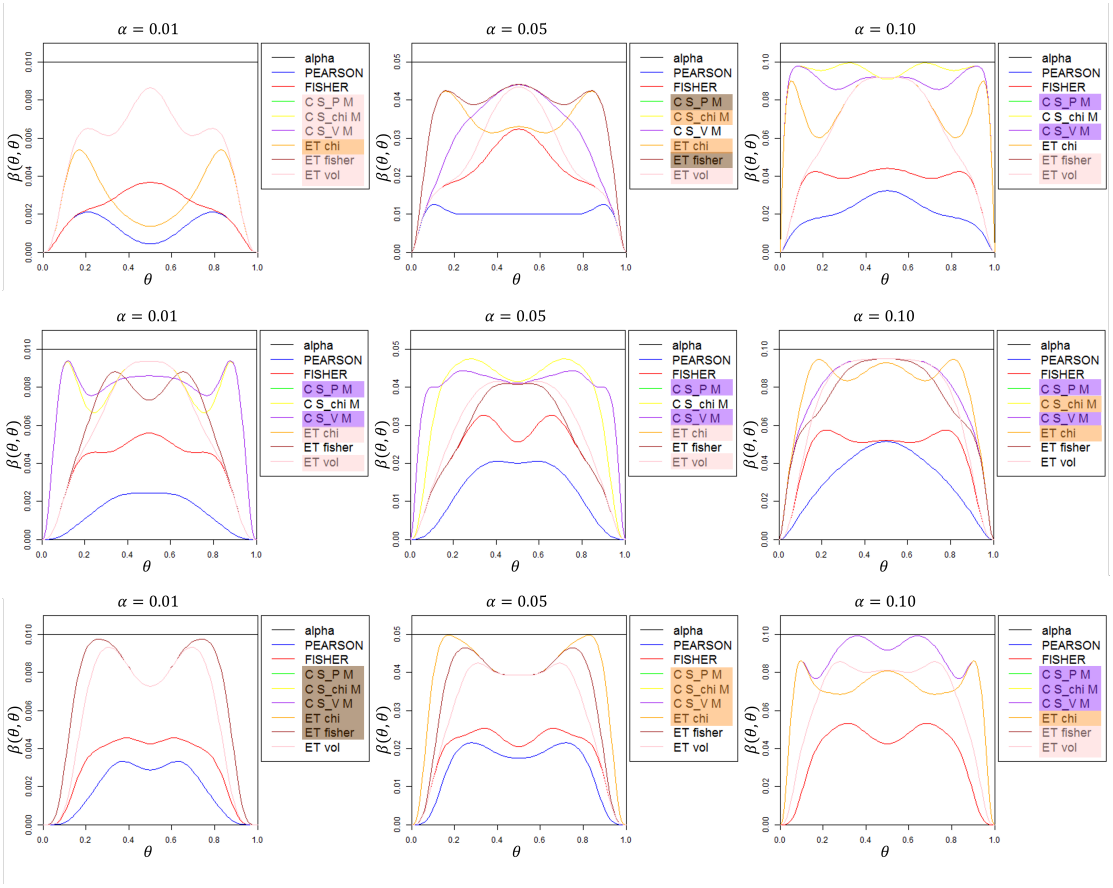


Figure C.2: $\beta(\theta, \theta)$ for indicated tests and α-values. Group sizes (from top to bottom) are $(20, 5)$, $(20, 10)$, and $(20, 20)$. Ambiguous overlaps are indicated in the legend.

C.2 Size functions on 3×2 tables

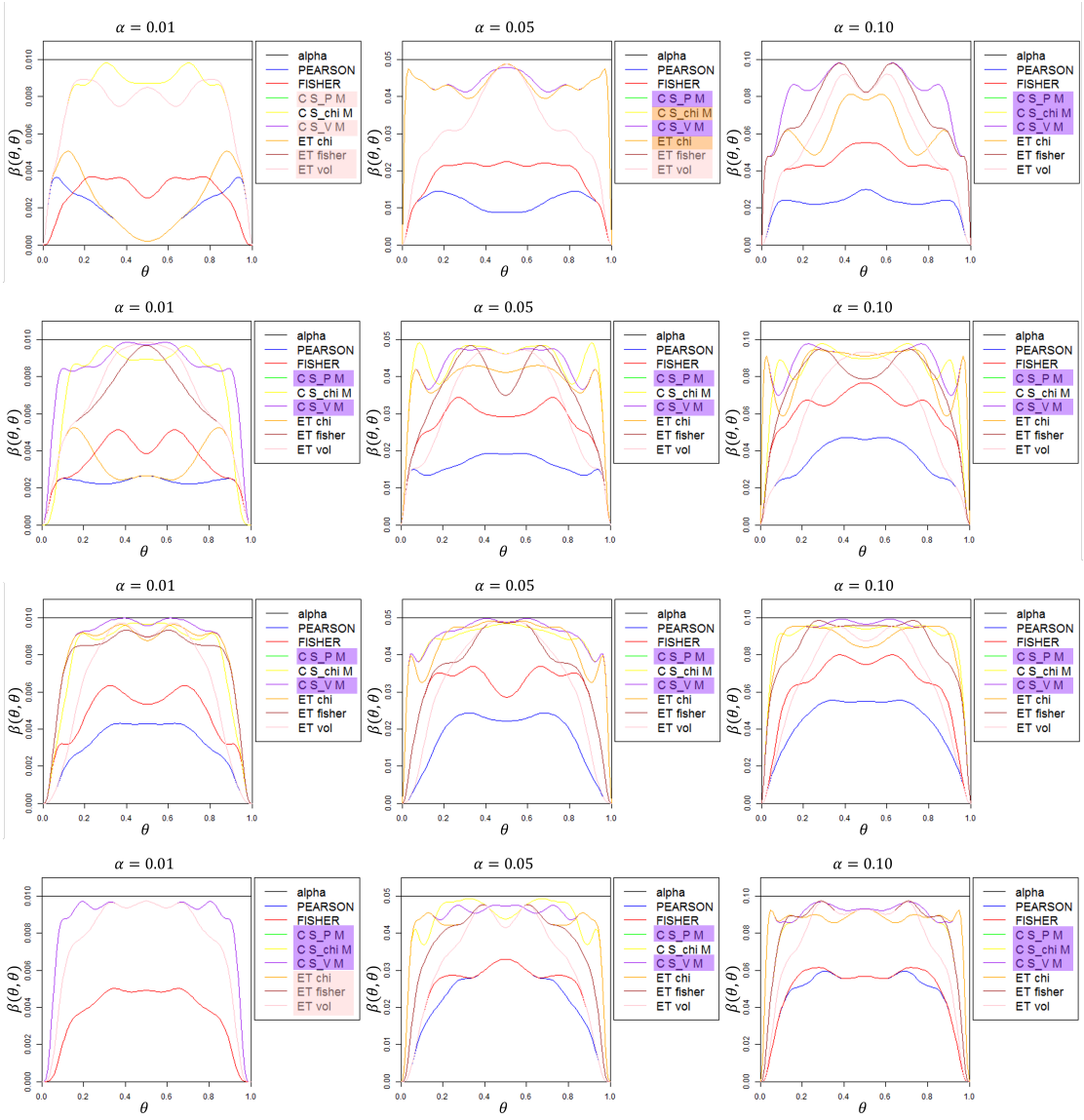


Figure C.3: $\beta(\theta, \theta)$ for indicated tests and α-values. Group sizes (from top to bottom) are $(40, 5)$, $(40, 10)$, $(40, 20)$, and $(40, 40)$. Ambiguous overlaps are indicated in the legend.

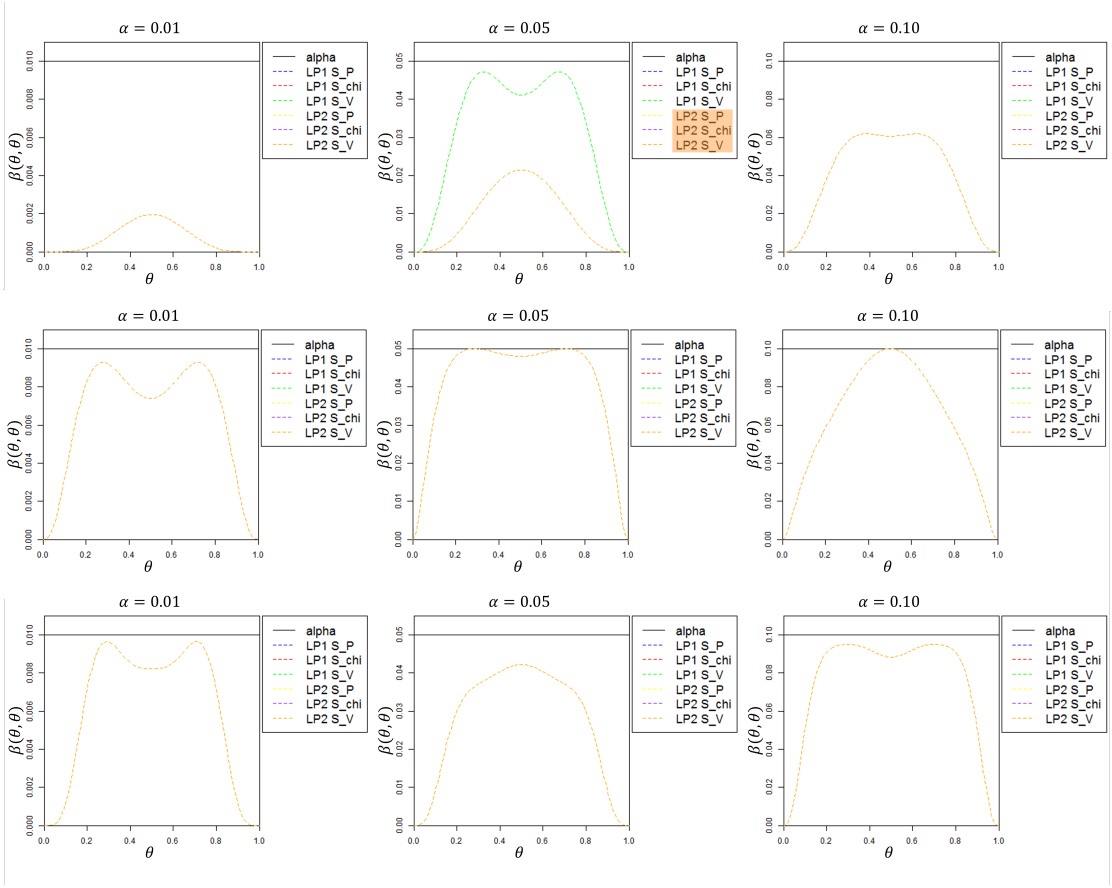


Figure C.4: $\beta(\theta, \theta)$ for indicated LP tests and α-values. Group sizes (from top to bottom) are $(5, 5)$, $(10, 5)$, and $(10, 10)$. Ambiguous overlaps are indicated in the legend.

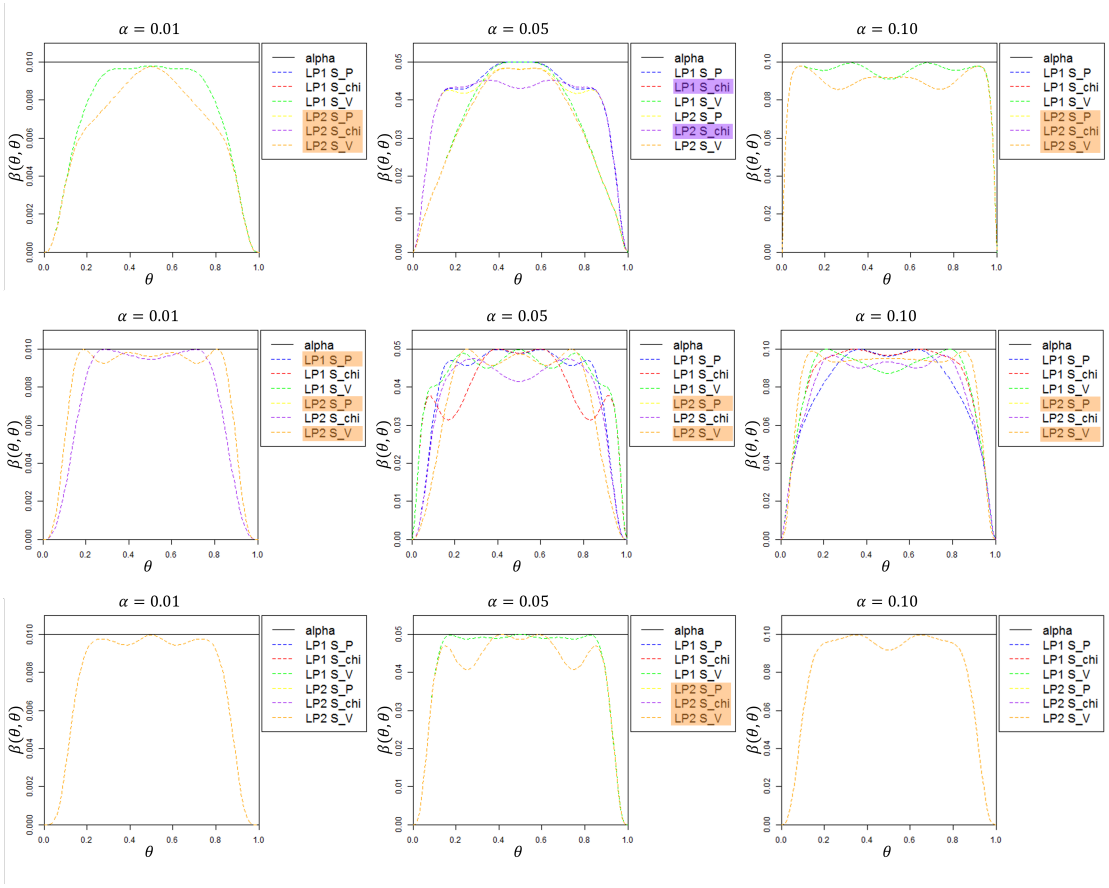


Figure C.5: $\beta(\theta, \theta)$ for indicated LP tests and α-values. Group sizes (from top to bottom) are $(20, 5)$, $(20, 10)$, and $(20, 20)$. Ambiguous overlaps are indicated in the legend.

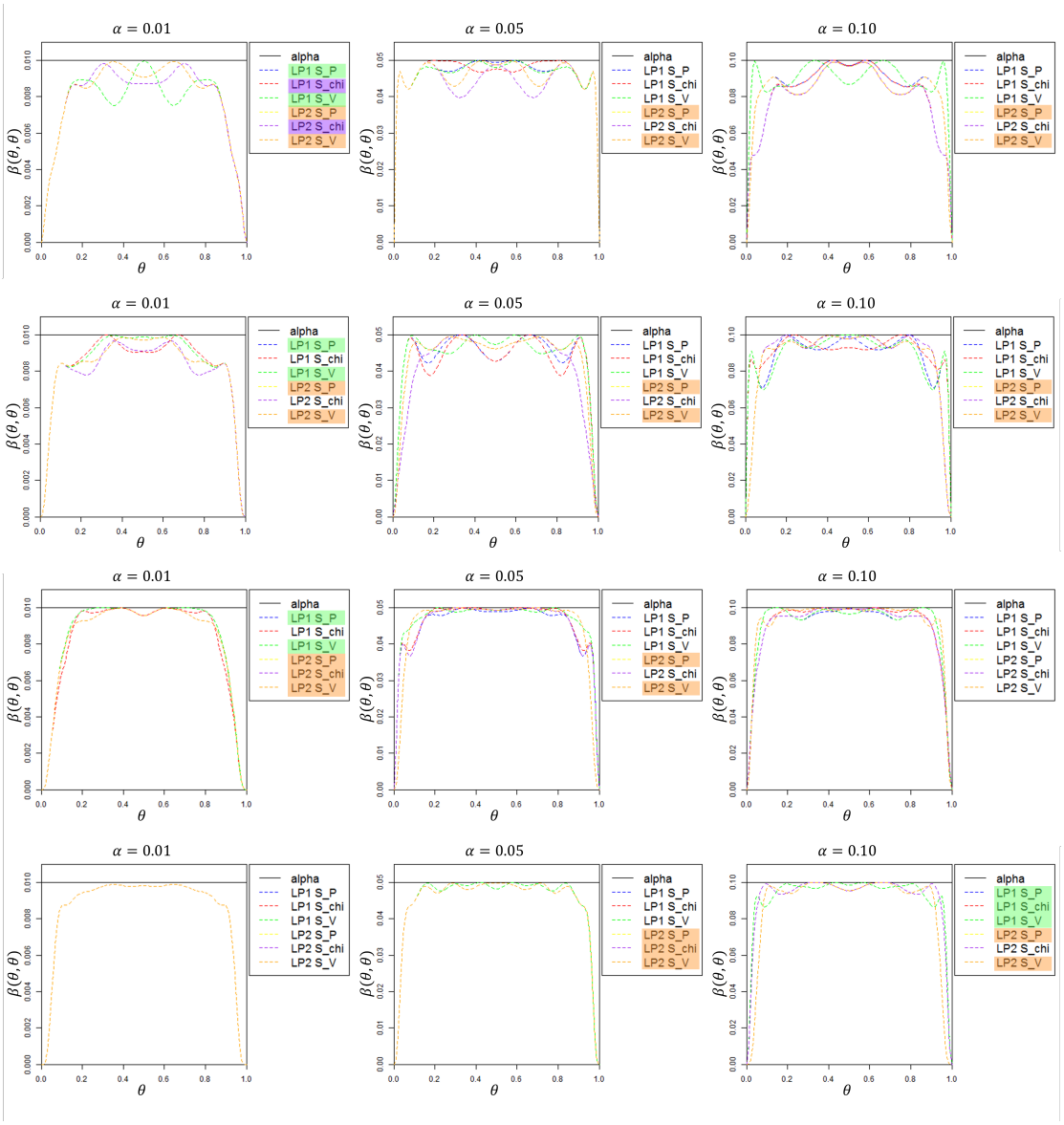


Figure C.6: $\beta(\theta, \theta)$ for indicated LP tests and α-values. Group sizes (from top to bottom) are $(40, 5)$, $(40, 10)$, $(40, 20)$, and $(40, 40)$. Ambiguous overlaps are indicated in the legend.

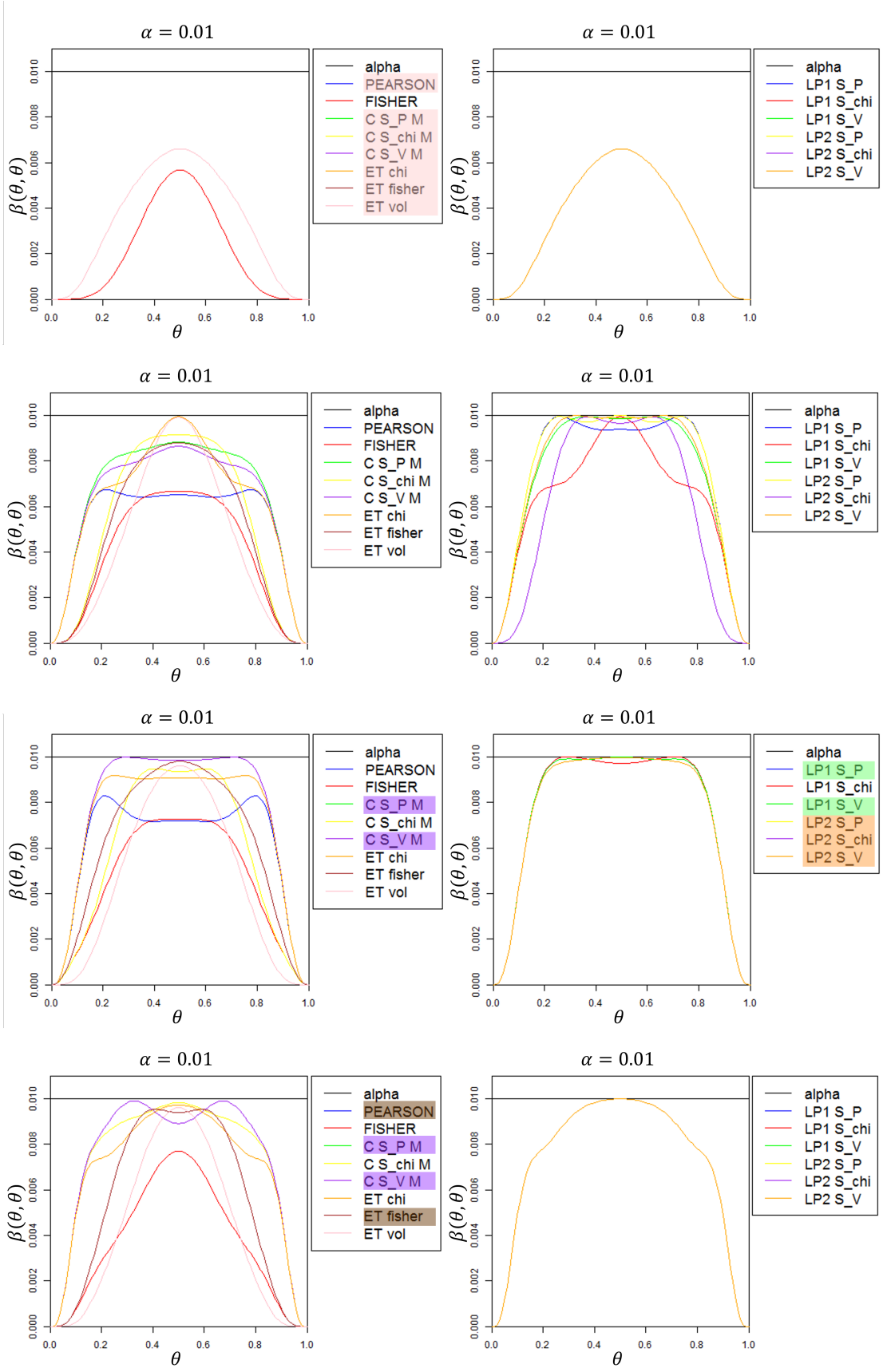


Figure C.7: $\beta(\theta, \theta, \theta)$ for indicated tests and $\alpha = 0.01$. Group sizes (from top to bottom) are $(5, 5, 5)$, $(10, 5, 5)$, $(10, 10, 5)$,and $(10, 10, 10)$. Ambiguous overlaps are indicated in the legend.

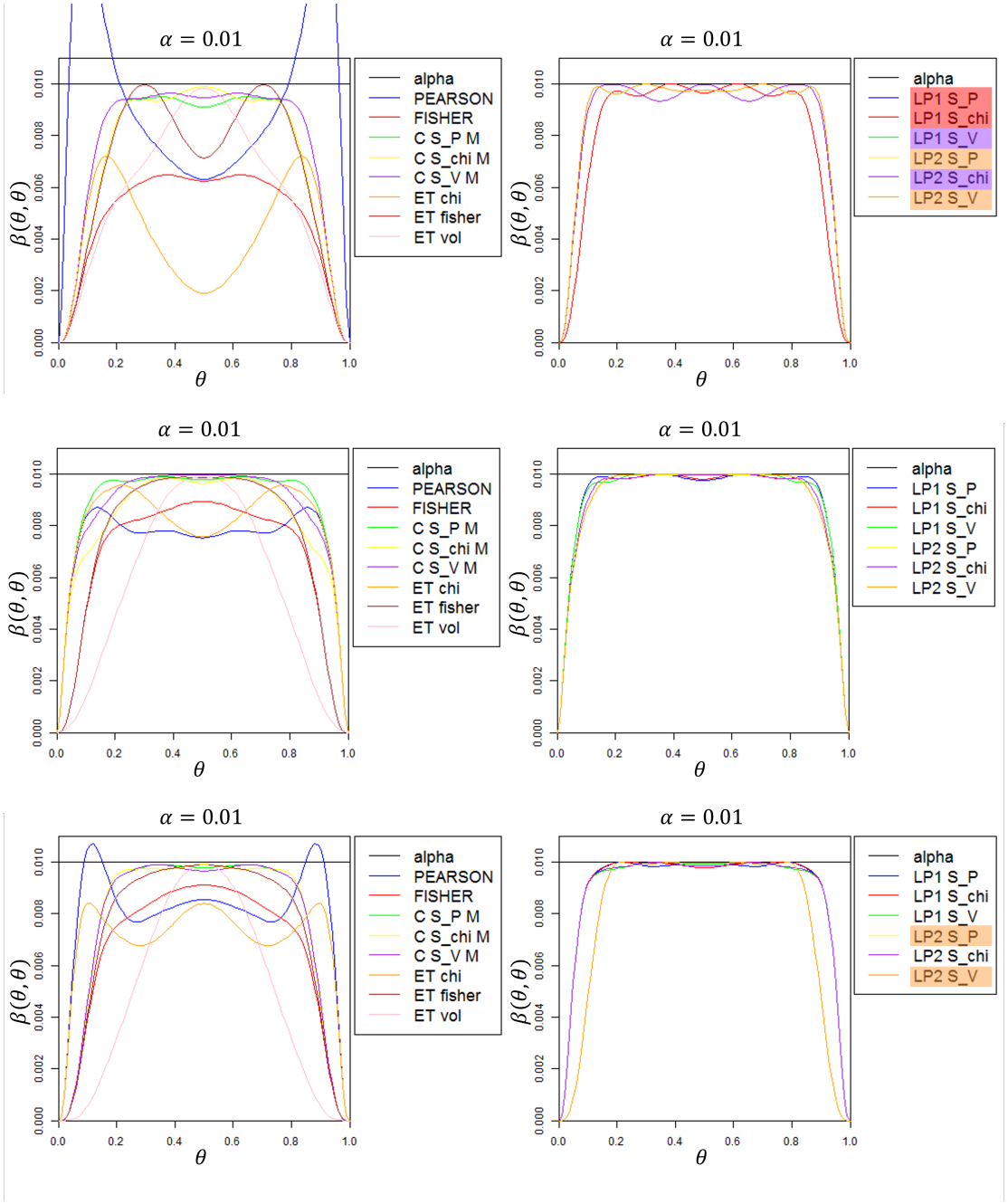


Figure C.8: $\beta(\theta, \theta, \theta)$ for indicated tests and $\alpha = 0.01$. Group sizes (from top to bottom) are $(20, 5, 5)$, $(20, 10, 5)$, and $(20, 20, 5)$. Ambiguous overlaps are indicated in the legend.

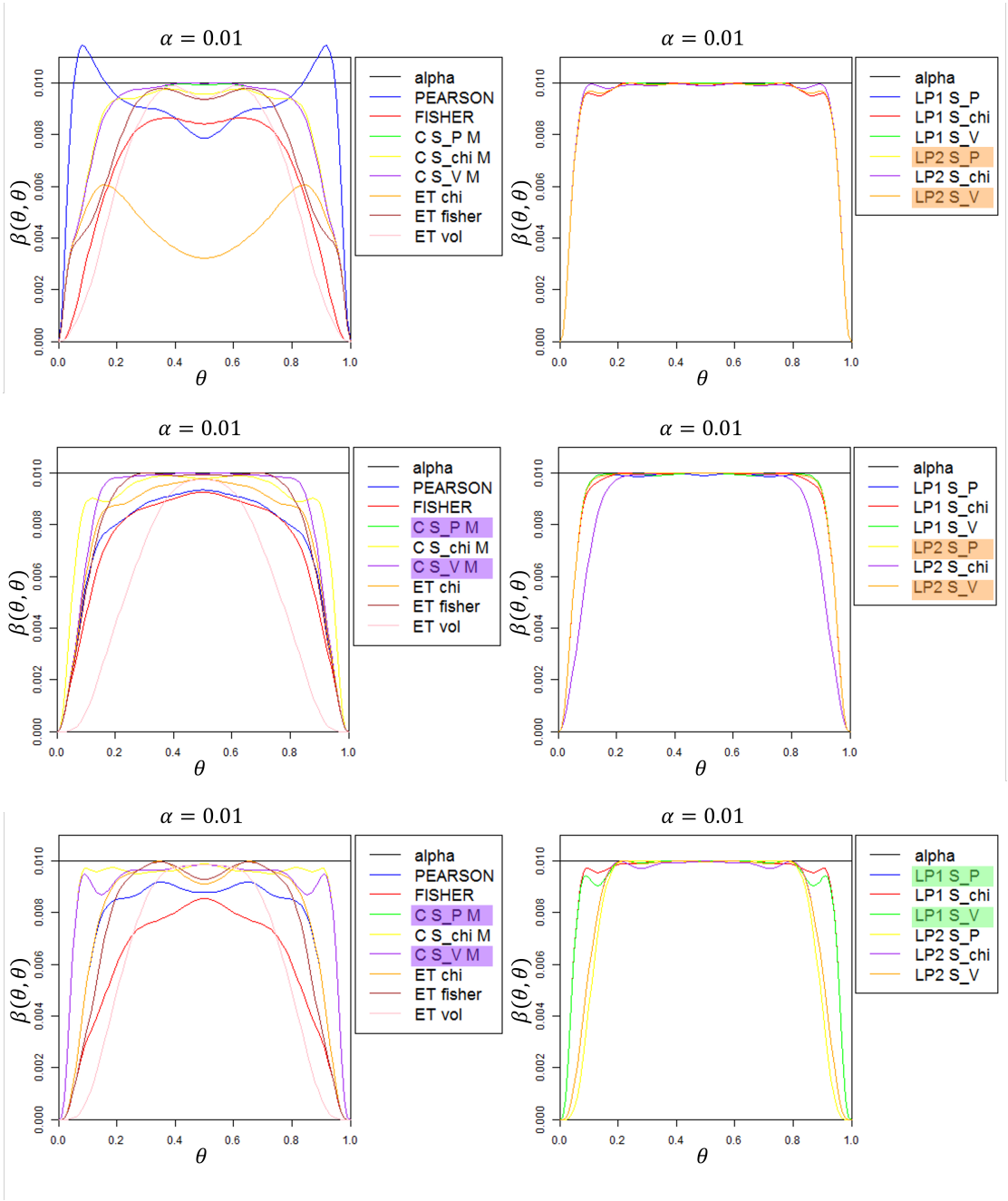


Figure C.9: $\beta(\theta, \theta, \theta)$ for indicated tests and $\alpha = 0.01$. Group sizes (from top to bottom) are $(20, 10, 10)$, $(20, 20, 10)$, and $(20, 20, 20)$. Ambiguous overlaps are indicated in the legend.

C.3 Tables for power comparison on 2×2 tables

Table C.1: Power comparison of tests on 2×2 tables with sample sizes (5.6) (`FISHER`, `C S_P M`, `C S_chi M`, and `C S_V M`).

	FISHER			C S_P M			C S_chi M			C S_V M		
PEARSON	< (-0.0590)	< (-0.0312)	=	< (-0.0590)	< (-0.0935)	< (-0.0863)	< (-0.0590)	< (-0.0935)	< (-0.0863)	< (-0.0590)	< (-0.0935)	< (-0.0863)
	< (-0.0327)	< (-0.0533)	< (-0.0282)	< (-0.0807)	< (-0.0899)	< (-0.0745)	< (-0.0807)	< (-0.0894)	< (-0.0745)	< (-0.0807)	< (-0.0899)	< (-0.0745)
	0.2023	< (-0.0274)	< (-0.0219)	=	< (-0.0763)	< (-0.0546)	< (-0.0403)	0.0855	< (-0.0468)	< (-0.0426)	< (-0.0403)	< (-0.0825)	< (-0.0763)	< (-0.0546)	< (-0.0403)
	(-0.0233)			(-0.0825)	(-0.0763)	(-0.0546)		(-0.0825)	(-0.0468)	(-0.0426)		=	=	=	
FISHER				=	< (-0.0623)	< (-0.0863)		=	< (-0.0623)	< (-0.0863)			< (-0.0623)	< (-0.0863)	
				< (-0.0480)	< (-0.0365)	< (-0.0463)		< (-0.0480)	< (-0.0360)	< (-0.0463)		< (-0.0480)	< (-0.0365)	< (-0.0463)	
				< (-0.0592)	< (-0.0490)	< (-0.0327)	< (-0.0403)	< (-0.0592)	< (-0.0195)	0.0857 (-0.0206)	< (-0.0403)	< (-0.0592)	< (-0.0490)	< (-0.0327)	< (-0.0403)
								=	=	=		=	=	=	
C S_P M								=	0.3643 (0.0005)	=		=	=	=	
								0.4464 (0.0000)	0.7024 (0.0295)	0.7375 (0.0121)	=	=	=	=	=
C S_chi M												=	=	=	
												=	0.6157 (-0.0005)	=	
												0.5144 (0.0000)	0.2962 (-0.0295)	0.2387 (-0.0121)	=
C S_V M															
ET chi															
ET fisher															
ET vol															
LP1 S_P															
LP1 S_chi															
LP1 S_V															
LP2 S_P															
LP2 S_chi															

Table C.2: Power comparison of tests on 2×2 tables with sample sizes (5.6) (`ET chi`, `ET fisher`, `ET vol`, and `LP1 S_P`)

	ET chi				ET fisher				ET vol				LP1 S_P			
PEARSON	< (-0.0590)	< (-0.0935)	< (-0.0679)		< (-0.0590)	< (-0.0935)	< (-0.0679)		< (-0.0590)	< (-0.0615)	< (-0.0679)		< (-0.0590)	< (-0.0935)	< (-0.0863)	
	< (-0.0322)	< (-0.0798)	< (-0.0745)		< (-0.0807)	< (-0.0803)	< (-0.0745)		< (-0.0807)	< (-0.0798)	< (-0.0644)		0.0076 (-0.0803)	< (-0.0889)	0.0763 (-0.0745)	
	< (-0.0081)	< (-0.0133)	< (-0.0523)	< (-0.0350)	< (-0.0825)	< (-0.0631)	< (-0.0500)	< (-0.0350)	< (-0.0825)	< (-0.0675)	< (-0.0426)	< (-0.0350)	< (-0.0825)	< (-0.0763)	< (-0.0549)	< (-0.0403)
	=	< (-0.0623)	< (-0.0679)		=	< (-0.0623)	< (-0.0679)		=	< (-0.0303)	< (-0.0679)		=	< (-0.0623)	< (-0.0863)	
FISHER	0.5334 (0.0005)	< (-0.0265)	< (-0.0463)		< (-0.0480)	< (-0.0269)	< (-0.0463)		< (-0.0480)	< (-0.0265)	< (-0.0362)		0.0288 (-0.0476)	0.0871 (-0.0356)	0.0931 (-0.0463)	
	0.6828 (0.0152)	0.6549 (0.0140)	< (-0.0304)	< (-0.0350)	< (-0.0592)	< (-0.0357)	< (-0.0281)	< (-0.0350)	< (-0.0592)	< (-0.0401)	0.0872 (-0.0206)	< (-0.0350)	< (-0.0592)	< (-0.0490)	0.0336 (-0.0330)	< (-0.0403)
	=	=	0.6632 (0.0184)		=	=	0.6632 (0.0184)		=	0.9802 (0.0320)	0.6632 (0.0184)		=	=	=	
C S_P M	0.9802 (0.0486)	0.5458 (0.0101)	=		=	0.6267 (0.0096)	=		=	0.5458 (0.0101)	0.9957 (0.0101)		0.2237 (0.0005)	0.2557 (0.0010)	0.3641 (0.0000)	
	0.9802 (0.0744)	0.9792 (0.0630)	0.4579 (0.0023)	0.9527 (0.0053)	=	0.73 (0.0132)	0.8196 (0.0047)	0.9527 (0.0053)	=	0.7142 (0.0088)	0.9466 (0.0121)	0.9527 (0.0053)	0.6092 (0.0000)	0.5766 (0.0000)	0.5864 (-0.0002)	0.4113 (0.0000)
	=	=	0.6632 (0.0184)		=	=	0.6632 (0.0184)		=	0.9802 (0.0320)	0.6632 (0.0184)		=	=	=	
C S_chi M	0.9802 (0.0486)	0.9802 (0.0096)	=		=	0.7871 (0.0091)	=		=	0.9802 (0.0096)	0.9957 (0.0101)		0.2237 (0.0005)	0.4888 (0.0005)	0.3641 (0.0000)	
	0.9802 (0.0744)	0.8618 (0.0335)	0.1395 (-0.0098)	0.9527 (0.0053)	0.5144 (0.0000)	0.5356 (-0.0162)	0.417 (-0.0074)	0.9527 (0.0053)	0.5144 (0.0000)	0.5068 (-0.0207)	0.5191 (0.0000)	0.9527 (0.0053)	0.3211 (0.0000)	0.4086 (-0.0295)	0.3999 (-0.0123)	0.4113 (0.0000)
	=	=	0.6632 (0.0184)		=	=	0.6632 (0.0184)		=	0.9802 (0.0320)	0.6632 (0.0184)		=	=	=	
C S_V M	0.9802 (0.0486)	0.5458 (0.0101)	=		=	0.6267 (0.0096)	=		=	0.5458 (0.0101)	0.9957 (0.0101)		0.2237 (0.0005)	0.2557 (0.0010)	0.3641 (0.0000)	
	0.9802 (0.0744)	0.9792 (0.0630)	0.4579 (0.0023)	0.9527 (0.0053)	=	0.73 (0.0132)	0.8196 (0.0047)	0.9527 (0.0053)	=	0.7142 (0.0088)	0.9466 (0.0121)	0.9527 (0.0053)	0.6092 (0.0000)	0.5766 (0.0000)	0.5864 (-0.0002)	0.4113 (0.0000)
					=	=	=		=	0.9802 (0.0320)	=		=	=	0.3364 (-0.0184)	
ET chi					< (-0.0486)	0.6102 (-0.0005)	=		< (-0.0486)	0.9957 (0.0101)			0.0186 (-0.0481)	0.3509 (-0.0091)	0.3641 (0.0000)	
					< (-0.0744)	< (-0.0498)	0.5778 (0.0023)	=	< (-0.0744)	0.6201 (0.0098)	=		< (-0.0744)	< (-0.0630)	0.5188 (-0.0026)	0.1221 (-0.0053)
									=	0.9802 (0.0320)	=		=	=	0.3364 (-0.0184)	
ET fisher									=	0.3696 (0.0005)	0.9957 (0.0101)		0.2237 (0.0005)	0.2559 (-0.0086)	0.3641 (0.0000)	
									=	< (-0.0044)	0.3053 (0.0074)	=	0.6092 (0.0000)	0.0035 (-0.0132)	0.3107 (-0.0049)	0.1221 (-0.0053)
													=	< (-0.0320)	0.3364 (-0.0184)	
ET vol													0.2237 (0.0005)	0.3509 (-0.0091)	0.2259 (-0.0101)	
													0.6092 (0.0000)	0.0345 (-0.0088)	0.2038 (-0.0123)	0.1221 (-0.0053)
LP1 S_P																
LP1 S_chi																
LP1 S_V																
LP2 S_P																
LP2 S_chi																

Table C.3: Power comparison of tests on 2×2 tables with sample sizes (5.6) (`LP1 S_chi`, `LP1 S_V`, `LP2 S_P`, and `LP2 S_chi`)

	LP1 S_chi				LP1 S_V				LP2 S_P				LP2 S_chi			
PEARSON	< (-0.0590)	(-0.0935)	< (-0.0863)		< (-0.0590)	(-0.0935)	< (-0.0863)		< (-0.0590)	(-0.0935)	< (-0.0863)		< (-0.0590)	(-0.0935)	< (-0.0863)	
	0.0076		0.0763		0.0076		0.0763		< (-0.0807)	< (-0.0889)	0.0763		< (-0.0807)	< (-0.0884)	0.0763	
	(-0.0803)	(-0.0884)	(-0.0745)		(-0.0803)	(-0.0884)	(-0.0745)				(-0.0745)				(-0.0745)	
	< (-0.0825)	(-0.0674)	< (-0.0549)	< (-0.0403)	< (-0.0825)	< (-0.0763)	0.0018 (-0.0549)	< (-0.0403)	(-0.0825)	(-0.0763)	< (-0.0546)	< (-0.0403)	< (-0.0825)	< (-0.0674)	< (-0.0546)	< (-0.0403)
FISHER	=	(-0.0623)	< (-0.0863)		=	(-0.0623)	< (-0.0863)		=	(-0.0623)	< (-0.0863)		=	(-0.0623)	< (-0.0863)	
	0.0288	< (-0.0351)	0.0931 (-0.0463)		0.0288	< (-0.0476)	0.0931 (-0.0463)		< (-0.0480)	0.0871 (-0.0356)	0.0931 (-0.0463)		< (-0.0480)	< (-0.0351)	0.0931 (-0.0463)	
	< (-0.0476)	0.0263	< (-0.0330)	< (-0.0403)	< (-0.0592)	< (-0.0490)	0.0336 (-0.0330)	< (-0.0403)	(-0.0592)	< (-0.0490)	< (-0.0327)	< (-0.0403)	< (-0.0592)	0.0137 (-0.0401)	< (-0.0327)	< (-0.0403)
C S_P M	=	=	=		=	=	=		=	=	=		=	=	=	
	0.2237	0.4517	0.3641		0.2237	0.4517	0.3641		0.3686	0.2557	0.3641		0.3686	0.4517	0.3641	
	(0.0005)	(0.0014)	(0.0000)		(0.0005)	(0.0014)	(0.0000)		(0.0000)	(0.0010)	(0.0000)		(0.0000)	(0.0014)	(0.0000)	
	0.4464	0.6805	0.5054	0.4113	0.6092	0.5766	0.5864	0.4113	0.4852	=	=	0.4113	0.4464	0.7988	=	0.4113
	(0.0000)	(0.0089)	(0.0003)	(0.0000)	(0.0000)	(0.0000)	(-0.0002)	(0.0000)	(0.0000)			(0.0000)	(0.0000)	(0.0089)		(0.0000)
C S_chi M	=	=	=		=	=	=		=	=	=		=	=	=	
	0.2237	0.2098	0.3641		0.2237	0.2098	0.3641		0.3686	0.4888	0.3641		0.3686	0.2098	0.3641	
	(0.0005)	(0.0009)	(0.0000)		(0.0005)	(0.0009)	(0.0000)		(0.0000)	(0.0005)	(0.0000)		(0.0000)	(0.0009)	(0.0000)	
	=	0.4402	0.304	0.4113	0.3211	0.4086	0.3999	0.4113	0.2517	0.2962	0.2387	0.4113	=	0.4923	0.2387	0.4113
		(-0.0206)	(-0.0123)	(0.0000)	(0.0000)	(-0.0295)	(-0.0123)	(0.0000)	(0.0000)	(-0.0295)	(-0.0121)	(0.0000)		(-0.0206)	(-0.0121)	(0.0000)
C S_V M	=	=	=		=	=	=		=	=	=		=	=	=	
	0.2237	0.4517	0.3641		0.2237	0.4517	0.3641		0.3686	0.2557	0.3641		0.3686	0.4517	0.3641	
	(0.0005)	(0.0014)	(0.0000)		(0.0005)	(0.0014)	(0.0000)		(0.0000)	(0.0010)	(0.0000)		(0.0000)	(0.0014)	(0.0000)	
	0.4464	0.6805	0.5054	0.4113	0.6092	0.5766	0.5864	0.4113	0.4852	=	=	0.4113	0.4464	0.7988	=	0.4113
	(0.0000)	(0.0089)	(-0.0003)	(0.0000)	(0.0000)	(0.0000)	(-0.0002)	(0.0000)	(0.0000)			(0.0000)	(0.0000)	(0.0089)		(0.0000)
ET chi	=	=	0.3364 (-0.0184)		=	=	0.3364 (-0.0184)		=	=	0.3364 (-0.0184)		=	=	0.3364 (-0.0184)	
	0.0186	< (-0.0087)	0.3641 (0.0000)		0.0186	< (-0.0087)	0.3641 (0.0000)		< (-0.0486)	0.3509 (-0.0091)	0.3641 (0.0000)		< (-0.0486)	< (-0.0087)	0.3641 (0.0000)	
	(-0.0481)		0.5197	0.1221	(-0.0481)		0.5188	0.1221	<	<	0.4946	0.1221	<	<	0.4946	0.1221
	< (-0.0744)	< (-0.0541)	(-0.0026)	(-0.0053)	< (-0.0744)	< (-0.0630)	(-0.0026)	(-0.0053)	(-0.0744)	(-0.0630)	(-0.0023)	(-0.0053)	(-0.0744)	< (-0.0541)	(-0.0023)	(-0.0053)
ET fisher	=	=	0.3364 (-0.0184)		=	=	0.3364 (-0.0184)		=	=	0.3364 (-0.0184)		=	=	0.3364 (-0.0184)	
	0.2237	0.1965	0.3641		0.2237	0.1965	0.3641		0.3686	0.2559	0.3641		0.3686	0.1965	0.3641	
	(0.0005)	(-0.0082)	(0.0000)		(0.0005)	(-0.0082)	(0.0000)		(0.0000)	(-0.0086)	(0.0000)		(0.0000)	(-0.0082)	(0.0000)	
	0.4464	0.3307	0.1719	0.1221	0.6092	0.0035	0.3107	0.1221	0.4852	0.2308	0.1394	0.1221	0.4464	0.3183	0.1394	0.1221
	(0.0000)	(-0.0044)	(-0.0049)	(-0.0053)	(0.0000)	(-0.0132)	(-0.0049)	(-0.0053)	(0.0000)	(-0.0132)	(-0.0047)	(-0.0053)	(0.0000)	(-0.0044)	(-0.0047)	(-0.0053)
ET vol	=	< (-0.0320)	0.3364 (-0.0184)		=	< (-0.0320)	0.3364 (-0.0184)		=	< (-0.0320)	0.3364 (-0.0184)		=	< (-0.0320)	0.3364 (-0.0184)	
	0.2237	<	0.2259		0.2237	<	0.2259		0.3686	0.3509	0.2259		0.3686	<	0.2259	
	(0.0005)	(-0.0087)	(-0.0101)		(0.0005)	(-0.0087)	(-0.0101)		(0.0000)	(-0.0091)	(-0.0101)		(0.0000)	(-0.0087)	(-0.0101)	
	0.4464	0.4316	0.0304	0.1221	0.6092	0.0345	0.2038	0.1221	0.4852	0.2465	0.0304	0.1221	0.4464	0.511	0.0304	0.1221
	(0.0000)	(0.0001)	(-0.0123)	(-0.0053)	(0.0000)	(-0.0088)	(-0.0123)	(-0.0053)	(0.0000)	(-0.0088)	(-0.0121)	(-0.0053)	(0.0000)	(0.0001)	(-0.0121)	(-0.0053)
LP1 S_P	=	=	=		=	=	=		=	=	=		=	=	=	
	=	0.5302	=		=	0.5302	=		0.6583	=	=		0.6583	0.5302	=	
		(0.0005)				(0.0005)			(-0.0005)				(-0.0005)	(0.0005)		
	0.6587	0.5006	0.4521	=	=	=	=	=	0.6802	0.3842	0.3884	=	0.6587	0.8141	0.3884	=
	(0.0000)	(0.0089)	(0.0000)						(0.0000)	(0.0000)	(0.0002)		(0.0000)	(0.0089)	(0.0002)	
LP1 S_chi					=	=	=		=	=	=		=	=	=	
					=	=	=		0.6583	0.4498	=		0.6583	=	=	
									(-0.0005)	(-0.0005)			(-0.0005)			
					0.3211	0.4786	0.5064	=	0.2517	0.2837	0.4593	=	=	0.6931	0.4593	=
					(0.0000)	(-0.0089)	(0.0000)		(0.0000)	(-0.0089)	(0.0003)			(0.0000)	(0.0003)	
LP1 S_V									=	=	=		=	=	=	
									0.6583	0.4498	=		0.6583	=	=	
									(-0.0005)	(-0.0005)			(-0.0005)			
									0.6802	0.3842	0.3884	=	0.6587	0.8141	0.3884	=
									(0.0000)	(0.0000)	(0.0002)		(0.0000)	(0.0089)	(0.0002)	
LP2 S_P													=	=	=	
													=	0.5302	=	
														(0.0005)		
													0.709	0.7988	=	=
													(0.0000)	(0.0089)		
LP2 S_chi																

Table C.4: Power comparison of tests on 2×2 tables with sample sizes (5.6) (`LP2 S_V`)

	LP2 S_V			
PEARSON	< (-0.0590)	< (-0.0935)	< (-0.0863)	
	< (-0.0807)	< (-0.0889)	0.0763 (-0.0745)	< (-0.0403)
	= (-0.0825)	< (-0.0763)	< (-0.0546)	
FISHER		< (-0.0623)	< (-0.0863)	
	< (-0.0480)	0.0871 (-0.0356)	0.0931 (-0.0463)	< (-0.0403)
	< (-0.0592)	< (-0.0490)	< (-0.0327)	
C S_P M	= 0.3686 (0.0000)	= 0.2557 (0.0010)	= 0.3641 (0.0000)	0.4113 (0.0000)
	0.4852 (0.0000)	=	=	
C S_chi M	= 0.3686 (0.0000)	= 0.4888 (0.0005)	= 0.3641 (0.0000)	
	0.2517 (0.0000)	0.2962 (-0.0295)	0.2387 (-0.0121)	0.4113 (0.0000)
C S_V M	= 0.3686 (0.0000)	= 0.2557 (0.0010)	= 0.3641 (0.0000)	
	0.4852 (0.0000)	=	=	0.4113 (0.0000)
ET chi	=	=	0.3364 (-0.0184)	
	< (-0.0486)	0.3509 (-0.0091)	0.3641 (0.0000)	
	< (-0.0744)	< (-0.0630)	0.4946 (-0.0023)	0.1221 (-0.0053)
ET fisher	= 0.3686 (0.0000)	= 0.2559 (-0.0086)	0.3364 (-0.0184) 0.3641 (0.0000)	
	0.4852 (0.0000)	0.2308 (-0.0132)	0.1394 (-0.0047)	0.1221 (-0.0053)
ET vol	= 0.3686 (0.0000)	< (-0.0320) 0.3509 (-0.0091)	0.3364 (-0.0184) 0.2259 (-0.0101)	
	0.4852 (0.0000)	0.2465 (-0.0088)	0.0304 (-0.0121)	0.1221 (-0.0053)
LP1 S_P	= 0.6583 (-0.0005)	=	=	
	0.6802 (0.0000)	0.3842 (0.0000)	0.3884 (0.0002)	=
LP1 S_chi	= 0.6583 (-0.0005)	= 0.4498 (-0.0005)	=	
	0.2517 (0.0000)	0.2837 (-0.0089)	0.4593 (0.0003)	=
LP1 S_V	= 0.6583 (-0.0005)	= 0.4498 (-0.0005)	=	
	0.6802 (0.0000)	0.3842 (0.0000)	0.3884 (0.0002)	=
LP2 S_P	=	=	=	
	=	=	=	
	=	=	=	=
LP2 S_chi	=	= 0.4498 (-0.0005)	=	
	0.2517 (0.0000)	0.1804 (-0.0089)	=	=

C.4 Tables for power comparison on 3×2 tables

Table C.5: Power comparison of tests on 3×2 tables with sample sizes (5.8) (`FISHER`, `C S_P M`, `C S_chi M`, and `C S_V M`).

FISHER			C S_P M			C S_chi M			C S_V M		
0.9455	0.6444	0.7960	0.3232	0.0889		0.2909	0.4465	0.2970		0.3273	0.0889
(0.3246)	(0.0620)	(0.2148)	(0.00077)		=	(-0.00517)	(-0.00018)	(-0.00635)	=	(0.00113)	(-0.01219)
	0.5091	0.1899	0.4566		<	0.2343	0.4222	0.7717	<	0.1879	0.2101
	(0.00633)	(-0.01000)	(-0.00523)		(-0.02460)	(-0.01410)	(0.00060)	(0.01219)	(-0.01770)	(-0.01963)	(-0.01382)
	0.9455	0.2707	0.5818		<	0.2303	0.3838	0.2485	<	0.1576	0.1758
	(0.03099)	(-0.01316)	(0.00344)		(-0.03210)	(-0.00839)	(-0.00356)	(-0.00785)	(-0.03210)	(-0.01871)	(-0.00879)
			0.1455	0.3232	0.0889	0.1576	0.1434	0.3535	0.0889	0.1455	<
			(-0.02418)	(-0.00270)	(-0.01155)	(-0.01137)	(-0.02166)	(-0.00712)	(-0.02390)	(-0.02035)	(-0.01296)
			0.2606	0.1636	0.2101	0.0364	0.6061	0.7556	<	0.2202	0.1576
			(-0.01196)	(-0.02195)	(-0.01382)	(-0.02043)	(0.01059)	(0.01742)	(-0.03843)	(-0.00963)	(-0.00859)
			0.0990	0.1576	0.1758	<	0.4808	0.3212	0.0485	0.3131	<
			(-0.00759)	(-0.02075)	(-0.00879)	(-0.03938)	(0.00960)	(-0.01129)	(-0.03807)	(-0.00555)	(-0.01223)
						0.6384	0.4222	0.7030	0.6545	0.5111	0.2667
						(0.01943)	(0.00252)	(0.00520)	(0.00691)	(0.00383)	(-0.00064)
						0.7152	0.7232	0.8404	=	0.3859	=
						(0.01800)	(0.02255)	(0.02601)		(0.00232)	
						0.5818	0.7030	0.6424	=	0.6121	=
						(-0.00130)	(0.01719)	(0.00094)		(0.00205)	
									0.3677	0.5576	0.1818
									(-0.01253)	(0.00131)	(-0.00585)
									0.2303	0.2162	0.1051
									(-0.01800)	(-0.02023)	(-0.02601)
									0.3636	0.2545	0.3030
									(0.00130)	(-0.01514)	(-0.00094)

Table C.6: Power comparison of tests on 3×2 tables with sample sizes (5.8) (`ET chi`, `ET fisher`, `ET vol`, and `LP1 S_P`)

ET chi			ET fisher			ET vol			LP1 S_P		
<	>	0.9455	=	0.2586	0.4343	0.2141	=	0.3556	0.4485	0.2424	=	<	0.3273	0.0848
(-0.02299)	(0.05965)	(0.05339)		(-0.01096)	(0.00255)	(-0.00485)		(-0.00725)	(0.00610)	(0.00042)		(-0.03136)	(-0.00931)	(-0.01499)
<	<	>		0.1414	0.1515	0.0808		0.3111	0.3192	0.4687		0.0162	0.0990	0.1172
(-0.02034)	(-0.00522)	(-0.00350)		(-0.02110)	(-0.01981)	(-0.01506)		(-0.00146)	(0.00085)	(0.00648)		(-0.03268)	(-0.02539)	(-0.01717)
=	0.7677	<		0.4364	0.0768	0.1455		0.6303	0.5131	0.2788		0.2061	0.1414	0.1333
	(0.00414)	(-0.00418)		(-0.00978)	(-0.02006)	(-0.01108)		(0.01828)	(0.00668)	(0.00145)		(-0.01902)	(-0.02606)	(-0.01568)
<	0.8303	0.8889	<	<	<	<	<	0.2586	0.2465	0.3960	<	<	<	0.1293
(-0.03246)	(-0.02919)	(0.05262)	(-0.03246)	(-0.01716)	(-0.01893)	(-0.00562)	(-0.03246)	(-0.01345)	(-0.01538)	(-0.00036)	(-0.03246)	(-0.03756)	(-0.03078)	(-0.01576)
<	<	0.4566		<	<	>		0.2990	0.4929	0.6303		0.0081	0.0970	>
(-0.02666)	(0.00478)	(0.00173)		(-0.02742)	(-0.00981)	(-0.00983)		(-0.00778)	(0.01085)	(0.01171)		(-0.03901)	(-0.01539)	(-0.01194)
<	0.7758	0.3394		<	<	<		0.2061	0.4970	0.4606		<	0.0364	<
(-0.03099)	(0.01730)	(-0.00762)		(-0.04076)	(-0.00689)	(-0.01452)		(-0.01270)	(0.01985)	(-0.00199)		(-0.05001)	(-0.01290)	(-0.01912)
=	0.4485	0.9455	=	0.7030	0.6263	0.8000	=	0.5960	0.4444	0.8525	=	0.3596	0.2586	0.2970
	(0.00162)	(0.08248)	(0.06494)	(0.01364)	(0.00525)	(0.00670)		(0.01735)	(0.00879)	(0.01197)		(-0.00676)	(-0.00661)	(-0.00344)
	0.7030	0.7253	0.6424	0.6263	0.5434	0.4687		0.7636	0.7010	0.7879		0.1172	0.4162	0.1293
	(0.01176)	(0.01674)	(0.01033)	(0.01100)	(0.00215)	(-0.00124)		(0.03064)	(0.02280)	(0.02031)		(-0.00058)	(-0.00344)	(-0.00335)
	0.5697	0.8727	0.5818	0.2788	0.5778	0.2909		0.6788	0.6465	0.6061		0.1697	0.3798	0.0970
	(0.00709)	(0.02489)	(0.00461)	(-0.00269)	(0.00070)	(-0.00229)		(0.02537)	(0.02744)	(0.01024)		(-0.01193)	(-0.00531)	(-0.00689)
=	0.1899	0.9273	0.8949	0.3899	0.4646	0.4626	=	0.3616	0.6364	0.4424	=	0.3556	0.3051	0.2586
	(-0.01782)	(0.07847)	(0.05974)	(-0.00579)	(0.00273)	(0.00150)		(-0.00209)	(0.00627)	(0.00676)		(-0.02619)	(-0.00913)	(-0.00864)
	0.4162	0.4303	0.0970	0.4040	0.1798	0.0202		0.6889	0.4101	0.2788		0.2141	0.1838	0.0364
	(-0.00623)	(-0.00581)	(-0.01568)	(-0.00700)	(-0.02040)	(-0.02725)		(0.01264)	(0.00025)	(-0.00570)		(-0.01858)	(-0.02599)	(-0.02936)
	0.7152	0.6465	0.4242	0.4727	0.1778	0.1455		0.9455	0.4848	0.3515		0.0848	0.1253	0.0242
	(0.00839)	(0.00770)	(0.00367)	(-0.00139)	(-0.01649)	(-0.00323)		(0.02667)	(0.01025)	(0.00930)		(-0.01063)	(-0.02250)	(-0.00783)
=	0.2101	>	0.9455	0.5253	0.6384	0.7354	=	0.5071	0.4202	0.7717	=	0.3434	0.2061	0.3273
	(-0.00529)	(0.07561)	(0.06558)	(0.00674)	(0.00143)	(0.00734)		(0.01044)	(0.00497)	(0.01261)		(-0.01366)	(-0.01043)	(-0.00280)
	0.7030	0.6707	0.6424	0.6263	0.5010	0.4687		0.7636	0.7253	0.7879		0.1172	0.3394	0.1293
	(0.01176)	(0.01441)	(0.01033)	(0.01100)	(-0.00018)	(-0.00124)		(0.03064)	(0.02048)	(0.02031)		(-0.00058)	(-0.00576)	(-0.00335)
	0.5697	0.8727	0.5818	0.2788	0.3677	0.2909		0.6788	0.6182	0.6061		0.1697	0.3475	0.0970
	(0.00709)	(0.02284)	(0.00461)	(-0.00269)	(-0.00135)	(-0.00229)		(0.02537)	(0.02539)	(0.01024)		(-0.01193)	(-0.00735)	(-0.00689)
			=	0.7475	0.0687	<	=	0.7273	0.0889	0.0808	=	0.3717	<	<
				(0.01203)	(-0.07574)	(-0.05551)		(0.01573)	(-0.07220)	(-0.05297)		(-0.00837)	(-0.08760)	(-0.06838)
				0.3879	0.2424	0.1535		0.6162	0.4364	0.5818		0.2424	0.1616	0.2465
				(-0.00076)	(-0.01459)	(-0.01157)		(0.01888)	(0.00607)	(0.00998)		(-0.01234)	(-0.02018)	(-0.01367)
				0.4364	0.0323	0.2182		0.6303	0.4444	0.4242		0.2061	<	0.1697
				(-0.00978)	(-0.02419)	(-0.00690)		(0.01828)	(0.00255)	(0.00563)		(-0.01902)	(-0.03020)	(-0.01150)
							=	0.4444	0.4970	0.6263	=	0.3313	0.1657	0.2424
								(0.00370)	(0.00354)	(0.00527)		(-0.02040)	(-0.01186)	(-0.01014)
								0.9455	0.7677	0.9051		0.2424	0.2869	0.4121
								(0.01964)	(0.02066)	(0.02155)		(-0.01158)	(-0.00558)	(-0.00210)
								0.7636	0.6949	0.7030		0.2182	0.3475	0.1455
								(0.02806)	(0.02674)	(0.01253)		(-0.00925)	(-0.00600)	(-0.00460)
											=	0.3475	0.2586	0.1980
												(-0.02410)	(-0.01540)	(-0.01541)
												0.1212	0.1677	<
												(-0.03122)	(-0.02624)	(-0.02365)
												>	0.1273	0.1333
												(-0.03730)	(-0.03274)	(-0.01713)

Table C.7: Power comparison of tests on 3×2 tables with sample sizes (5.8) (`LP1 S_chi`, `LP1 S_V`, `LP2 S_P`, and `LP2 S_chi`)

LP1 S_chi			LP1 S_V			LP2 S_P			LP2 S_chi		
=	<	0.3273	0.1333	=	<	0.2707	0.0848	=	0.0323	0.2283	0.1374	=	0.1495	0.2707	0.3919
		(-0.02299)	(-0.00906)			(-0.01244)			(-0.02839)	(-0.01494)	(-0.01499)		(-0.03401)	(-0.01504)	(-0.01505)		(-0.02249)	(-0.01494)	(-0.00505)
	0.0545	<	0.0929		0.0162	0.0828	0.1010		0.0404	0.1475	0.2222		0.0404	0.0485	0.1091
	(-0.02917)	(-0.02100)	(-0.01089)		(-0.03268)	(-0.02560)	(-0.01756)		(-0.03047)	(-0.02161)	(-0.00634)		(-0.03047)	(-0.01909)	(-0.00594)
	0.2061	0.0606	0.0848		0.2061	0.1091	0.1333		0.2061	0.0444	0.1333		0.2061	0.0404	0.0121
	(-0.01902)	(-0.02107)	(-0.01244)		(-0.01902)	(-0.02556)	(-0.01568)		(-0.01902)	(-0.01819)	(-0.00329)		(-0.01902)	(-0.01701)	(-0.00680)
<	0.0808	<	0.0081	<	0.0525	0.1293	<	0.0444	<	0.0323	<	0.0525	0.1616	0.3071
(-0.03246)	(-0.02919)	(-0.03054)	(-0.01321)	(-0.03246)	(-0.03459)	(-0.03641)	(-0.01576)	(-0.03246)	(-0.04021)	(-0.03652)	(-0.01582)	(-0.03246)	(-0.02869)	(-0.03641)	(-0.00582)
	0.0667	0.3758	0.3192		0.0081	0.0202	0.0889		0.0566	0.0444	0.4202		0.0566	0.3354	0.4081
	(-0.03550)	(-0.01101)	(-0.00566)		(-0.03901)	(-0.01560)	(-0.01233)		(-0.03680)	(-0.01161)	(-0.00111)		(-0.03680)	(-0.00910)	(-0.00071)
	<	0.1576	<		<	0.0040	<		<	0.3051	0.3394		<	0.3535	0.2667
	(-0.05001)	(-0.00791)	(-0.01588)		(-0.05001)	(-0.01240)	(-0.01912)		(-0.05001)	(-0.00503)	(-0.00673)		(-0.05001)	(-0.00385)	(-0.01024)
=	0.4485	0.3838	0.3455	=	0.3515	0.1616	0.2970	=	0.2828	0.4465	0.2061	=	0.2626	0.1616	0.6465
	(0.00162)	(-0.00636)	(-0.00089)		(-0.00379)	(-0.01224)	(-0.00344)		(-0.00941)	(-0.01234)	(-0.00350)		(0.00211)	(-0.01224)	(0.00650)
	0.4283	0.5091	0.5293		0.1172	0.3677	0.2061		0.6343	0.3939	0.6182		0.6343	0.5273	0.5758
	(0.00293)	(0.00095)	(0.00293)		(-0.00058)	(-0.00365)	(-0.00374)		(0.00163)	(0.00034)	(0.00748)		(0.00163)	(0.00286)	(0.00788)
	0.1697	0.4566	0.2545		0.1697	0.3394	0.0970		0.1697	0.4444	0.5576		0.1697	0.5737	0.4606
	(-0.01193)	(-0.00031)	(-0.00365)		(-0.01193)	(-0.00481)	(-0.00689)		(-0.01193)	(0.00256)	(0.00550)		(-0.01193)	(0.00374)	(0.00199)
=	0.1899	0.3212	0.2343	=	0.3596	0.2242	0.2586	=	0.1616	0.3838	0.1414	=	0.1616	0.2242	0.3515
	(-0.01782)	(-0.00888)	(-0.00610)		(-0.02322)	(-0.01476)	(-0.00864)		(-0.02884)	(-0.01486)	(-0.00870)		(-0.01732)	(-0.01476)	(0.00130)
	0.3152	0.1192	0.0687		0.2141	0.1354	0.0242		0.2303	0.1333	0.1091		0.2303	0.1414	0.0889
	(-0.01507)	(-0.02160)	(-0.02308)		(-0.01858)	(-0.02620)	(-0.02975)		(-0.01637)	(-0.02221)	(-0.01853)		(-0.01637)	(-0.01969)	(-0.01813)
	0.0848	0.1293	<		0.0848	0.0808	0.0242		0.0848	0.1899	0.6061		0.0848	0.2182	0.4121
	(-0.01063)	(-0.01750)	(-0.00459)		(-0.01063)	(-0.02200)	(-0.00783)		(-0.01063)	(-0.01463)	(0.00456)		(-0.01063)	(-0.01345)	(0.00105)
=	0.2101	0.1818	0.3475	=	0.3071	0.1253	0.3273	=	0.2626	0.4000	0.2707	=	0.1657	0.1253	0.5859
	(-0.00529)	(-0.01019)	(-0.00025)		(-0.01070)	(-0.01606)	(-0.00280)		(-0.01631)	(-0.01617)	(-0.00286)		(-0.00479)	(-0.01606)	(0.00714)
	0.4283	0.4929	0.5293		0.1172	0.2586	0.2061		0.6343	0.3758	0.6182		0.6343	0.5293	0.5758
	(0.00293)	(-0.00137)	(0.00293)		(-0.00058)	(-0.00597)	(-0.00374)		(0.00163)	(-0.00198)	(0.00748)		(0.00163)	(0.00054)	(0.00788)
	0.1697	0.4081	0.2545		0.1697	0.2970	0.0970		0.1697	0.4465	0.5576		0.1697	0.5556	0.4606
	(-0.01193)	(-0.00236)	(-0.00365)		(-0.01193)	(-0.00685)	(-0.00689)		(-0.01193)	(0.00052)	(0.00550)		(-0.01193)	(0.00170)	(0.00199)
=	=	<	0.1414	=	0.3596	<	<	=	0.2182	<	<	=	0.4424	<	<
		(-0.08735)	(-0.06583)		(-0.00541)	(-0.07807)	(-0.06838)		(-0.01102)	(-0.08244)	(-0.06844)		(0.00050)	(-0.00563)	(-0.05844)
	0.3293	0.1556	0.1818		0.2424	0.1253	0.1535		0.1333	0.1899	0.3697		0.1333	0.1717	0.2949
	(-0.00883)	(-0.01579)	(-0.00739)		(-0.01234)	(-0.02038)	(-0.01406)		(-0.01013)	(-0.01639)	(-0.00285)		(-0.01013)	(-0.01388)	(-0.00245)
	0.2061	0.0162	0.1455		0.2061	0.0061	0.1697		0.2061	0.0323	0.4970		0.2061	0.0040	0.1697
	(-0.01902)	(-0.02520)	(-0.00826)		(-0.01902)	(-0.02682)	(-0.01150)		(-0.01902)	(-0.02233)	(0.00089)		(-0.01902)	(-0.02115)	(-0.00262)
=	0.1980	0.1455	0.1414	=	0.3636	0.2545	0.2424	=	0.1778	0.2263	0.1778	=	0.0848	0.2545	0.5232
	(-0.01203)	(-0.01161)	(-0.00759)		(-0.01743)	(-0.01749)	(-0.01014)		(-0.02305)	(-0.01759)	(-0.01020)		(-0.01153)	(-0.01749)	(-0.00020)
	0.3879	0.5636	0.6061		0.2424	0.3232	0.3838		0.2788	0.4040	0.6384		0.2788	0.5414	0.6747
	(-0.00807)	(-0.00120)	(0.00417)		(-0.01158)	(-0.00579)	(-0.00249)		(-0.00937)	(-0.00180)	(0.00872)		(-0.00937)	(0.00071)	(0.00912)
	0.2182	0.3717	0.4485		0.2182	0.3636	0.1455		0.2182	0.4566	0.6424		0.2182	0.4485	0.5455
	(-0.00925)	(-0.00101)	(-0.00136)		(-0.00925)	(-0.00550)	(-0.00460)		(-0.00925)	(0.00187)	(0.00779)		(-0.00925)	(0.00305)	(0.00428)
=	0.2182	0.2869	0.1697	=	0.3596	<	0.1980	=	0.1939	0.2667	0.0768	=	0.2747	<	0.5192
	(-0.01573)	(-0.01516)	(-0.01286)		(-0.02114)	(-0.02103)	(-0.01541)		(-0.02675)	(-0.02114)	(-0.01547)		(-0.01523)	(-0.02103)	(-0.00546)
	0.2283	0.1798	0.1939		0.1212	0.1091	0.0848		0.2000	0.2545	0.3253		0.2000	0.2707	0.3434
	(-0.02771)	(-0.02185)	(-0.01737)		(-0.03122)	(-0.02645)	(-0.02404)		(-0.02901)	(-0.02246)	(-0.01283)		(-0.02901)	(-0.01994)	(-0.01243)
	<	0.2869	0.1091		<	0.0808	0.1333		<	0.2626	0.4727		<	0.2949	0.4242
	(-0.03730)	(-0.02775)	(-0.01389)		(-0.03730)	(-0.03224)	(-0.01713)		(-0.03730)	(-0.02487)	(-0.00474)		(-0.03730)	(-0.02369)	(-0.00825)
=	0.5717	0.5394	0.6444	=	0.3475	0.3677	=	=	0.5152	0.4970	0.2101	=	0.4525	0.3677	0.5980
	(0.00837)	(0.00024)	(0.00255)		(0.00297)	(-0.00563)			(-0.00265)	(-0.00574)	(-0.00006)		(0.00887)	(-0.00563)	(0.00994)
	0.4202	0.5576	0.6303		=	0.3495	0.2909		0.6101	0.6162	0.6646		0.6101	0.5313	0.6424
	(0.00351)	(0.00439)	(0.00628)			(-0.00021)	(-0.00039)		(0.00221)	(0.00378)	(0.01082)		(0.00221)	(0.00630)	(0.01122)
	=	0.4667	0.6606		=	0.3152	=		=	0.5091	0.7030		=	0.4848	0.6667
		(0.00499)	(0.00324)			(0.00050)				(0.00787)	(0.01239)			(0.00905)	(0.00888)
				=	0.3596	0.3798	0.3010	=	0.2182	0.4606	0.2505	=	0.4424	0.3798	0.6505
					(-0.00541)	(-0.00587)	(-0.00255)		(-0.01102)	(-0.00598)	(-0.00261)		(0.00050)	(-0.00587)	(0.00740)
					0.2343	0.3717	0.3394		0.5313	0.3596	0.6626		0.5313	0.3616	0.6990
					(-0.00351)	(-0.00460)	(-0.00667)		(-0.00130)	(-0.00061)	(0.00455)		(-0.00130)	(0.00191)	(0.00495)
					=	0.2424	0.2848		=	0.5495	0.7030		=	0.5172	0.6667
						(-0.00449)	(-0.00324)			(0.00288)	(0.00915)			(0.00406)	(0.00564)
								=	0.3838	0.4727	0.2101	=	0.5899	=	0.5980
									(-0.00562)	(-0.00011)	(-0.00006)		(0.00590)		(0.00994)
									0.6101	0.6081	0.6828		0.6101	0.6465	0.6909
									(0.00221)	(0.00399)	(0.01122)		(0.00221)	(0.00651)	(0.01161)
									=	0.5091	0.7030		=	0.5091	0.6667
										(0.00737)	(0.01239)			(0.00855)	(0.00888)
												=	0.4949	0.4727	0.6768
													(0.01152)	(0.00011)	(0.01000)
													=	0.5616	0.5293
														(0.00252)	(0.00040)
														0.7394	0.2364
														(0.00118)	(-0.00351)

Table C.8: Power comparison of tests on 3×2 tables with sample sizes (5.8) (`LP2 S_V`)

		LP2 S_V		
PEARSON	=	0.0485	0.2283	0.1374
		(-0.02627)	(-0.01504)	(-0.01505)
		0.0404	0.1394	0.2222
		(-0.03047)	(-0.02158)	(-0.00634)
		0.2061	0.0444	0.0970
		(-0.01902)	(-0.01819)	(-0.00463)
FISHER	< (-0.03246)	0.1333	<	0.0323
		(-0.03247)	(-0.03652)	(-0.01582)
		0.0566	0.0727	0.4202
		(-0.03680)	(-0.01158)	(-0.00111)
		<	0.3051	0.3152
		(-0.05001)	(-0.00503)	(-0.00807)
C S_P M	=	0.3596	0.4465	0.2061
		(-0.00166)	(-0.01234)	(-0.00350)
		0.6343	0.3919	0.6182
		(0.00163)	(0.00037)	(0.00748)
		0.1697	0.4444	0.5091
		(-0.01193)	(0.00256)	(0.00416)
C S_chi M	=	0.2586	0.3838	0.1414
		(-0.02110)	(-0.01486)	(-0.00870)
		0.2303	0.1697	0.1091
		(-0.01637)	(-0.02218)	(-0.01853)
		0.0848	0.1899	0.5939
		(-0.01063)	(-0.01463)	(0.00322)
C S_V M	=	0.1495	0.4000	0.2707
		(-0.00857)	(-0.01617)	(-0.00286)
		0.6343	0.3192	0.6182
		(0.00163)	(-0.00195)	(0.00748)
		0.1697	0.4465	0.5091
		(-0.01193)	(0.00052)	(0.00416)
ET chi	=	0.3596	<	<
		(-0.00328)	(-0.08224)	(-0.06844)
		0.1333	0.2020	0.3697
		(-0.01013)	(-0.01636)	(-0.00285)
		0.2061	0.0323	0.4485
		(-0.01902)	(-0.02233)	(-0.00045)
ET fisher	=	0.2586	0.2263	0.1778
		(-0.01531)	(-0.01759)	(-0.01020)
		0.2788	0.3152	0.6384
		(-0.00937)	(-0.00177)	(0.00872)
		0.2182	0.4566	0.5939
		(-0.00925)	(0.00187)	(0.00645)
ET vol	=	0.3596	0.2667	0.0768
		(-0.01901)	(-0.02114)	(-0.01547)
		0.2000	0.1778	0.3253
		(-0.02901)	(-0.02243)	(-0.01283)
		<	0.2626	0.4606
		(-0.03730)	(-0.02487)	(-0.00608)
LP1 S_P	=	0.4566	0.4970	0.2101
		(0.00509)	(-0.00574)	(-0.00006)
		0.6101	0.5677	0.6646
		(0.00221)	(0.00381)	(0.01082)
		=	0.5091	0.7152
			(0.00787)	(0.01105)
LP1 S_chi	=	0.3596	0.4606	0.2505
		(-0.00328)	(-0.00598)	(-0.00261)
		0.5313	0.3737	0.6626
		(-0.00130)	(-0.00057)	(0.00455)
		=	0.5495	0.7091
			(0.00288)	(0.00781)
LP1 S_V	=	0.5899	0.4727	0.2101
		(0.00213)	(-0.00011)	(-0.00006)
		0.6101	0.5414	0.6828
		(0.00221)	(0.00402)	(0.01122)
		=	0.5091	0.7152
			(0.00737)	(0.01105)
LP2 S_P	=	0.4848	=	=
		(0.00774)		
		=	0.3293	=
			(0.00003)	
		=	=	0.2182
				(-0.00134)
LP2 S_chi	=	0.4263	0.4727	0.2687
		(-0.00378)	(-0.00011)	(-0.01000)
		=	0.3798	0.4162
			(-0.00249)	(-0.00040)
		=	0.2061	0.7394
			(-0.00118)	(0.00217)

C.5 Tables for power comparison on 2×3, 3×3, and 2×4 tables

Table C.9: Power comparison of tests on 2×3, 3×3, and 2×4 tables with sample sizes (5.9) (non-LP tests)

	FISHER		C S_P M		C S_V M		ET chi		ET fisher		ET vol	
PEARSON	<	0.5877	<	0.1654	<	0.5199	<	0.9425	<	0.2411	<	0.3695
	(-0.0064)	(-0.0018)	(-0.0155)	(-0.0133)	(-0.0064)	(0.0012)	(-0.0044)	(0.0067)	(-0.0108)	(-0.0109)	(-0.0108)	(-0.0074)
	<	0.06	<	0.02	<	0.0257	<	<	<	0.0143	<	0.0771
	(-0.0168)	(-0.0192)	(-0.0287)	(-0.0224)	(-0.0179)	(-0.0208)	(-0.0146)	(-0.0144)	(-0.0224)	(-0.0263)	(-0.0180)	(-0.0132)
	0.1848	<	<	<	0.024	<	<	<	<	<	0.0721	<
	(-0.0100)	(-0.0185)	(-0.0298)	(-0.0246)	(-0.0229)	(-0.0229)	(-0.0083)	(-0.0121)	(-0.0254)	(-0.0236)	(-0.0124)	(-0.0229)
FISHER			<	<	=	0.4126	0.4158	0.5757	<	<	<	0.1645
			(-0.0091)	(-0.0115)		(0.0030)	(0.0020)	(0.0085)	(-0.0044)	(-0.0091)	(-0.0044)	(-0.0056)
			<	0.1943	0.415	0.2829	0.5487	0.5343	<	<	0.2575	0.6171
			(-0.0120)	(-0.0032)	(-0.0011)	(-0.0016)	(0.0021)	(0.0049)	(-0.0057)	(-0.0071)	(-0.0012)	(0.0060)
			<	<	0.068	<	0.3619	0.7029	<	<	0.3789	<
			(-0.0198)	(-0.0061)	(-0.0129)	(-0.0044)	(0.0017)	(0.0064)	(-0.0154)	(-0.0051)	(-0.0024)	(-0.0044)
C S_P M					>	0.7909	0.9413	0.9411	0.7988	0.7123	0.7988	0.8765
					(0.0091)	(0.0146)	(0.0111)	(0.0200)	(0.0047)	(0.0025)	(0.0047)	(0.0060)
					0.8443	0.5229	0.8806	0.6286	0.7525	0.1486	0.8106	0.6857
					(0.0109)	(0.0016)	(0.0141)	(0.0081)	(0.0063)	(-0.0038)	(0.0107)	(0.0092)
					0.7718	0.8636	0.9566	>	0.7466	0.8052	0.9472	0.8636
					(0.0069)	(0.0017)	(0.0215)	(0.0125)	(0.0044)	(0.0010)	(0.0174)	(0.0017)
C S_V M							0.4158	0.6669	<	0.229	<	0.3578
							(0.0020)	(0.0054)	(-0.0044)	(-0.0121)	(-0.0044)	(-0.0086)
							0.6434	0.5457	0.3598	0.02	0.4906	0.7
							(0.0032)	(0.0065)	(-0.0046)	(-0.0055)	(-0.0001)	(0.0076)
							0.8534	>	0.4504	<	0.8065	=
							(0.0146)	(0.0108)	(-0.0025)	(-0.0006)	(0.0105)	
ET chi									<	0.088	<	0.19
									(-0.0064)	(-0.0176)	(-0.0064)	(-0.0141)
									0.2176	0.3	0.3446	0.6371
									(-0.0078)	(-0.0119)	(-0.0034)	(0.0011)
									0.1067	<	0.546	<
									(-0.0171)	(-0.0115)	(-0.0041)	(-0.0108)
ET fisher											=	0.8762
												(0.0035)
											0.9044	0.9114
											(0.0044)	(0.0130)
											0.8897	0.8084
											(0.0130)	(0.0006)
ET vol												
LP1 S_P												
LP1 S_chi												
LP1 S_V												
LP2 S_P												
LP2 S_chi												

Table C.10: Power comparison of tests on 2×3, 3×3, and 2×4 tables with sample sizes (5.9) (LP tests)

		LP1 S_P	LP1 S_chi	LP1 S_V	LP2 S_P	LP2 S_chi	LP2 S_V
PEARSON	LP1 S_P	< 0.146 (-0.0180) < (-0.0150) (-0.0298) < (-0.0280) (-0.0340) < (-0.0227)	< 0.3499 (-0.0044) < (-0.0029) (-0.0185) < (-0.0198) 0.0188 (-0.0184) < (-0.0126)	< 0.1959 (-0.0108) < (-0.0123) 0.0296 (-0.0254) < (-0.0265) (-0.0309) < (-0.0229)	< 0.1595 (-0.0155) < (-0.0130) 0.0114 (-0.0298) < (-0.0286) (-0.0326) < (-0.0246)	< 0.3792 (-0.0044) < (-0.0018) (-0.0185) < (-0.0203) (-0.0188) < (-0.0126)	< 0.2053 (-0.0108) < (-0.0112) (-0.0263) < (-0.0251) (-0.0315) < (-0.0229)
	FISHER	< 0.4158 (-0.0116) < (-0.0132) (-0.0131) < (-0.0087) (-0.0240) < (-0.0042)	< 0.2848 (0.0020) < (-0.0011) 0.4065 (-0.0017) < (-0.0006) 0.1889 (-0.0084) < (0.0059)	< 0.0566 (-0.0044) < (-0.0105) 0.2367 (-0.0087) < (-0.0072) 0.0141 (-0.0209) < (-0.0044)	< 0.0167 (-0.0091) < (-0.0112) 0.0636 (-0.0130) < (-0.0093) < 0.1795 (-0.0226) < (-0.0060)	< 0.3205 (0.0020) < (0.0000) 0.4065 (-0.0017) < (-0.0011) 0.1795 (-0.0088) < (0.0059)	< 0.0472 (-0.0044) < (-0.0094) 0.1372 (-0.0095) < (-0.0058) 0.0006 (-0.0215) < (-0.0044)
	C S_P M	< 0.4317 (-0.0025) < (-0.0017) 0.3015 (-0.0011) < (-0.0055) 0.2422 (-0.0042) < (0.0019)	< 0.293 (0.0111) < (0.0105) 0.8434 (0.0102) < (0.0026) 0.6839 (0.0114) < (0.0120)	< 0.7988 (0.0047) < (0.0010) 0.6953 (0.0033) < (-0.0040) 0.3302 (-0.0011) < (0.0017)	= 0.3457 < (-0.0010) 0.3326 (-0.0028) < (0.0000)	< 0.9413 (0.0111) < (0.0116) 0.8434 (0.0102) < (0.0021) 0.7496 (0.0110) < (0.0120)	< 0.7988 (0.0047) < (0.0116) 0.6035 (0.0024) < (0.0026) 0.3408 (-0.0017) < (0.0017)
	C S_V M	< 0.4158 (-0.0116) < (-0.0162) 0.1097 (-0.0120) < (0.02) 0.1689 (-0.0111) < (0.0002)	< 0.3554 (0.0020) < (-0.0041) 0.4554 (-0.0007) < (0.0010) 0.4651 (0.0045) < (0.0103)	< 0.1554 (-0.0044) < (-0.0136) 0.2836 (-0.0076) < (-0.0057) 0.1566 (-0.0080) =	< 0.1789 (-0.0091) < (-0.0142) 0.1463 (-0.0119) < (-0.0078) 0.1466 (-0.0097) < (-0.0016)	< 0.4158 (0.0020) < (-0.0030) 0.4554 (-0.0007) < (0.0005) 0.4733 (0.0041) < (0.0103)	< 0.4018 (-0.0044) < (-0.0124) 0.2774 (-0.0085) < (-0.0043) 0.149 (-0.0086) =
	ET chi	< 0.2616 (-0.0136) < (-0.0217) 0.0836 (-0.0152) < (-0.0136) 0.0358 (-0.0257) < (-0.0106)	< 0.0349 (-0.0095) < (-0.0095) 0.1968 (-0.0039) < (-0.0055) 0.1255 (-0.0102) < (-0.0005)	< 0.0683 (-0.0064) < (-0.0190) 0.1642 (-0.0108) < (-0.0121) 0.0381 (-0.0226) < (-0.0108)	< 0.0449 (-0.0111) < (-0.0197) 0.0434 (-0.0151) < (-0.0142) 0.0246 (-0.0243) < (-0.0125)	< 0.0915 (-0.0085) < (-0.0085) 0.1968 (-0.0039) < (-0.0060) 0.1144 (-0.0105) < (-0.0005)	< 0.0777 (-0.0064) < (-0.0179) 0.1217 (-0.0117) < (-0.0107) 0.0199 (-0.0232) < (-0.0108)
	ET fisher	< 0.2915 (-0.0072) < (-0.0041) 0.0466 (-0.0074) < (-0.0017) 0.0845 (-0.0086) < (0.0009)	< 0.0833 (0.0064) < (0.0080) 0.2143 (0.0039) < (0.0065) 0.5886 (0.0070) < (0.0110)	= 0.3132 (-0.0014) 0.3809 (-0.0030) < (-0.0002) 0.1179 (-0.0055) < (0.0006)	< 0.305 (-0.0047) < (-0.0021) 0.2041 (-0.0073) < (-0.0023) 0.105 (-0.0072) < (-0.0010)	< 0.9413 (0.0064) < (0.0091) 0.5886 (0.0039) < (0.0060) 0.5155 (0.0066) < (0.0110)	= 0.7375 (0.0091) < (-0.0003) 0.3311 (-0.0039) < (0.0012) 0.1185 (-0.0061) < (0.0006)
	ET vol	< 0.2915 (-0.0072) < (0.0443) 0.0771 (-0.0118) < (-0.0147) 0.0487 (-0.0216) < (0.0002)	< 0.9413 (0.0064) < (0.0045) 0.4721 (-0.0005) < (-0.0066) 0.2792 (-0.0060) < (0.0103)	= 0.1012 (-0.0047) < (-0.0056) 0.3164 (-0.0074) < (-0.0132) 0.0481 (-0.0185) =	< 0.1012 (-0.0047) < (-0.0056) 0.1557 (-0.0118) < (-0.0153) 0.0499 (-0.0202) < (-0.0016)	< 0.9413 (0.0064) < (0.0056) 0.4721 (-0.0005) < (-0.0071) 0.2604 (-0.0064) < (0.0103)	< 0.571 (0.0056) = 0.24 (-0.0071) < (-0.0083) 0.9643 (0.0103) =
	LP1 S_P	0.7009 (0.0136) 0.8282 (0.0113) 0.8428 (0.0155)	0.8223 (0.0121) 0.8314 (0.0081) 0.9156 (0.0101)	0.6657 (0.0072) 0.7041 (0.0044) 0.6182 (0.0031)	0.5255 (0.0025) 0.5021 (0.0001) 0.624 (0.0014)	0.7009 (0.0136) 0.8282 (0.0113) 0.8557 (0.0152)	0.844 (0.0072) 0.6692 (0.0035) 0.5924 (0.0025)
	LP1 S_chi			< 0.1897 (-0.0064) < (-0.0095) 0.3082 (-0.0069) < (-0.0066) 0.2622 (-0.0125) < (-0.0103)	< 0.1935 (-0.0111) < (-0.0101) 0.0757 (-0.0113) < (-0.0087) 0.1859 (-0.0142) < (-0.0120)	= 0.6367 (0.0136) = 0.06 (-0.0006) 0.3133 (-0.0018) 0.4258 (-0.0004) = 0.4205	< 0.6469 (0.0011) 0.4571 (-0.0005) = 0.8062 (0.0106)
	LP1 S_V						
	LP2 S_P						
	LP2 S_chi						

C.6 Long-term power comparison

Table C.11: Long-term power comparison for indicated group and table sizes.

	PEARSON	FISHER	C S_P M	C S_chi M	C S_V M	ET chi	ET fisher	ET vol	LP1 S_P	LP1 S_chi	LP1 S_V	LP2 S_P	LP2 S_chi	LP2 S_V
$(5, 5)$; 2	0.0000	**0.0556**	**0.0556**	**0.0556**	**0.0556**	**0.0556**	**0.0556**	**0.0556**	**0.0556**	**0.0556**	**0.0556**	**0.0556**	**0.0556**	**0.0556**
$(10, 5)$; 2	0.0909	0.1212	**0.1818**	**0.1818**	**0.1818**	**0.1818**	**0.1818**	0.1515	**0.1818**	**0.1818**	**0.1818**	**0.1818**	**0.1818**	**0.1818**
$(10, 10)$; 2	0.1653	0.1653	**0.2479**	**0.2479**	**0.2479**	0.2314	0.2314	0.2314	**0.2479**	**0.2479**	**0.2479**	**0.2479**	**0.2479**	**0.2479**
$(20, 5)$; 2	0.1429	0.1746	**0.2222**	**0.2222**	**0.2222**	0.1746	**0.2222**	**0.2222**	**0.2222**	**0.2222**	**0.2222**	**0.2222**	**0.2222**	**0.2222**
$(20, 10)$; 2	0.2424	0.2944	**0.3290**	0.3117	**0.3290**	0.3203	0.3203	0.3203	**0.3290**	**0.3290**	**0.3290**	**0.3290**	**0.3290**	**0.3290**
$(20, 20)$; 2	0.3447	0.3719	**0.4172**	**0.4172**	**0.4172**	**0.4172**	**0.4172**	0.4082	**0.4172**	**0.4172**	**0.4172**	**0.4172**	**0.4172**	**0.4172**
$(40, 5)$; 2	0.1870	0.2114	**0.2683**	**0.2683**	**0.2683**	0.1951	**0.2683**	**0.2683**	**0.2683**	**0.2683**	**0.2683**	**0.2683**	**0.2683**	**0.2683**
$(40, 10)$; 2	0.3104	0.3370	**0.3858**	0.3725	**0.3858**	0.3237	0.3725	0.3769	**0.3858**	0.3769	**0.3858**	**0.3858**	0.3769	**0.3858**
$(40, 20)$; 2	0.4367	0.4576	**0.4901**	0.4344	**0.4901**	0.4878	0.4855	0.4785	**0.4901**	**0.4901**	**0.4901**	**0.4901**	**0.4901**	**0.4901**
$(40, 40)$; 2	0.5306	0.5306	**0.5699**	**0.5699**	**0.5699**	0.5651	0.5651	0.5651	**0.5699**	**0.5699**	**0.5699**	**0.5699**	**0.5699**	**0.5699**
$(5, 5, 5)$; 2	**0.1667**	0.1389	**0.1667**	**0.1667**	**0.1667**	**0.1667**	**0.1667**	**0.1667**	**0.1667**	**0.1667**	**0.1667**	**0.1667**	**0.1667**	**0.1667**
$(10, 5, 5)$; 2	0.2475	0.2424	0.2677	0.2576	0.2626	0.2677	0.2576	0.2525	**0.2727**	0.2677	**0.2727**	**0.2727**	0.2677	0.2677
$(10, 10, 5)$; 2	0.3306	0.3251	**0.3554**	0.3416	**0.3554**	0.3471	0.3471	0.3306	**0.3554**	0.3526	**0.3554**	0.3526	0.3526	0.3526
$(10, 10, 10)$; 2	0.4192	0.3877	0.4192	0.4237	0.4192	0.4192	0.4192	0.4012	**0.4282**	**0.4282**	**0.4282**	**0.4282**	**0.4282**	**0.4282**
$(20, 5, 5)$; 2	0.3386	0.3201	0.3492	0.3439	0.3466	0.2751	0.3413	0.3360	**0.3545**	**0.3545**	**0.3545**	0.3519	**0.3545**	0.3519
$(20, 10, 5)$; 2	0.4170	0.4242	0.4343	0.4199	0.4329	0.4228	0.4343	0.4170	**0.4387**	0.4372	**0.4387**	0.4372	0.4358	0.4372
$(20, 20, 5)$; 2	0.5079	0.5042	0.5185	0.5117	0.5185	0.4512	0.5117	0.5072	**0.5208**	0.5185	**0.5208**	0.5200	0.5163	0.5200
$(20, 10, 10)$; 2	0.4935	0.4990	0.5061	0.4896	0.5045	0.4880	0.5069	0.4856	**0.5116**	0.5092	**0.5116**	0.5085	0.5077	0.5085
$(20, 20, 10)$; 2	0.5760	0.5772	0.5846	0.5677	0.5846	0.5801	0.5863	0.5669	**0.5879**	0.5854	**0.5879**	0.5821	0.5821	0.5821
$(20, 20, 20)$; 2	0.6459	0.6382	0.6524	0.6498	0.6524	0.6511	0.6518	0.6427	**0.6563**	0.6550	**0.6563**	0.6485	0.6524	0.6505
$(5, 5)$; 3	0.0136	0.0170	**0.0238**		0.0170	0.0170	0.0204	0.0204	**0.0238**	0.0170	0.0204	**0.0238**	0.0170	0.0204
$(10, 5)$; 3	0.0292	0.0400	0.0492		0.0422	0.0395	0.0438	0.0411	**0.0498**	0.0427	0.0471	**0.0498**	0.0427	0.0471
$(10, 10)$; 3	0.0561	0.0606	0.0723		0.0689	0.0620	0.0692	0.0623	**0.0744**	0.0682	0.0730	0.0733	0.0682	0.0730
$(20, 5)$; 3	0.0631	0.0592	0.0685		0.0606	0.0581	0.0665	0.0643	**0.0694**	0.0645	0.0677	0.0686	0.0639	0.0674
$(5, 5, 5)$; 3	0.0136	0.0194	0.0226		0.0210	0.0207	0.0232	0.0192	**0.0247**	0.0227	0.0237	0.0243	0.0226	0.0235
$(5, 5)$; 4	0.0000	0.0018	**0.0027**		0.0022	0.0013	0.0024	0.0022	**0.0027**	0.0015	0.0022	**0.0027**	0.0015	0.0022