NTNU
Norwegian University of
Science and Technology

Testing equality of the success probabilities in two independent binomial distributions

Fredrik Lohne Aanes

Master of Science in Physics and Mathematics
Submission date: June 2016
Supervisor: Mette Langaas, MATH
Co-supervisor: Øyvind Bakke, MATH

Norwegian University of Science and Technology
Department of Mathematical Sciences

Abstract

In this thesis we compare the power of tests of equality of success probabilities in two independent binomial distributions for different sample sizes and significance levels. We calculate the realisations of different p-values using enumeration and the power functions are also evaluated using enumeration. We consider four different methods for creating a p-value, the asymptotic (A) method, the estimation (E) method, the conditional (C) method and the maximization (M) method. We also consider combining the C, E, A or M method with the M method, resulting in the C $\circ$ M, E $\circ$ M, A $\circ$ M or M^2 method. We emphasize that each p-value is a random variable and also stress the concept of validity. A valid p-value may be used to construct a test that never exceeds the significance level under H_0. We found the E and A p-values not to be valid.

For the power study under the alternative hypothesis we only considered the power functions of the tests based on the remaining valid p-values. The power functions where only evaluated at points at which at least one power function is above or equal to 80%. The C $\circ$ M p-value and the E $\circ$ M p-value are found to give level α tests that are the most powerful. The C $\circ$ M p-value is guaranteed to give a test with power that uniformly equal to or greater than the test based on the C p-value. The power increase is quite substantial under the alternative hypothesis for the smallest sample sizes studied, both for the unbalanced design and the balanced design. For the largest sample sizes studied the power functions of the test based on the C $\circ$ M, E $\circ$ M and C p-values are found to take nearly the same values in a majority of the parameters considered. Since the C method is less computer intensive than the M method, we recommend using the C method for large sample sizes. We recommend using difference plots to easily establish where in parameter space the differences between the power functions occur and check if the differences are in interesting parts of the parameter space. We also recommend using the cumulative difference function in power studies.

Sammendrag

I denne masteroppgaven sammenligner vi styrken til tester av likhet av suksessannsynlighetene i to uavhengige binomiske fordelinger for forskjellige utvalgsstørrelser og for forskjellige signifikansnivå. Vi beregner realisasjonene til forskjellige p-verdier ved å bruke enumerering. Styrkefunksjonene blir også evaluert ved å bruke enumerering. Vi betrakter fire metoder for å generere en p-verdi, den asymptotiske (A) metoden, estimeringsmetoden (E), den betingede (C) metoden og maksimeringsmetoden (M). Vi ser også på kombinasjoner av A, C, E eller metoden med M metoden, som resulterer i A$\circ$M, C$\circ$M, E$\circ$M eller M^2 metoden. Vi understreker at p-verdien er en stokastisk variabel og framhever også begrepet validitet. En "valid" p-verdi kan brukes til å lage en test som aldri overstiger signifikansnivået under nullhypotesen. Både A og E p-verdiene er funnet til å være ikke "valid".

I styrkestudien under den alternative hypotesen betraktet vi kun styrkefunksjonene til tester basert på de resterende "valid" p-verdiene. Vi betraktet kun styrkefunksjonene i parameterpunkter hvor styrken til minst en av funksjonene er 80 % eller høyrere. Vi har funnet at C$\circ$M og E$\circ$M p-verdiene gir nivå α tester med størst styrke. Det er garantert at C$\circ$M p-verdien gir en test med uniformt minst like høy styrke som styrken til testen basert på C p-verdien. Økningen i teststyrke under den alternative hypotesen er ganske stor for de minste studerte utvalgsstørrelsene, både for balanserte og ubalanserte design. For de største studerte utvalgsstørrelsene er det funnet at styrkefunksjonene til testene basert på C, C$\circ$M og E$\circ$M p-verdiene tar tilnærmet de samme verdiene i storparten av de studerte punktene i parameterrommet. Siden C metoden er mindre beregningskrevende enn M metoden, anbefaler vi å bruke C metoden for store utvalgsstørrelser. Vi anbefaler å bruke differanseplott for enkelt å finne ut hvor i parameterrommet det er forskjeller i styrken til de ulike testene og finne ut om forskjellene er i interessante deler av parameterrommet. Vi anbefaler også å bruke den kumulative differansefunksjonen i styrkestudier.

Preface

This thesis finishes my Master's degree in Applied Physics and Mathematics with specialization in Industrial Mathematics. The work has taken place in the tenth semester at the Norwegian University of Science and Technology, from the end of January 2016 to the end of June 2016.

This thesis is not an extension of the master project carried out in the fall semester of 2015. The results of my thesis are to be part of Bakke & Langaas (n.d.). I would like to thank my supervisors Mette Langaas and Øyvind Bakke for this opportunity. Throughout the semester I have received excellent guidance from them as well and would like to thank them for the great collaboration.

Contents

List of abbreviations

A	asymptotic method
AM	A∘M, combination of the asymptotic method with the maximization method
C	conditional method
CM	C∘M, combination of the conditional method with the maximization method
E	estimation method
EM	E∘M, combination of the estimation method with the maximization method
M	maximization method
mle	maximum likelihood estimator
pmf	probability mass function
pp	percentage points

Chapter 1

Introduction

In this chapter we present the objective of this thesis and give an overview of the different chapters.

1.1 Problem description

Background The simplest situation in hypothesis testing is when the null hypothesis specifies only one point. Such a null hypothesis is called simple. However, when testing equality of the success probabilities in two independent binomial distributions, θ_1 and θ_2 respectively, the null hypothesis takes the form $H_0 : \theta_1 = \theta_2 = \theta$. Now infinitely many values are possible for the common value θ. Such a null hypothesis is called composite.

There exist different methods for creating tests of composite hypotheses in discrete distributions. All of them are based on a test statistic. It is also possible to transform this original statistic. The p-value is an example of such a transformation. The p-value can be computed using different methods. We first consider four different methods of creating p-values when testing the specified composite null hypothesis. One method uses the asymptotic distribution of the original test statistic, which we call the A method. Another method, the E method, calculates the maximum likelihood estimator of θ under the null hypothesis and analyse the original test statistic based on this value. When using the tests resulting from these methods there is no guarantee that the maximum probability of the type I error is below or equal to a specified significance level α. One method that gives a test with this property is given in Theorem 8.3.27 in Casella & Berger (2002, p. 397). We call it the M method. Another method that gives a test with this

property conditions the original test statistic on a sufficient statistic for the parameter θ under the null hypothesis. We call this the C method. It is also possible to combine two methods. You then use the p-value generated from one method as test statistic in the other method. You can for instance let the p-value from the A method serve as test statistic in the M method. We consider combining the A, E, C or M method with the M method.

Objective In the master thesis we consider testing equality of the success probability in two independent binomial distributions. We base the level α tests on the p-values generated from the four mentioned methods (A, E, C and M) and also the p-values resulting from letting the p-value from either the A, E, M or C method serve as test statistic in the M method (i.e. we combine the A, C, M or E method with the M method). We study and compare the power functions of these tests for different α.

1.2 Overview of the different chapters

In Chapter 2 we review the traditional Neyman–Pearson approach to statistical hypothesis testing and in Chapter 3 we review the p-value approach to hypothesis testing when the null hypothesis is simple. In Chapter 4 we present the theory needed for the power study in Chapter 5. We start Chapter 4 with presenting and giving examples of the main topic of this thesis, testing equality of the success probabilities in two independent binomial distributions. We then present the four different methods (A, C, E, M) of creating p-values. One of the main goals of this chapter is to familiarise the different methods of creating a p-value. We do this for instance by showing how to calculate the p-value using each method, by showing how to evaluate the power functions of tests based on the p-values and by comparing the p-values from the different methods and trying to see how the differences in the realisations of the p-values affect the power functions. We also present the theory behind the methods and also show how we can simplify the calculations in the power study.

The other main goal of Chapter 4 is to develop an understanding of the role of the test statistic. We do this in several steps. We will see that the p-value from one method can be regarded as a reasonable test statistic on its own. This means we can use the p-value from one method as test statistic in another method. We also consider how we should combine different methods and look at properties of some of the combinations.

Chapter 2

Neyman–Pearson approach to statistical hypothesis testing: a review

In this chapter we give a review of the steps in the traditional Neyman–Pearson approach to statistical hypothesis testing.

2.1 Overview

The main steps of the Neyman–Pearson approach to statistical hypothesis testing may be summarised in the following steps (which are loosely inspired from Oyana & Margai (2016, p. 107–110) and from Borgan (2007)):

1. Formulate research hypothesis
2. Plan an experiment
3. Specify the null and alternative hypothesis
4. Choose an appropriate test and select the significance level
5. Perform the experiment and reject or accept the null hypothesis based on the outcome of the test

In the next sections we take a closer look at the different steps.

2.2 Research hypothesis

The research hypothesis is what the investigator(s) thinks is the relationship between two or more variables (Kiess 2007). We give two examples that we study in this section. In the first example, we refer to it as example (a), two populations are considered. One of the populations has brain cancer of type I and the other has brain cancer of type II (Bland & Altman 2004). The investigators believe that the lifetime of the two populations are different (i.e. the research hypothesis). In the other example, example (b), a manufacturer of an ointment against a hand rash has developed a new ointment (Borgan 2007). The manufacturer thinks that the new product is better than the old one.

2.3 Plan an experiment

There are two main types of "experiments": observational experiments and randomized experiments (Altman 1991, p. 74–76). In observational experiments the investigators only observe the subjects under consideration and measure variables of interest without assigning treatments to the subjects. Differences between subjects already exist and are beyond the control of the investigators. In randomized experiments the investigators assign treatments to the subjects randomly and observe the effect of the treatments on them.

In example (a), the investigators follow 20 subjects with type I tumour and 31 subjects with the other type of tumour. This is thus an observational study. It would be considered unethical to inflict study subjects with either kind of brain cancer. In example (b) the manufacturer studies 16 subjects and asks them to apply a chemical agent on both hands so that they develop rashes. Afterwards for each study subject, one of the hands is randomly treated with one of the ointments and the other hand is treated with the other ointment. The manufacturer keeps track of which ointment has been used on each hand. This study is therefore a randomized experiment.

To be able to use statistical hypothesis testing the investigators *must* consider the outcomes in the experiments as realisations of random variables. In example (a) the investigators treat the observed lifetimes Y_{1i} of the study subjects from population with tumour of type I as independent realisations of a random variable T_1, which has survival function $R_1(t) = \Pr(T_1 > t)$. Similarly they treat the observed lifetimes Y_{2i} of the study subjects from population with tumour of type II as independent realisations of a random variable T_2, where T_2 has survival function $R_2(t) = \Pr(T_2 > t)$.

In example (b) the manufacturer hires a neutral doctor to decide after a specified amount of time which of the two hands have recovered the best from the rashes for each of the individuals in the experiment. The manufacturer defines a indicator variable for each individual i where $X_i = 1$ if the new ointment is best and 0 otherwise. It also lets $X = X_1 + \cdots + X_{16}$, i.e X is the number of individuals where the new ointment is best. In one run of the experiment the value of X could be 9, but if the manufacturer had repeated the experiment, it would most likely have obtained another value, say the number 7. It is therefore natural to consider X as a random variable, and this is necessary to be able to use statistical hypothesis testing to answer the research hypothesis. One possible model for X would be the binomial model with parameters $n = 16$ and unknown success probability $\theta \in [0, 1]$.

2.4 Null and alternative hypothesis

After modelling the outcomes in the experiment as random variables, the experimenters "translate" the research hypothesis into two contradictory hypotheses about quantities of the random variables (Borgan 2007). The two hypotheses are called the *null hypothesis* and the *alternative hypothesis*, denoted H_0 and H_1 respectively. Usually the values of the quantities specified in H_1 are the values which supports the research hypothesis and the values specified in the null hypothesis do not support the research hypothesis and usually these values support the hypothesis of "no difference" or "no improvement" (Devore et al. 2012, p. 427).

In example (a) the investigators formulate the null hypothesis and the alternative hypothesis as

$$H_0 : R_1(t) = R_2(t), \text{ for all } t, H_1 : R_1(t) \neq R_2(t) \text{ for at least one } t, \tag{2.1}$$

so that the null hypothesis is that there is no difference in survival between the two populations with different kinds of tumours while the alternative hypothesis is that there is a difference in survival between them.

In example (b) the manufacturer formulates the hypotheses

$$H_0 : \theta \leq \frac{1}{2}, H_1 : \theta > \frac{1}{2}, \tag{2.2}$$

where H_0 specifies no improvement (the old ointment is at least as good) and H_1 that the new product is better.

As a note, we have chosen to differentiate between the research hypothesis and the alternative hypothesis (Kiess 2007). This is not done in all textbooks, see for

instance Devore et al. (2012, p. 426), Casella & Berger (2002, p. 373) or Walpole et al. (2012, p. 321). There are two main reasons for this. The first one is that the statistical hypothesis test is pure mathematical; it only concerns quantities of theoretical models while the research hypothesis is about quantities from the real world. This means we differentiate between hypotheses about quantities in the real world and hypotheses about quantities in the chosen models of them. The model you choose may not model the real world quantities well enough and then the results of the hypothesis test tell you nothing about the research hypothesis. The second reason is that there may be more than one possible model of the real world quantities. Depending on which model you choose, the outcome of the hypothesis test may change. In theory, one model may provide test results that support the research hypothesis and another model may give test results that do not favour the research hypothesis.

2.5 Parametric and non-parametric statistical hypothesis tests

The hypotheses in Equation (2.1) and Equation (2.2) are of different kinds. The null hypothesis and alternative hypothesis in Equation (2.2) are about a parameter in a specified distribution, the binomial, while the hypotheses in Equation (2.1) do not involve parameters in a specified distribution. Statistical hypotheses about parameters in distributions are called parametric hypotheses (Stuart & Ord 1991, p. 795), while statistical hypotheses where no special distribution parameters are specified are called non-parametric hypotheses. This means the hypotheses in Equation (2.1) are non-parametric and the hypotheses in Equation (2.2) are parametric.

2.6 Choose an appropriate test and select the significance level

A test is a rule that tells us for which outcomes we should reject the null hypothesis and for which outcomes we should accept it (Casella & Berger 2002, p. 374). In example (a) the logrank-test may be appropriate to test the hypotheses in Equation (2.1) (Bland & Altman 2004). However, in this thesis we will be concerned with parametric hypothesis tests and will therefore not show how the log-rank test is carried out.

In example (b), a statistical test may be based on the value of X. A real valued function of the data that is used to tell for which outcomes we should reject the null hypothesis and for which outcomes we should accept it is called a test statistic (Casella & Berger 2002, p. 374). We denote the test statistic by the letter T and since is a function of the outcomes in the experiment $T = T(X)$. In this example we therefore have $T(X) = X$. Furthermore we divide the *sample space* S, i.e the space of all possible outcomes in the experiment, into two different regions, the *acceptance region* and the *rejection region* (Casella & Berger 2002, p. 374). (Some would call the set of all possible realisations of the random variable X the support of X. This is for example done in Taboga (2010), but we follow the convention made of Casella & Berger (2002, p. 27) where we regard the random variable as a function from the original sample space to a new sample space, the space of all the possible realisations of the random variable.). We denote the rejection region by R and the acceptance region by R^c (Casella & Berger 2002, p. 382-383). The rejection region consists of the outcomes for which the test rejects the null hypothesis and the acceptance region consists of the outcomes for which the test accepts the null hypothesis. We use the values of $T(X)$ to decide which outcomes should be in R or R^c. Large values of X indicate that the alternative hypothesis is true while small values indicate that the null hypothesis is true, so it seems reasonable to reject the null hypothesis for large values of X and not reject the null hypothesis for small values of X. Therefore small values of X should be in R^c while large values should be in R. Before be describe what the significance level is, we should take a closer look at the different situations that can occur when the test either accepts or rejects the null hypothesis.

Depending on whether H_0 or H_1 is true and the outcome of the test (reject H_0 or accept H_0) one of four possible situations will occur. They are shown in Table 2.1. We see that two of the situations are correct decisions while the other two are wrong decisions. We say that a *type I error* occurs when H_0 is correct but the test rejects the null hypothesis (Casella & Berger 2002, p. 382). A *type II error* occurs when H_1 is correct but the test accepts the null hypothesis (Casella & Berger 2002, p. 382). The two types of errors are common for both parametric and non-parametric hypothesis tests. In example (a) the logrank test commits a type II error if the survival functions of the two groups really are different but the test does not detect it. The logrank test commits a type I error if the survival functions of the two populations are the same but the test rejects the null hypothesis. In example (b) the test the manufacturer uses commits a type I error if it rejects the null hypothesis (so that the manufacturer thinks the new ointment is better and possibly accepts the research hypothesis) but really $\theta \in [0, \frac{1}{2}]$. The test commits a type II error if it accepts the null hypothesis (so that the manufacturer thinks the old ointment is the best and possibly rejects the research hypothesis), but

Table 2.1: The four different states that can occur when a statistical hypothesis test either rejects or accepts H_0 and in reality H_0 is either true or false. Only one of the states can be true for a given situation.

Test result	Truth	
	H_0 true	H_1 true
Accept H_0	No error	Type II error
Accept H_1	Type I error	No error

$\theta \in (\frac{1}{2}, 1]$.

When considering parametric hypotheses it is possible to specify the hypotheses as $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$ (Casella & Berger 2002, p. 373), where $\Theta_0 \cap \Theta_1 = \varnothing$ ($\varnothing$ is the empty set) and Θ_0 and Θ_1 are subsets of the parameter space. In many cases $\Theta_1 = \Theta_0^c$, i.e the complement of Θ_0, so that the union of Θ_0 and Θ_1 is the whole of the parameter space. We also let

$$\Pr_{\theta_0}(X \in C)$$

denote the probability that X is in the set C when θ_0 is the true value of the parameter θ (Hogg et al. 2014, p. 440). Then the probability of a type I error for each value of $\theta \in \Theta_0$ is $\Pr_\theta(\text{reject } H_0) = \Pr_\theta(X \in R)$, while the probability of a type II error for each value of $\theta \in \Theta_1$ is $\Pr_\theta(\text{accept } H_0) = \Pr_\theta(X \in R^c) \overset{(\star)}{=} 1 - \Pr_\theta(X \in R)$, where the equality in transition $(\star)$ follows since $S = R \cup R^c$ and $R \cap R^c = \varnothing$ (so that $\Pr_\theta(X \in S) = 1 = \Pr_\theta(X \in R) + \Pr_\theta(X \in R^c)$). Then

$$\Pr_\theta(X \in R) = \begin{cases} \text{probability of a type I error for each fixed } \theta \in \Theta_0, \\ 1 - \text{probability of a type II error for each fixed } \theta \in \Theta_1. \end{cases} \tag{2.3}$$

Due to the previous observation we make the following definition: the *power function* of a parametric hypothesis test with rejection region R is a function of θ and is defined to be (Casella & Berger 2002, p. 383)

$$\gamma(\theta) = \Pr_\theta(X \in R). \tag{2.4}$$

Since we do not know the value of θ before we do the hypothesis test, the power function should be small, ideally 0, for most $\theta \in \Theta_0$ and large, ideally 1, for most $\theta \in \Theta_1$ (Casella & Berger 2002, p. 383). We discuss the ideal situation $\gamma(\theta) \approx 1$ for most $\theta \in \Theta_1$ in Section 2.7.

In Section 2.4 we stressed the difference between the research hypothesis and the statistical hypotheses. We said that the statistical model could be wrong and

that the results from the statistical hypothesis test may tell nothing about the research hypothesis. How then can then the statistical hypothesis test make a correct or wrong decision in such cases? Should not the statistical test always be wrong? When we use a statistical hypothesis test we assume that the underlying statistical model is true. When H_0 is correct the outcome of the experiment might be a "very unlikely" realisation(s) from the probability distribution so that the test wrongly rejects the null hypothesis. We always regard the outcome in the experiment as a realisation from the probability distribution we have chosen to model the real quantities with.

In the Neyman–Pearson approach to hypothesis testing we demand that the upper bound on the maximum of the type I error probabilities of the test is below some limit. This upper bound is called the *significance level* (Casella & Berger 2002, p. 385) of the test. We say that a test with power function $\gamma(\theta)$ is a (significance) level α test if

$$\sup_{\theta \in \Theta_0} \gamma(\theta) \leq \alpha \tag{2.5}$$

Common significance levels are 0.05 or 0.01. From the definition of level α test, a level 0.01 test is also a level 0.05 test. However, it is not common to say that a level 0.01 test is a level 0.05 test. One of the reasons we define the level α test as we have done becomes evident when we work with discrete distributions. Say that you want a level 0.05 test. Due to the discreteness of the probability distribution it may be that for a specific rejection region R you get $\sup_{\theta \in \Theta_0} \gamma(\theta) = 0.04$, but when you include one or more outcomes in R you get $\sup_{\theta \in \Theta_0} \gamma(\theta) = 0.06$. We then keep the first rejection region (assuming that is wisely constructed) since we want a level 0.05 test. We also refer to $\sup_{\theta \in \Theta_0} \gamma(\theta)$ as the *size* of the test (Hogg et al. 2014, p. 440).

The manufacturer in example (b) wants to do a traditional statistical hypothesis test and sets the statistical significance level to 0.05. We know that large values of X should be in R. We also know that we need to pick the $\theta \in \Theta_0$ that maximises $\Pr_\theta(X \in R) = \sum_{x \in R} \Pr_\theta(X = x)$ when calculating the size of the test. These two tasks, pick a rejection region and calculate $\sup_{\theta \in \Theta_0} \Pr_\theta(X \in R) = \sup_{\theta \in \Theta_0} \sum_{x \in R} \Pr_\theta(X = x)$, might seem difficult to do at the same time. However, the value of θ that will give the most probability mass on the outcomes that potentially will be in the rejection region is $\theta = \frac{1}{2}$. We have tried to illustrate this in Figure 2.1. Here we have plotted the binomial distribution with 16 trials for two different values of the success probability θ, $\theta = 0.4$ and $\theta = 0.5$. When you decrease the value of θ from 0.5 less and less probability mass is placed on the points that will be in the rejection region. We therefore only need to consider $\gamma(\theta) = \sum_{x \in R} \Pr_\theta(X = x)$ for $\theta = \frac{1}{2}$.

We have that

$$\Pr_{0.5}(X = 11) + \Pr_{0.5}(X = 12) + \Pr_{0.5}(X = 13) + \Pr_{0.5}(X = 14) + \Pr_{0.5}(X = 15) + \Pr_{0.5}(X = 16) \leq 0.03841,$$

and also that

$$\Pr_{0.5}(X = 10) + \Pr_{0.5}(X = 11) + \Pr_{0.5}(X = 12) + \Pr_{0.5}(X = 13) + \Pr_{0.5}(X = 14) + \Pr_{0.5}(X = 15) + \Pr_{0.5}(X = 16) \leq 0.1051,$$

which means we should let the rejection region be $R = \{11, 12, 13, 14, 15, 16\}$ to get a (significance) level 0.05 test. The reason we select the significance level to be as low as 0.05 is that when the test rejects the null hypothesis, we can be quite certain that H_1 is true (Borgan 2007) (if the statistical model is an adequate model of the experiment). This also depends on the construction of the rejection region and only holds if the rejection region consists of outcomes indicating that H_1 is correct. We could of course select outcomes with low value of X to be in the rejection region, i.e possibly 0, 1 etc. However, since $\Pr_0(X = 0) = 1$ and $\Pr_{0.065}(X = 1) = 0.74$, we would need to set $R = \varnothing$. Then the test would always accept the null hypothesis and would never make a type I error. Even in cases where $0 < \sup_{\theta \in \Theta_0} \gamma(\theta) \leq \alpha$ this idea would not be a good one. We do control the type I error probabilities, but the power will be much lower than if the rejection rejection instead consists of points indicating that H_1 is true.

As a note, if we choose to reject the null hypothesis we do not know the probability of committing an error, since we do not know the true value of θ. Only under H_0 we now the maximum type I error probability.

2.7 Perform the experiment and reject or accept the null hypothesis

After choosing a test to be used in the parametric statistical hypothesis testing procedure, which means selecting a test statistic and choosing an appropriate rejection region so that the level of the test is below the significance level, and after carrying out the experiment the test will tell you whether to accept or reject the null hypothesis. In example (b), the outcome of the experiment is 10, which means that the test at significance level 0.05 does not reject the null hypothesis. However, if the new ointment really is better than the old one, how likely is it that the manufacturer will discover it using the given test? To answer this question we must look at $\gamma(\theta)$ for $\theta \in \Theta_0^c$. The power function is plotted in Figure 2.2

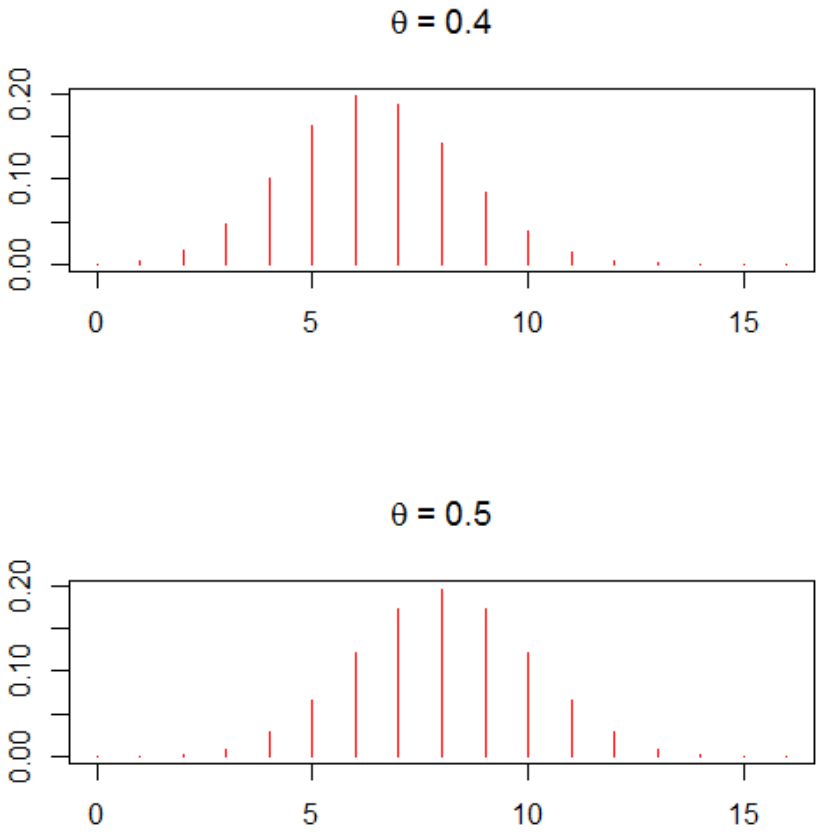


Figure 2.1: Illustrations of the binomial probability mass function for two different values of the success probability θ, but with the same number of trials, 16. The value of θ used is given in the title of each subplot.

for $\theta \in \Theta_0^c$. We see that the power function is a non-decreasing function for $\theta \in \Theta_0$. This makes intuitively sense: it should be easier to detect a larger success probability compared to a smaller one (both above $\frac{1}{2}$). We also see that the power is above 80% when $p \geq 0.8$ and it is about 1 when $p \geq 0.9$. In Section 2.6 we said that a theoretically desired property of the power function is that $\gamma(\theta) \approx 1$ for most $\theta \in \Theta_1$. Should then the manufacturer be dissatisfied with the test? To answer this question, we first look at three ways the manufacturer can increase the power. The first way is to use a different test statistic resulting in a test with better power properties. The second way is be to increase the significance level α of the original test, since then the rejection region would be become bigger so that we sum over more outcomes when we calculate the power. The third way is to increase the number of trials in the experiment, i.e increase the sample size. With the third approach the power still increases even if the significance level remains the same. This also makes intuitively sense, since once you have more data it should be easier to detect also "smaller" success probabilities (in Θ_0^c) compared to when you have less data.

In Figure 2.3 we have plotted the power function $\gamma(\theta)$ for $\theta \in \Theta_0^c$ of the same test (same test statistic $T = T(X)$, the same significance level α, construction region constructed similarly, same H_0 and H_1) but with $n = 1000$. We see that $\gamma(\theta) \approx 1$ for $\theta \geq 0.6$ and that $\gamma(\theta) \geq 0.8$ for $\theta \geq 0.58$. From a pure mathematical view, the power function depicted in Figure 2.3 is better than the one shown in Figure 2.2 since the power is greater (or to be more exact; at least as great) at each $\theta \in \Theta_0^c$. For instance, the power at $\theta = 0.6$ is 1 when $n = 1000$ but below 0.20 when $n = 16$. It is therefore unlikely that the test with sample size 16 will detect that the success probability is 0.6, but when $n = 1000$ this will be detected. However, the manufacturer might prefer the sample size 16 over 1000. For starters it is much cheaper to test less people, but more importantly it might be that the manufacturer only wants the test to be able detect a success probability that is higher than 0.8.

It is common practise to the set sample size large enough so that the power of the test is at least 80 % at θ that are *scientifically meaningful* (Ambrosius 2007, p. 70-72) (Sakpal 2004), i.e for the $\theta \in \Theta_0^c$ that the investigators want their test to be able to detect. So in the ointment example when $\frac{1}{2} < \theta < 0.8$ we could get a statistically significant outcome when $n = 1000$, but if the manufacturer only considers $0.8 \leq \theta \leq 1$ to be of practical importance, the result of the test would not be considered practically significant. We do not know the true value of θ before we do the statistical hypothesis test, otherwise it would not be necessary to do it. We calculate power before we do the experiment to ensure that the power is high enough (as mentioned, at least 80 %) for scientifically meaningful values of $\theta \in \Theta_0^c$

and lower for the values that are not practically meaningful.

2.8 Simple and composite null hypotheses

In the null hypothesis in Equation (2.2) infinitely many values of the success probability θ are possible and we call it a *composite null hypothesis* (Stuart & Ord 1991, p. 795). If only one value is possible in the null hypothesis, such as $H_0 : \theta = \frac{1}{2}, H_1 : \theta \neq \frac{1}{2}$, we call the null hypothesis for *simple* (Stuart & Ord 1991, p. 795). Note that we do not make the same definitions for the alternative hypothesis. The reason is that when we construct rejection regions we only consider the distribution of the test statistic under the null hypothesis and the construction is simpler when the null hypothesis is simple. Also later when we consider the *p-value*, we will see that we calculate the p-value assuming the null hypothesis is true and that it is easier to calculate the p-value when the null hypothesis is simple compared to when it is composite.

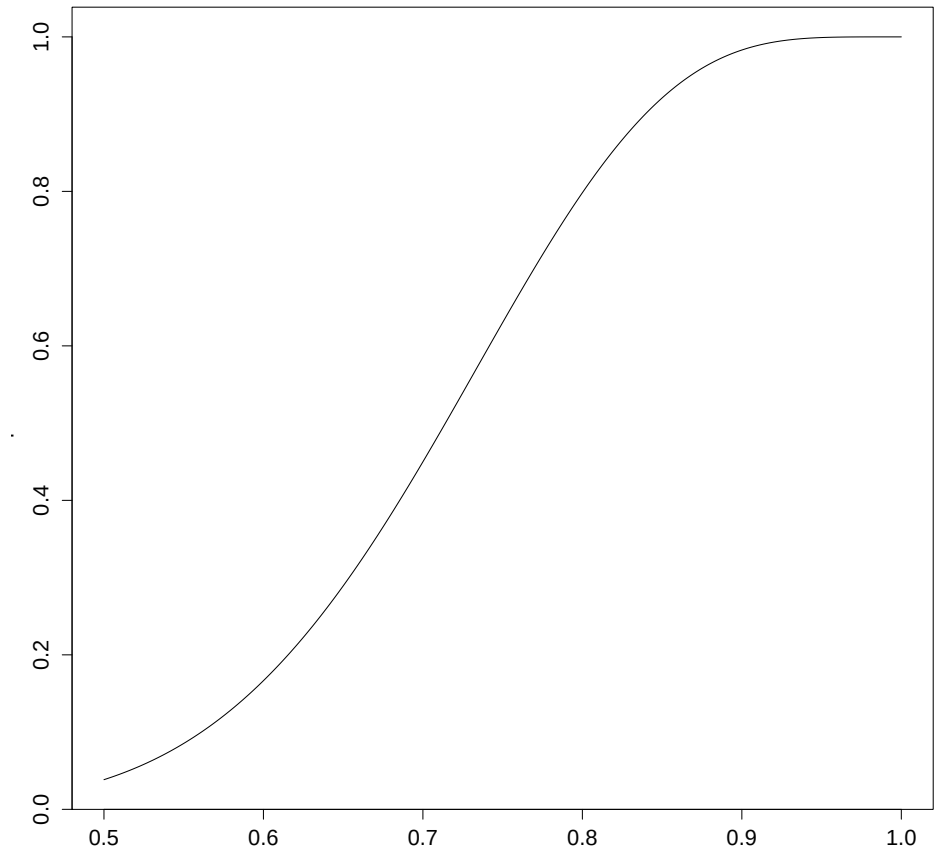


Figure 2.2: Power function for the hypothesis test with test statistic $T(X) = X$, $X \sim$ Binom(16, θ), and rejection region $R = \{x \mid 11 \leq x \leq 16\}$. The function is plotted for $\theta \in \Theta_0^c$.

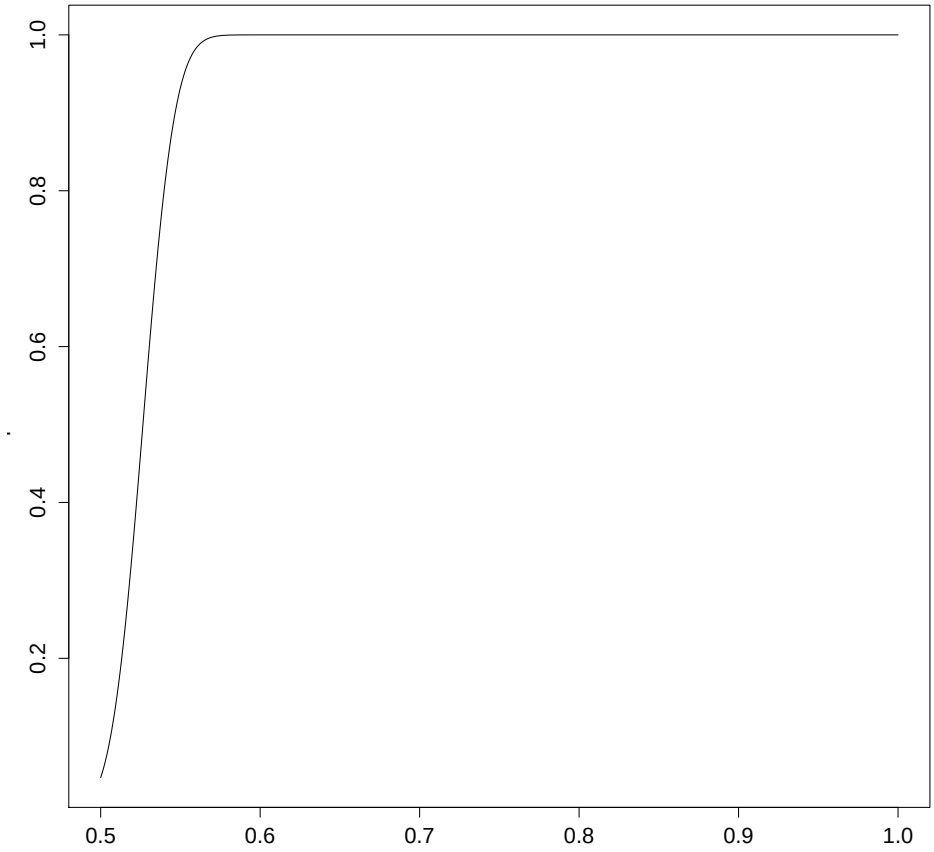


Figure 2.3: Power function for the hypothesis test with test statistic $T(X) = X$, $X \sim \text{Binom}(1000, \theta)$, and rejection region $R = \{x \mid 526 \leq x \leq 1000\}$. The function is plotted for $\theta \in \Theta_0^c$.

Chapter 3

The p-value approach to statistical hypothesis testing when the null hypothesis is simple

The p-value approach to statistical hypothesis testing is an alternative to the Neyman-Pearson approach. We start this section with presenting this method. We then show how enumeration can be used to calculate the realisations of a p-value and illustrate that the p-value is a random variable. We also introduce the concept of validity.

3.1 Overview

An alternative approach to the traditional Neyman–Pearson approach to hypothesis testing is the p-value approach. The main steps are the same as for traditional hypothesis testing. However, we do not explicitly construct a rejection region and do not reject or accept the null hypothesis depending on the value of the test statistic. Instead we calculate the *p-value*. In this chapter we consider simple null hypotheses when we define the p-value, i.e. we study null hypotheses of the form

$$H_0 : \theta = \theta_0. \tag{3.1}$$

The p-value can then be calculated as the probability (assuming the value of θ in H_0 to be the true value) of obtaining a value of the test statistic as least as

contradictory to H_0 as the value of the test statistic we get when we evaluate it in the outcome of the experiment (Devore et al. 2012, p. 456). If the p-value is below or equal to the significance level the test accepts the null hypothesis otherwise the test rejects it. If we assume that large values of the test statistic $T(X)$ indicate that H_1 is correct and the larger the value the stronger the indication, then the p-value can be calculated as

$$p(x) = \Pr_{\theta_0}(T(X) \geq T(x)). \tag{3.2}$$

Note that we have not explicitly specified an alternative hypothesis in Equation (3.1). The reason is that the definition of p-value works for all types of alternative hypothesis and that the calculation of the probability in Equation (3.2) does not explicitly depend on the values of θ under H_1. The choice of test statistic may depend on the form of the hypothesis (this will hopefully become clear in Section 4.3), for instance we may prefer to use a different test statistic when H_1 is *two-tailed*, $H_1 : \theta \neq \theta_0$, compared to when it is *one-tailed*, $H_1 : \theta < \theta_0$ or $H_1 : \theta > \theta_0$, but when the choice of test statistic has been made the values of θ in H_1 do not matter when we calculate the p-value

3.2 Use of enumeration to calculate p-values

When we consider discrete distributions where the sample space is finite we can calculate the realisation of the p-value in Equation (3.2) by use of the exact distribution of the test statistic, and not some large sample (asymptotic) distribution of it, by using a method called *enumeration*, see for instance Verbeek & Kroonenberg (1985). The steps in this method for calculating the p-value for a given outcome x in the experiment are

1. List all the outcomes in the experiment (or list all elements in the sample space of the random variable X).
2. Evaluate the test statistic in each outcome (and store these values).
3. For each outcome with test statistic greater than or equal to $T(x)$, calculate the probability of the outcome.
4. The realisation of the p-value is the sum of the probabilities calculated in the previous step.

Mathematically we write the enumeration procedure as

$$p(x) = \Pr_{\theta_0}(T(X) \geq T(x)) = \sum_{T(x') \geq T(x)} \Pr_{\theta_0}(X = x'). \tag{3.3}$$

Table 3.1: Illustration of the enumeration method given in Section 3.2 used on example (b) in Chapter 2 with simple null hypothesis $H_0 : \theta = \frac{1}{2}$ instead of composite. The result of step 1 is the leftmost table, the result of step 2 is the table in the middle and the result of step 3 is the rightmost table. In step 4 the result is $p(x) = p(10) = 0.227$

x'
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

x'	$T(x')$
0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10
11	11
12	12
13	13
14	14
15	15
16	16

x'	$T(x')$	$\Pr_{\theta_0}(X = x')$
0	0	
1	1	
2	2	
3	3	
4	4	
5	5	
6	6	
7	7	
8	8	
9	9	
10	10	0.12219
11	11	0.06665
12	12	0.02777
13	13	0.00854
14	14	0.00183
15	15	0.00024
16	16	0.00002
		Sum 0.227

We illustrate this method when testing an adapted version of the hypotheses in Equation (2.2) where we make H_0 simple, i.e

$$H_0 : \theta = \frac{1}{2}, H_1 : \theta > \frac{1}{2}. \tag{3.4}$$

The realisation in the experiment is still $x = 10$. The different steps are illustrated in Table 3.1. If the manufacturer were to do the same calculations as we have done it would get that the p-value is $p(x) = p(10) = 0.227$ so that the test would not reject the null hypothesis at the 5% significance level.

As mentioned, the enumeration procedure works for all discrete distributions where the sample space is finite. So the method can be used for instance either with the multinomial or hypergeometric distribution. This also means that the procedure does not work when the distribution is continuous such as when the outcome follows the normal or chi square distribution.

Table 3.2: The p-values for the different outcomes in the adapted version of the ointment example where we test $H_0 : \theta = \frac{1}{2}, H_1 : \theta \geq \frac{1}{2}$.

x'	$\Pr_{\theta_0}(X = x')$
0	0.99998
1	0.99974
2	0.99791
3	0.98936
4	0.96159
5	0.89494
6	0.77275
7	0.59819
8	0.40181
9	0.22725
10	0.10506
11	0.03841
12	0.01064
13	0.00209
14	0.00026
15	0.00002
16	0

3.3 The p-value as a random variable

In Table 3.2 we have calculated all the possible p-values in the adapted version of the ointment experiment (where we consider the hypotheses in Equation (3.4)) using Equation (3.3)). Although we calculate the p-value as a probability, we see from Table 3.2 that the p-value acts like a random variable. The reason we observe this behaviour is that the "summation limit" $T(x)$ in Equation (3.3) is a random variable. We calculate the p-value using Equation (3.3) once we have observed the realisation of $T(x)$ in the experiment. This means that we calculate different p-values if we in one run of the experiment observed $T(x) = 10$ compared to if we observed $T(x) = 13$ in another run. Therefore the p-value should be considered a random variable and not a fixed probability.

Since we calculate the p-value using Equation (3.3) we can view the p-value as a special transformation of the test statistic $T(X)$. Because the test statistic is a function of the data, the p-value is also a function of the data. So it is natural to consider the p-value as a test statistic, i.e $p = p(X)$. Since we calculate the realisations of $p(X)$ as probabilities, $p(X)$ takes values between 0 and 1. Is is therefore more correct to refer to the calculated values $p(1)$ or $p(10)$ as realisations

of the p-value $p(X)$ and not as the p-value itself. This is the same as differentiating between the observed values of a random variable and the random variable itself, which is common practise in statistics (see for instance Verzani (2013, p. 195) for a brief note or Georgi (2014, p. 211) for a longer note). This practise does not seem to be common for the p-value. For instance in Walpole et al. (2012, p. 334) they say that "The P-value can be viewed as simply the probability of obtaining these data given that both samples come from the same distribution.". If we talk of the p-value as a probability, then it should not change from experiment to experiment. In this book they also say that "Compute the P-value based on the computed value of the test statistic" (Walpole et al. 2012, p. 233). So we compute *the* p-value but calculate the *value* of the test statistic. From now on we refer to the different calculated values of $p(X)$ as realisations or observations of $p(X)$ to hopefully make it clearer that $p(X)$ is a random variable and not a probability. One could argue that writing $p = p(X)$ should be enough (since if X is a random variable and p is a function of the random variable, p should be a random variable as well), but we want to make the distinction between the two views as clear as possible.

At the beginning of Section 3.1 we said that after having calculated the p-value in an experiment, the test rejects H_0 if $p(x) \leq \alpha$. We want this procedure to provide a valid level α-test. We know from Equation (2.5) that a (significance) level α-test satisfies

$$\Pr_\theta(\text{reject } H_0) \leq \alpha$$

for all $\theta \in \Theta_0$. Since $\Pr_\theta(\text{reject } H_0) = \Pr_\theta(p(X) \leq \alpha)$ in the p-value procedure, we want that

$$\Pr_\theta(p(X) \leq \alpha) \leq \alpha, \tag{3.5}$$

to hold for all $\theta \in \Theta_0$ and all significance levels $\alpha \in [0, 1]$. A p-value $p(X)$ that satisfies Equation (3.5) is said to be a *valid* p-value (Casella & Berger 2002, p. 397). In Section 3.5 we show that the p-value defined in Equation (3.2) is valid.

We evaluate tests based on p-values in the same fashion as we evaluate other hypothesis tests, i.e by considering the power function of the test. For a test based on the p-value $p(X)$ the power function takes the form

$$\gamma(\theta) = \Pr_\theta(\text{reject } H_0) = \Pr_\theta(p(X) \leq \alpha). \tag{3.6}$$

We noted at the start of Section 3.1 that we do not construct a rejection region when we use the p-value approach to statistical hypothesis testing. Furthermore, we are not interested in knowing the exact distribution of the p-value for different values of θ. So how do we evaluate the power function in Equation (3.6)? The answer is that we use a process similar to that of enumeration:

1. Calculate all the realisations of the p-value for each outcome using the enumeration procedure outlined in Section 3.2
2. For each realisation of the p-value equal to or below α calculate the probability of the corresponding outcome where you use the value of θ at which you want to evaluate the power function.
3. The value of the power function at θ is the sum of the probabilities in the previous step.

Mathematically we write the previous procedure as

$$\gamma(\theta) = \Pr_\theta(p(X) \leq \alpha) = \sum_{p(x)\leq\alpha} \Pr_\theta(X = x).$$

If we use the outlined procedure to calculate the power in the test of Equation (3.4), we see that we sum over the same outcomes as when we calculated the power in the original example, see Table 3.2, since we consider the outcomes $x = 11$ to $x = 16$ in both cases. This means the power functions are the same for $\theta \in \Theta_0^c$ and therefore Figure 2.2 also illustrates the power of both tests.

3.4 Interpretation of the p-value when the null hypothesis is simple

Since we work in a frequentist setting we can interpret each realisation of the p-value defined in Equation (3.2) as the long run proportion of experiments where we get a test statistic as least as extreme as the original value of the test statistic when H_0 is true. We explain this interpretation in the setting of the ointment experiment. Imagine that the manufacturer of the ointment is to do n independent runs of the experiment. The probability $\Pr_\theta(T(X) \geq T(10))$ is then the long-run proportion of the independent trials where a test statistic as least as large as $T(10)$ is observed (i.e $x \geq 10$). Since $p(10)$ equals $\Pr_\theta(T(X) \geq T(10))$, $p(10)$ must have the same interpretation as $\Pr_\theta(T(X) \geq T(10))$.

We now make some comments about p-values that we hope clarify how they should be interpreted when the null hypothesis is simple. The first comment is that the p-value indicates how much evidence the data or outcome of the experiment gives against the null hypothesis (Wasserstein & Lazar 2016). The smaller the realisation of the p-value, the more evidence there is that the null hypothesis is false. So a p-value of 0.0002 is stronger evidence against the null hypothesis than a p-value of 0.01. The reason is that a lower p-value means the probability of observing

a outcome as least as contradictory to H_0 in a future run of the experiment as the value we already have observed is lower compared to when the p-value is higher (where we calculate the probabilities assuming the null hypothesis is true). However, we cannot compare a p-value in an experiment with a p-value in another experiment if the number of trials are different even if we compute both the p-values using Equation (3.2) and use the same test statistic. The reason is that the sample spaces are different so that the power functions may be different. Furthermore, if we use different test statistics in the same experiment we cannot compare the p-values since they order the sample space differently.

In Section 2.7 we said that a statistically significant result is not the same as practically significant result. The same holds for the p-value procedure for hypothesis testing, so even if a p-value indicates how much evidence there is against the null hypothesis, it does does not measure the the importance or practically significance of a result (Wasserstein & Lazar 2016). We can demonstrate this fact in the ointment example. Let us say that the manufacturer, hypothetically, performs the experiment on 1,000,000 persons and the true value of θ is 0.51 (so that the binomial model is indeed an appropriate statistical model and the true value of the success probability is 0.51). After specifying a significance level we can calculate the test power. Even if the manufacturer specifies the significance level as low as for instance $5 \cdot 10^{-10}$ the test power at 0.51 is 1. This means we observe would observe a very small realisation of the p-value in the experiment (given that $\theta = 0.51$) (Lin et al. 2013). However, the success probability is most likely of no practical importance.

Our last comment is about another quantity the realisations of p-values do not measure. They do not tell you anything about the probability of the studied null or alternative hypothesis being true (Wasserstein & Lazar 2016). Either $\theta \in \Theta_0$ or $\theta \in \Theta_0^c$ and θ is not a random variable (since we work in a frequentist setting). This means statements such as $\Pr(H_0) = 0.05$, $\Pr(\theta \in \Theta_0) = 0.05$ or $\Pr(\theta \in \Theta_0^c) = 0.95$ based on the realisation of the p-value being 0.05 are wrong. One could say that $\Pr(\theta \in \Theta_0)$ or $\Pr(H_0)$ is 1 or 0 depending on whether the null hypothesis is true or false (the same holds for $\Pr(\theta \in \Theta_0^c)$ or $\Pr(H_1)$), but the realisation of the p-value does not tell you anything about these probabilities.

3.5 Proof of Equation (3.2) defining a valid p-value

We want to show that the p-value in Equation (3.2) is a valid p-value, i.e that

$$\Pr_{\theta_0}(p(X) \leq \alpha) \leq \alpha, \tag{3.7}$$

holds for all $\alpha \in [0, 1]$. The proof is valid for all simple null hypotheses where we use a test statistics for which large values indicate that the alternative hypothesis H_1 is true and the larger the value the stronger the indication.

We start by selecting a significance level $0 \leq \alpha \leq 1$ and set

$$t_L = \min\{t \mid \Pr_{\theta_0}(T(X) \geq t) \leq \alpha\}. \tag{3.8}$$

There are two situations that can occur for a given α

1. The set on the right hand side of Equation (3.8) is non-empty
2. The set on the right hand side of Equation (3.8) is empty

Situation (2) can occur since $\Pr_{\theta_0}(T(X) = t) > \alpha$ is possible for the largest possible value of t when we consider discrete distributions.

We first consider situation (1). In Equation (3.8) we select the t so that $\Pr_{\theta_0}(T(X) \geq t)$ is closest to α among the t so that $\Pr_{\theta_0}(T(X) \geq t) \leq \alpha$. We then have that the test with rejection region $R = \{x \mid T(X) \geq t_L\}$ is a level α-test and have as large rejection region as possible (that is also wisely constructed). The realisations of the p-value for the outcomes in the rejection region are lower than or equal to α. For the outcomes in the acceptance region, the realisations of the p-value are higher than α. The probability that the p-value is below or equal to α is therefore given by

$$\Pr_{\theta_0}(p(X) \leq \alpha) = \Pr_{\theta_0}(T(X) \geq t_L) = \Pr_{\theta_0}(\text{reject } H_0) = \Pr_{\theta_0}(X \in R) \leq \alpha,$$

since we have a level α-test.

In situation (2) it could for instance be that there exists t_l and t_u so that

$$\sum_{T(t_l) \leq T(x) \leq T(t_u)} \Pr_{\theta_0}(X = x) \leq \alpha,$$

However, we construct the rejection region by starting with the empty set and then "adding" the outcomes by descending order of the test statistic. When we add elements we begin with the most extreme value of the test statistic and then

add outcomes until we have the test with size closet to (i.e smaller than) or equal to α. In situation (2) the rejection region is empty since $\Pr_{\theta_0}(T(X) = t) > \alpha$ for the largest possible value of t. This means that all p-values are greater than α so that

$$\Pr_{\theta_0}(p(X) \leq \alpha) = 0 \leq \alpha.$$

The two situations we have considered for a given α will hold for all $\alpha \in [0, 1]$. This means we have shown Equation (3.7) to hold for all $\alpha \in [0, 1]$, which means the p-value is valid.

Chapter 4

Theory: Testing equality of binomial proportions

In this chapter we construct tests of equality of success probabilities in two independent binomial distributions.

4.1 Introduction

This section is organized as follows: we start by introducing the hypotheses studied when testing equality of success probabilities in two independent binomial distributions, then we give examples where it is possible to study these hypotheses and finally we introduce methods for creating p-values.

4.1.1 Null and alternative hypothesis and examples

When testing the equality of the success probabilities in two independent binomial distributions the null and alternative hypothesis can be formulated as

$$H_0 : \theta_1 = \theta_2, H_1 : \theta_1 \neq \theta_2, \tag{4.1}$$

where $X_1 \sim \text{Binom}(\theta_1, n_1)$, $X_2 \sim \text{Binom}(\theta_1, n_2)$ and X_1 and X_2 are independent. This means $\Theta_0 = \{(\theta_1, \theta_2) \mid \theta_1 = \theta_2, 0 \leq \theta_1, \theta_2 \leq 1\}$ and $\Theta_0^c = \{(\theta_1, \theta_2) \mid \theta_1 \neq \theta_2, 0 \leq \theta_1, \theta_2 \leq 1\}$.

We now consider two examples where it is natural to compare two independent binomial proportions. In the first example, example (c), researchers have developed

Table 4.1: Reccurence of tumors in the patiens recieving tratment and the patients reciveing placebo in example (c). The coloumn titles "Yes" and "No" referes to recurrence of tumor.

	Yes	No	Total
Treatment	17	121	138
Placebo	53	76	129

a new treatment to prevent recurrence of benign tumours in the big intestine and the rectum (Meyskens et al. 2008). The researchers assign the treatment randomly to 138 of the 267 study subjects and give placebo to the rest. This is thus a randomized experiment. The results are summarised in the 2×2 table in Table 4.1. The number of study subjects receiving the treatment and the number of subjects receiving the placebo are fixed by design. If the researchers were to do another run of the experiment, the only quantities that can vary are the numbers of subjects with recurrent tumours and non-recurrent tumours. It is therefore natural to consider these numbers as realisations of random variables. One possibility is to model each of the subjects receiving treatment as independent realisations of Bernoulli random variables with same success probability θ_1, where $I_{i1} = 1$ if a tumour recurs in subject i receiving treatment and 0 otherwise and let $X_1 = I_{11} + \cdots + I_{n_1 1}$ where $n_1 = 138$. Similarly they can model each of the subjects receiving placebo as independent realisations of Bernoulli random variables, with the different success probability θ_2 and let $X_2 = I_{12} + I_{22} + I_{32} + \cdots + I_{n_2 2}$, where $I_{i2} = 1$ if the tumour recurs in subject i receiving placebo and 0 if no tumour recurs and $n_2 = 129$. Then $X_1 \sim \text{Binom}(\theta_1, n_1)$ and $X_2 \sim \text{Binom}(\theta_2, n_2)$ independently. Since they want to know if the treatment has an effect on the recurrence of tumours in the rectum and big intestine, they should consider the same null hypothesis as in Equation (4.1). They could also study the same alternative hypothesis.[1]

In the other example, example (d), researches want to decide if certain genes can cause severe near-sightedness (myopia) (Zhao et al. 2011). If you are unfamiliar with genetics and want to understand this example, a short introduction to the relevant concepts needed are given in Appendix A. The researches do a case control study. This is an observational study. Here we sample individuals with and without severe near-sightedness. The individuals with near-sightedness are called cases and

[1]However, it does not seem plausible that the treatment should be worse than placebo, so they should not consider a two-sided alternative hypothesis but instead consider

$$H_1 : \theta_1 > \theta_2.$$

This means the researchers should consider the same null hypothesis as in Equation (4.1) but a different alternative hypothesis.

those subjects without are called controls (Silva 1999). So when we sample cases or controls we know beforehand that they have severe near-sightedness or not. Afterwards we measure if cases and controls have unequal level of exposure to a factor. In our case this means we want to find out if at some loci, the genotype frequencies depend on the case/control status. If so then different genotypes are likely to give contributions to the phenotype depending on the case/control status and some of the genotypes could cause severe near-sightedness. Near sightedness is a quantitative trait, which means that this trait is caused by both several genes and environmental conditions that act together. This means for instance a single gene cannot cause severe near sightedness by itself. The researchers have genotyped several SNP-loci, but we only consider one of these. Furthermore, the researches have studied 103 cases and 97 controls. The results are shown in Table 4.2.

If we were to genotype the same number of cases and controls as done in example (d) but use different individuals, then we would most likely not obtain the same values as in Table 4.2. It is therefore natural to consider the numbers as realisations of random variables. One possibility is to model the data in each row as realisations from two multinomial distributions (Balding et al. 2003, p. 944). Normally we assume the SNP is in gametic disequilibrium with a causal locus and hope the SNP is linked to the causal locus. However, if we assume that the SNP is the causal loci and assume that the A-allele is dominant, then the individuals with the AA genotype and the AB genotype get the same contribution to the trait from this locus. This is called a dominant model (Ziegler & König 2010, p. 269) and the model is studied in Zhao et al. (2011). Since individuals with genotypes AA and AB get the same contribution to the trait from this locus, we should group the individuals with these two genotypes together. By grouping the mentioned genotypes together we get Table 4.3. One possibility is then to model each row in Table 4.3 with the binomial model. If we let X_1 denote the number of controls with AA or AB genotypes, then we assume $X_1 \sim \text{Binom}(\theta_1, 97)$. And if we let X_2 denote the number of cases with the AA or AB genotypes we assume that $X_2 \sim \text{Binom}(\theta_2, 103)$. We want to study whether or not it is more likely that one of the two genotype groups makes contributions depending on case/control status, i.e we want to find out if $\theta_1 = \theta_2$ or $\theta_1 \neq \theta_2$. If $\theta_1 \neq \theta_2$ this would indicate the SNP-locus is associated with severe near-sightedness. We therefore want to study the hypotheses in Equation (4.1).

Table 4.2: The different numbers of genotypes at the studied SNP-locus in the myopia study. We denote the alleles at the locus A and B.

	AA	AB	BB	Total
Controls	17	51	29	97
Cases	39	44	20	103

Table 4.3: A dominant model for the A allele is assumed at the studied locus in the severe near-sightedness study, which means we put the individuals with AB or AA genotypes in the same group.

	AA+AB	BB
Controls	68	29
Cases	83	20

4.1.2 Methods for calculating p-values when the null hypothesis is composite

Under the null hypothesis in Equation (4.1) $\theta = \theta_1 = \theta_2$ can take infinitely many values, since all points on the line from $(0, 0)$ to $(1, 1)$ are possible values of (θ_1, θ_2). In Figure 4.1 we have illustrated the parameter space. Since the null hypothesis is composite we cannot use Equation (3.3) when calculating the realisation of the p-value. We initially consider four different methods for generating a p-value. They are

1. the *maximisation (M) p-value*, where we calculate the realisations as

$$p_M(\boldsymbol{x}) = \sup_{\theta \in \Theta_0} \Pr_{\boldsymbol{\theta}}(T(\boldsymbol{X}) \geq T(\boldsymbol{x})) \tag{4.2}$$

   Casella & Berger (2002, p. 397). Lloyd (2008) refers to this as the worst case p-value since we calculate the realisation of the p-value in each simple null hypothesis and pick the one that gives the largest value.

2. the *estimation (E) p-value*, where we replace θ_0 in Equation (3.7) with the maximum likelihood estimate of $\boldsymbol{\theta} = (\theta_1, \theta_2)$ under H_0 using the observed outcome $\boldsymbol{x}$ in the experiment. Bayarri & Berger (2000) refer to this as the plug-in p-value and Lloyd (2008) refer to this as the estimated p-value. The realisations of the E p-value are therefore given by

$$p_E(\boldsymbol{x}) = \Pr_{\hat{\boldsymbol{\theta}}}(T(\boldsymbol{X}) \geq T(\boldsymbol{x})), \tag{4.3}$$

   where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$ under H_0 based on $\boldsymbol{x}$.

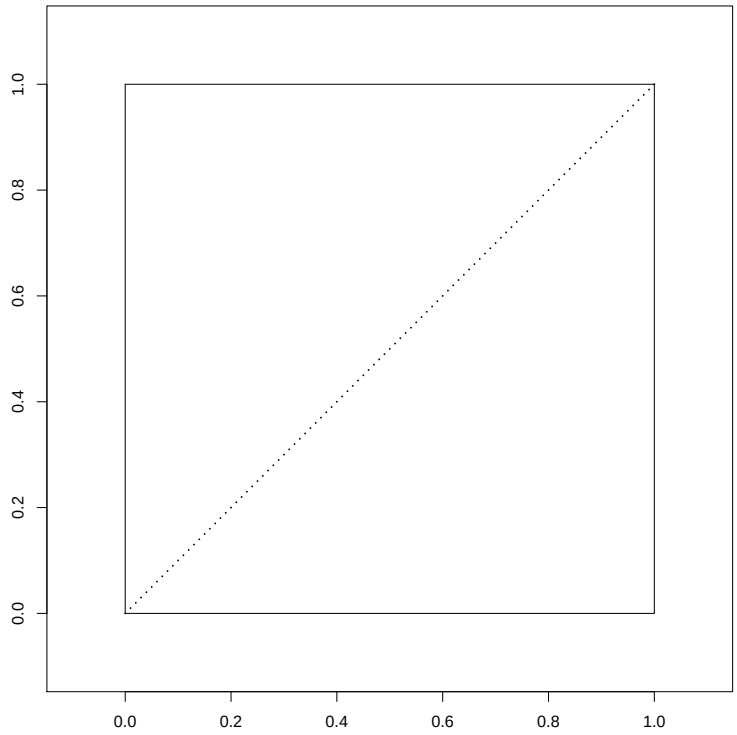


Figure 4.1: Illustration of the parameter space $\Theta = \{(\theta_1, \theta_2) \mid 0 \leq \theta_1 \leq 1, 0 \leq \theta_2 \leq 1)\}$ when $\boldsymbol{X} = (X_1, X_2)$, $X_1 \sim \text{Binom}(\theta_1, n_1)$ and $X_2 \sim \text{Binom}(\theta_2, n_2)$. The parameter space consists of the points on or inside the square with corners $(0,0), (1,0), (1,1), (0,1)$. Under H_0 in Equation (4.1) only the points on the dotted line are possible.

3. the *conditional (C) p-value* where we condition the test statistic on a sufficient statistic $S(\boldsymbol{X})$ for $\boldsymbol{\theta}$ under H_0 (Casella & Berger 2002, p. 399). So the realisations of the C p-value are given by

$$p_C(\boldsymbol{x}) = \Pr(T(\boldsymbol{X}) \geq T(\boldsymbol{x}) \mid S(\boldsymbol{X}) = S(\boldsymbol{x})), \tag{4.4}$$

where the probability does not depend on $\boldsymbol{\theta}$ since $S(\boldsymbol{X})$ is sufficient under H_0.

4. the *asymptotic (A) p-value* where we use a large sample distribution of $T(\boldsymbol{X})$ under H_0, which is free of $\boldsymbol{\theta}$ (see for instance Walpole et al. (2012, p. 363–364) or Høyland (1986, p. 80–81). We therefore calculate the realisations of this p-value as

$$p_A(\boldsymbol{x}) = \Pr(Y \geq T(\boldsymbol{x})), \tag{4.5}$$

where $T(\boldsymbol{X}) \xrightarrow{d} Y$.

In Section 4.6 we show that $p_A(X)$ and $p_E(X)$ are asymptotically valid and that $p_C(X)$ and $p_M(X)$ are valid.

Before we illustrate how we calculate the realisations of the different p-values we need a test statistic. We consider different candidates for the test statistic in Section 4.3. We want to apply the different methods on only one test statistic. In the next section we review some of the theory necessary to calculate the realisations of the p-values and also the theory needed to find the large sample distribution of a test statistic.

4.2 Review of theory

In this section we give a review of the theory needed in the different methods for creating a p-value and the theory needed to be able to find a large sample distribution of a test statistic.

4.2.1 Sufficiency

A statistic $S(\boldsymbol{X})$ is *sufficient* for a parameter θ if

$$\Pr_\theta(\boldsymbol{X} = \boldsymbol{x} \mid S(\boldsymbol{X}) = s) = \Pr(\boldsymbol{X} = \boldsymbol{x} \mid S(\boldsymbol{X}) = s), \tag{4.6}$$

holds for all values s such that $\Pr_\theta(S(\boldsymbol{X}) = s) > 0$ (Casella & Berger 2002, p. 272–273) i.e the conditional distribution of $\boldsymbol{X}$ given the value of the sufficient statistic $S(\boldsymbol{X})$ does not depend on the parameter θ.

When testing equality of independent binomial proportions the joint probability mass function (pmf) under H_0 is given by

$$\Pr_\theta(X_1 = x_1, X_2 = x_2) = \binom{n_1}{x_1}\binom{n_2}{x_2}\theta^{x_1+x_2}(1-\theta)^{n_1+n_2-(x_1+x_2)}. \tag{4.7}$$

The joint pmf depends on θ. Equivalently we can think that outcomes from the experiment (which are realisations distributed according to the stated pmf under H_0) carry information about θ. To understand this, we should try to picture the joint pmf. In the figure in the lower panel of Figure 2.1 we have plotted the pmf of X_1 when $\theta = 0.5$ and $n_1 = 16$. The pmf of X_2 will have a similar shape (in the sense that it has a peak around $n_2\theta$, and since θ is the same the peak will be around $n_2/2$. The shape can be a little different since n_2 need not equal n_1, so that it might appear either elongated or compressed.) When imaging the joint pmf it may help to think that for each fixed value of x_2, the shape of the pmf along the x_1-axis is exactly the same as in the figure in the bottom of Figure 2.1, but all the values are multiplied with $\Pr_\theta(X_2 = x_2)$. When $\theta \to 0$ more and more of the mass of the joint mass function is concentrated around the origin and when $\theta \to 1$ more and more of the mass is concentrated around $(x_1, x_2) = (n_1, n_2)$. This also means the long run frequency of each possible realisation from the joint distribution will change as θ changes. It is therefore natural to think that the outcomes in the experiment carry information about θ. This also means a function of the outcome carry information about θ. We would like that the function of the outcome captures all the information about θ there is, since when we know the value of this function

there is no need to know the outcome in the experiment in order to get more information about θ.

The sufficient statistic captures all the available information about θ there is in the outcome $\boldsymbol{X}$, since once we know this value the distribution of the outcome is independent of θ. This means once we know the value of the sufficient statistic, the outcome cannot tell us anything further about θ. To show that a statistic is sufficient we can either guess a statistic and condition the outcome in the experiment on it and show that this conditional distribution is independent of θ or we can use the *factorisation theorem*.

Theorem 4.2.1 (Factorisation theorem) *If $f(\boldsymbol{x};\theta)$ denotes the joint probability mass function of the outcome in an experiment $\boldsymbol{X}$ then*

$$S(\boldsymbol{X}) \text{ is sufficient for } \theta$$

if and only if

$$\begin{gathered}\text{There exists functions } g(s;\theta) \text{ and } h(\boldsymbol{x}) \text{ so that} \\ f(\boldsymbol{x};\theta) = g(S(\boldsymbol{x});\theta)h(\boldsymbol{x}) \qquad (4.8)\end{gathered}$$

for all outcomes $\boldsymbol{x}$ and values of the parameter θ, see e.g Casella & Berger (2002, p. 276)

If we try to factor the joint probability mass in Equation (4.7) as done in the factorisation theorem we get that

$$\begin{aligned} f(\boldsymbol{x};\theta) &= \binom{n_1}{x_1}\binom{n_2}{x_2}\theta^{x_1+x_2}(1-\theta)^{n_1+n_2-(x_1+x_2)} = \theta^{S(\boldsymbol{x})}(1-\theta)^{n_1+n_2-S(\boldsymbol{x})}h(\boldsymbol{x}) \\ &= g(S(\boldsymbol{x});\theta)h(\boldsymbol{x}), \end{aligned} \qquad (4.9)$$

where $S(\boldsymbol{x}) = x_1 + x_2$ and $h(\boldsymbol{x}) = \binom{n_1}{x_1}\binom{n_2}{x_2}$. This means that $T(\boldsymbol{X}) = X_1 + X_2$ is sufficient for θ under H_0. We also want the conditional distribution of $\boldsymbol{X}$ given the value of the sufficient statistic. We have that X_1+X_2 is binomially distributed with parameters θ and $n_1 + n_2$ since the three conditions for the binomial distribution are satisfied: 1) we have $n_1 + n_2$ independent trials since the n_1 trials where X_1 gives the number of successes are independent and independent of all the n_2 trials where X_2 gives the number of successes. The same holds for the n_2 trials. 2) In each trial there are two outcomes, either success or failure. And 3) in each trial the probability for success is θ. One should also note that $X_1 + X_2 = s$ specifies a line in the x_1x_2-plane. This line has slope -1, i.e $X_1 = s - X_2$ or $X_2 = s - X_1$. Therefore the distribution of $\boldsymbol{X} = (X_1, X_2)$ given $X_1 + X_2 = s$ is one dimensional.

We get

$$\begin{aligned}
\Pr(X_1 = x_1, X_2 = x_2 | S(\boldsymbol{X}) = s) &\overset{(\star)}{=} \frac{\Pr_\theta(X_1 = x_1, X_2 = x_2, X_1 + X_2 = s)}{\Pr_\theta(X_1 + X_2 = s)} \\
&\overset{\star\star}{=} \frac{\Pr_\theta(X_1 = x_1, X_2 = s - x_2)}{\Pr_\theta(X_1 + X_2 = s)} \\
&\overset{(\star\star\star)}{=} \frac{\Pr_\theta(X_1 = x_1)\Pr_\theta(X_2 = s - x_1)}{\Pr_\theta(X_1 + X_2 = s)} \\
&\overset{(\star\star\star\star)}{=} \frac{\binom{n_1}{x_1}\theta^{x_1}(1-\theta)^{n_1 - x_1}\binom{n_2}{s-x_1}\theta^{s-x_1}(1-\theta)^{n_2-(s-x_1)}}{\binom{n_1+n_2}{s}\theta^s(1-\theta)^{n_1+n_2-s}} \\
&= \frac{\binom{n_1}{x_1}\binom{n_2}{s-x_1}\theta^s(1-\theta)^{n_1+n_2-s}}{\binom{n_1+n_2}{s}\theta^s(1-\theta)^{n_1+n_2-s}} \\
&= \frac{\binom{n_1}{x_1}\binom{n_2}{s-x_1}}{\binom{n_1+n_2}{s}},
\end{aligned} \tag{4.10}$$

where we see that the resulting distribution is one dimensional, as previously explained. We use Bayes rule in transition $(\star)$. Transition $(\star\star)$ is only valid if $x_1 + x_2 = s$. We try to give some further insight into this transition. The event $\{X_1 + X_2 = s\}$ is the collection of all outcomes such that $x_1 + x_2 = s$. In the event $\{X_1 = x_1\}$ x_1 can be any possible value of X_1 and in the event $\{X_2 = x_2\}$ x_2 can be any possible value of x_2. When we intersect $\{X_1 = x_1\}$ with $\{X_2 = x_2\}$, x_1 and x_2 can still be any possible value of respectively X_1 and X_2 since X_1 and X_2 are independent random variables. However, when we intersect $\{X_1 = x_1, X_2 = x_2\}$ with $\{X_1 + X_2 = s\}$ the only possible outcome is (x_1, x_2) such that $x_1 + x_2 = s$. This means only one of X_1 or X_2 is free to vary. We choose X_1 to be the free random variable. When $X_1 = x_1$ we know X_2 must be $s - x_1$. We could of course have chosen X_2 as the free random variable. Then the result would be $\binom{n_1}{s-x_2}\binom{n_2}{x_2}/\binom{n_1+n_2}{s}$ instead of the stated expression. In transition $(\star\star\star)$ we use that X_1 and X_2 are unconditionally independent and finally in transition $(\star\star\star\star)$ we use that $X_1 \sim \text{Binom}(\theta, n_1)$, $X_2 \sim \text{Binom}(\theta, n_2)$ and $X_1 + X_2 \sim \text{Binom}(\theta, n_1 + n_2)$ under H_0.

4.2.2 Maximum likelihood estimators of $\theta, \theta_1, \theta_2$

When the outcome in an experiment is $\boldsymbol{X}$ and the joint probability mass function is given by $f(\boldsymbol{x}; \boldsymbol{\theta})$ the *likelihood function* is $L(\boldsymbol{\theta}; x) = f(\boldsymbol{x}; \boldsymbol{\theta})$, i.e the likelihood function has the same expression as the joint probability mass function but we

view it as a function of $\boldsymbol{\theta}$ and not as a function of $\boldsymbol{x}$. The *maximum likelihood estimator* (mle) is then defined as

$$\hat{\boldsymbol{\theta}}_{ML} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}; x) = \arg\max_{\boldsymbol{\theta}} f(\boldsymbol{x}; \boldsymbol{\theta}). \tag{4.11}$$

To be more correct we find the realisation of the maximum likelihood estimator by using Equation (4.11). The maximum likelihood estimator is found by replacing $\boldsymbol{x}$ with $\boldsymbol{X}$ in the result in Equation (4.11). The observed value of the maximum likelihood estimator is the value of the parameter $\boldsymbol{\theta}$ that maximises the probability of obtaining the observed outcome in the experiment (Christensen 1998, p. 252). This holds for *all* observed outcomes.

From Equation (4.7) the likelihood function when testing equality of independent binomial proportions under H_0 is

$$L(\theta; \boldsymbol{x}) = \binom{n_1}{x_1}\binom{n_1}{x_2}\theta^{x_1+x_2}(1-\theta)^{n_1+n_1-(x_1+x_2)}.$$

We know $\theta \in [0, 1]$, which is a closed and finite interval. Furthermore L is a continuous function of θ. From the extreme value theorem (Theorem 6) in Adams & Essex (2010, p. 233) $L(\theta, \boldsymbol{x})$ must have an absolute maximum and minimum on $[0, 1]$. We must look for candidates among 1) critical points, 2) singular points and 3) endpoints. Since $L(\theta; \boldsymbol{x})$ is an infinitely many times differentiable function of θ it has no singular points. Furthermore $L(0; \boldsymbol{x}) = L(1; \boldsymbol{x}) = 0$, which give the values of the likelihood function at the two endpoints.

We have that $\ln u$ is a strictly increasing function of u, which means $\ln u_1 < \ln u_2$ if and only if $u_1 < u_2$. Therefore critical points of u are the same as critical points of $\ln u$. So when we want to find the critical points of $L(\theta; \boldsymbol{x})$ we can instead find the critical points of $\ln L(\theta; \boldsymbol{x})$, which is easier. By differentiating $\ln L$ with respect to θ we get

$$\begin{aligned}
&\frac{\mathrm{d}\ \ln L(\theta; \boldsymbol{x})}{\mathrm{d}\theta} \\
&= \frac{\mathrm{d}}{\mathrm{d}\theta}\left(\ln\left(\binom{n_1}{x_1}\binom{n_2}{x_2}\right) + (x_1 + x_2)\ln\,\theta + (n_1 + n_2 - (x_1 + x_2))\ln(1-\theta)\right) \\
&= \frac{x_1 + x_2}{\theta} - \frac{n_1 + n_2 - (x_1 + x_2)}{1-\theta}.
\end{aligned}$$

Solving $\frac{\mathrm{d}\ \ln L(\theta;\boldsymbol{x})}{\mathrm{d}\theta} = 0$ for θ gives

$$\theta_{CP} = \frac{x_1 + x_2}{n_1 + n_2}.$$

We see that $L(\theta_{CP}; \boldsymbol{x}) > 0$. Since the likelihood function is 0 at the endpoints, θ_{CP} must give the absolute maximum of $L(\theta; \boldsymbol{x})$ on $[0, 1]$. The maximum likelihood estimator of θ under H_0 is then

$$\hat{\theta}_{ML} = \frac{X_1 + X_2}{n_1 + n_2}. \tag{4.12}$$

By similar calculations we can show that the maximum likelihood estimator of θ_1 is given by

$$\hat{\theta}_{1,ML} = \frac{X_1}{n_1}. \tag{4.13}$$

and that the maximum likelihood estimator of θ_2 is given by

$$\hat{\theta}_{2,ML} = \frac{X_2}{n_2}. \tag{4.14}$$

4.2.3 Types of convergence

Convergence in probability

A sequence of random variables $\{X_n, n \geq 1\} = \{X_1, X_2, \ldots\}$ *converges in probability* to the random variable X if for every $\epsilon > 0$

$$\lim_{n \to \infty} \Pr(|X_i - X| \geq \epsilon) \to 0, \tag{4.15}$$

or equivalently

$$\lim_{n \to \infty} \Pr(|X_i - X| < \epsilon) \to 1, \tag{4.16}$$

which we in shorthand notation write as $X_n \xrightarrow{p} X$ (Casella & Berger 2002, p. 232). It is important to realise that the limit in Equation (4.15) consists of real numbers between 0 and 1 and that this sequence of numbers either converges to 0 or 1 depending on which of the two equivalent formulations we use. Roughly speaking, when $X_n \xrightarrow{p} X$ then X_i and X tend to take more and more similar values as i gets larger.

Consistency

Consider a sequence $\{T_n, n \geq 1\} = \{T_1, T_2, \ldots\}$, where $T_i = T_i(X_1, \ldots, X_i)$, of estimators for some unknown real valued function $\phi(\theta)$. We have that the given sequence is *consistent* for $\phi(\theta)$ if and only if

$$T_n \xrightarrow{p} \phi(\theta)$$

(Mukhopadhyay 2000, p. 380).

Convergence in distribution

A sequence of random variables $\{X_1, X_2, \ldots\}$ *converges in distribution* to a random variables X if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

at all points x where $F_X(x)$ is continuous, which we write as

$$X_n \xrightarrow{d} X$$

(Casella & Berger 2002, p. 235). As noted in Casella & Berger (2002, p. 235) it is really the sequence of pmfs and not the sequence of random variables that converge. We also note that convergence in probability to a random variable X is stronger than convergence in distribution. The reason is that when a sequence of random variables converge in probability the X_i tend to take more and more similar values as X when i increases but when the sequence converges in distribution we only know that their distribution functions tend to be more and more similar as i increases, which does not does tell us anything about how similar the values of X and X_i are when i increases. In can be shown that convergence in probability implies convergence in distribution (see for instance Casella & Berger (2002, p. 236)).

4.2.4 Convergence results

Theorem 4.2.2 (The weak law of large numbers) *If $X_1, X_2, \ldots,$ is a sequence of iid random variables where $\mathrm{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) < \infty$, then*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} \bar{X} \xrightarrow{p} \mu$$

(Casella & Berger 2002, p. 233). According to Theorem 4.2.2 the sample mean $\bar{X}$ is a consistent estimator for μ and is thus a natural estimator for μ. Note that if $\bar{X} \xrightarrow{p} Y$ where for instance $Y \sim N(0, 1)$, then $\bar{X}$ would be useless as an estimator for μ since once given an infinite amount of information the estimator does not produce the true value of μ.

Theorem 4.2.3 (The central limit theorem) *If $X_1, X_2, X_3, \ldots$ is a sequence of iid random variables where $\mathrm{E}(X_i) = \mu$ and $\mathrm{Var}(X) = \sigma^2 < \infty$ and we let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ then*

$$\frac{\bar{X} - \mu}{\sqrt{\mathrm{Var}(\bar{X})}} = \sqrt{n}\frac{(\bar{X} - \mu)}{\sigma} \xrightarrow{d} N(0, 1)$$

(Casella & Berger 2002, p. 238). We try to give some insight into Theorem 4.2.3. From Theorem 4.2.2 we know that $\bar{X} \xrightarrow{p} \mu$. We have that $\text{Var}(\bar{X})$ tends to 0 as $n \to \infty$. This means we divide "oscillations" around the mean μ by a number that tends to 0, which means we blow the oscillations up. The result is random variable distributed as a standard normal variable. Alternatively, since we multiply $\frac{\bar{X}-\mu}{\sigma}$ by $\sqrt{n}$ in Equation (4.2.3) we can think that we blow the oscillations up (i.e the term $\frac{X-\mu}{\sigma}$) by multiplying with this term.

Theorem 4.2.4 (Slutsky's Theorem) *If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} a$ (a is a constant) then $Y_n X_n \xrightarrow{d} aX$ (Casella & Berger 2002, p. 239–240).*

Theorem 4.2.5 *If the sequence $\{X_1, X_2, \ldots\}$ converges in probability to the random variable X and if $h(u)$ is a continuous function of u then the sequence $\{h(X_1), h(X_2), \ldots\}$ converges to $h(X)$ (Casella & Berger 2002, p. 233).*

Proof:
Recall the definition of continuity. If f is continuous at a point a, then for every $\epsilon > 0$ there exists a $\delta > 0$ such that (Adams & Essex 2010, p. 88)

$$|x - a| < \delta \Rightarrow |f(x) - f(a)| < \epsilon.$$

This means if h is continuous then for very $\epsilon > 0$ there exists a $\delta > 0$ such that

$$\Pr(|X - Y|) < \delta \Rightarrow \Pr(|h(X) - h(Y)| < \epsilon.$$

Also recall the definition of a sequence that converges to a number L. We say that $\lim x_n = L$ if there for every $\epsilon > 0$ exists a positive number $k = k(\epsilon)$ such that $|x_n - L| < \epsilon$ holds when $n \geq k$ (Adams & Essex 2010, p. 498).

To prove the theorem we need to show that there for every ϵ_2 exists a number k_2 such that

$$|\Pr(|h(X_n) - h(X)|) - 1| < \epsilon_2$$

whenever $n \geq k_2$. Since h is continuous and X_n converges in probability to X, we know we can find ϵ_3 and k_3 such that

$$|\Pr(|X_n - X|) - 1| < \epsilon_3 \Rightarrow |\Pr(|h(X_n) - h(X)|) - 1| < \epsilon_2$$

holds whenever $n \geq k_3$. This means $h(X_n)$ converges in probability to $h(X)$.

The result in Theorem 4.2.5 is very intuitive. We know that X_n converges to X in probability so that X_n and X take more and more similar values as n grows, and since h is continuous we should be able to get $h(X_n)$ as "close" as we want to $h(X)$ by choosing n big enough.

Theorem 4.2.6 *We have that*

$$X_n \xrightarrow{p} \mu$$
$$\textit{if and only if}$$
$$X_n \xrightarrow{d} X$$

where

$$F_X(x) = \begin{cases} 0 & \textit{if } x < \mu \\ 1 & \textit{if } x > \mu \end{cases}$$

(Casella & Berger 2002, p. 236). Theorem 4.2.6 means convergence in probability and convergence in distribution are equivalent when the limit is a constant.

Theorem 4.2.7 (Consistency of maximum likelihood estimators) *Let $X_1, X_2, \ldots, X_n$ be iid X where X has density function $f(x;\theta)$, define the likelihood function by $L(\theta;\boldsymbol{x}) = \prod_{i=1}^{n} f(x_i;\theta)$ and assume that $\hat{\theta}$ is the maximum likelihood estimator of θ. Then, under some regularity conditions (see for instance Casella & Berger (2002, p. 516))*

$$\tau(\hat{\theta}) \xrightarrow{p} \tau(\theta).$$

(Casella & Berger 2002, p. 470). Theorem 4.2.7 states that $\tau(\hat{\theta})$ is a consistent estimator of $\tau(\theta)$. Intuitively, when we get more and more data an estimator should be able to estimate a parameter with better and better precision. It then seems intuitive that the maximum likelihood estimator should tend to take values closer and closer to $\tau(\theta)$ (i.e the function of θ it is estimating) as the sample size increases if it is to be regarded as a reasonable estimator for $\tau(\theta)$. This is exactly what the above result states.

4.3 Different test statistics

When using the different methods for creating a p-value we need a test statistic. In this section we consider different candidates for the test statistic and find a large sample distribution of each candidate. We want to use the same test statistic in each of the four mentioned methods for creating a p-value given in Section 4.1.2.

4.3.1 The test statistic $|D|$

When testing the hypotheses in Equation (4.1) an intuitive test statistic is $D(X_1, X_2) = \hat{\theta}_{1,ML} - \hat{\theta}_{2,ML}$ where $\hat{\theta}_{1,ML}$ is given by Equation (4.13) and $\hat{\theta}_{2,ML}$ is given by Equation (4.14), i.e we look at the difference between the maximum likelihood estimators of θ_1 and θ_2 (Høyland 1986, p. 80). However, both large and negative numbers with large magnitude indicate that H_0 is not true. We only want large numbers to indicate that H_0 is not true, and the larger the value the stronger the indication. One possible solution is to look at the absolute value of D, i.e consider

$$|D|(X_1, X_2) = |D(X_1, X_2)| = \left| \frac{X_1}{n_1} - \frac{X_2}{n_2} \right|$$

Purpose of the test statistic

Before we consider the asymptotic distribution of $|D|$, we take a closer look at the purpose of the test statistic. From Section 2.6 we know that the test statistic T is a real valued function of the outcome $\boldsymbol{X}$ in an experiment (i.e $T = T(\boldsymbol{X})$ and $\forall \boldsymbol{x} \in \mathcal{S}, T(\boldsymbol{x}) \in \mathbb{R}$) and the higher the value of the test statistic, the stronger the indication that the alternative hypothesis is true. Moreover, we say that the test statistic *orders* the outcomes in the sample space (Royall 1997, p. 63). To understand this property better, we compare the ordering given by the test statistic $|D|$ with the ordering induced by the test statistic $|D|/10(\boldsymbol{X}) = |D|(\boldsymbol{X})/10$. We set $n_1 = 2$ and $n_2 = 2$ and evaluate both test statistics in all of the possible outcomes. For each outcome we put the outcome and the test statistics evaluated in the outcome in a separate row in a table. If we order the rows according to increasing value of $|D|(\boldsymbol{x})$ from top to bottom we get Table 4.4.

Imagine that we for each test statistic create boxes. We create one box for each unique value of the test statistic. This means we create three boxes for each of the two test statistics. We then place outcomes with the same value of the test statistic in the same box. If we consider the boxes for $|D|(\boldsymbol{X})$, the outcomes $(0,0), (1,1)$ and $(2,2)$ are placed in the same box, the outcomes $(1,0), (0,1), (1,2)$ and $(2,1)$ are placed in another box and the remaining two outcomes $(2,0)$ and $(0,2)$ are placed in the last box. We then order the boxes so that the box where the outcomes have the smallest value of the test statistic comes first, then comes the box with the second smallest value of the test statistic and thereafter comes the box with the largest value of the test statistic. This means the box with the outcome $(0,0)$ comes first and the box with the outcome $(0,2)$ comes last.

If we do a similar procedure for the boxes for the $|D|/10$-statistic we see that both the elements of the boxes and the order of the boxes are the same as for the boxes

for the $|D|$-statistic. What does this mean? When we calculate the realisation of the p-value in an experiment and use the enumeration procedure to calculate the realisation, or to be more exact use either the E or M method, we start with computing the test statistic for all outcomes. Then we compare the value of the test statistic and consider the outcomes with a least as large test statistic as the outcome. In our box example it is possible to compare the value of the test statistic with the value of the test statistic in each of the boxes. We start with the first box. We compare the values of the test statistic until they are equal. Due to the ordering of the boxes, we need to consider the box where the values of the test statistic are equal and the remaining boxes when we calculate the realisation of the p-value. Since the boxes of $|D|(\boldsymbol{X})$ and of $|D|/10(\boldsymbol{X})$ contain the same elements and are ordered the same, the realisations of the p-values calculated using the different test statistics must be the same. This means it is not the magnitude of the value of the test statistic that matters when we calculate realisations of p-values using the enumeration procedure, only the ordering induced matters. We say that $|D|$ and $|D|/10$ order the sample space in the same way. If two test statistics induce the same ordering of the sample space and we use both test statistics in turn as test statistic when calculating the realisation of the p-value for an outcome in the experiment, we calculate the same realisation if we use the same method (E or M). As a note, if we use the C-method then we only look at tables (y_1, y_2) such that $y_1 + y_2 = x_1 + x_2$. This means we look at a subset of the original outcomes. We can still order the outcomes in boxes as previously done. The result will be the same: the number of boxes, the order of them and the outcomes in them will be the same for the boxes for $|D|$ and for $|D|/10$. This means we calculate the same p-value when we use the C-method on the test statistics $|D|$ and $|D|/10$.

We have seen that two test statistics that orders the sample space the same need not take the same values. What matters is if the created imaginary boxes for the unique values of the test statistics are ordered in the same way and contain the same elements when we compare the boxes from one test statistic with the boxes from the other. Mathematically we can write the condition as $T_1(x_1) > T_1(x_2)$ if and only of $T_2(x_1) > T_2(x_2)$ for all outcomes x_1, x_2 and $T_1(x_1) = T_1(x_2)$ if and only if $T_2(x_1) = T_1(x_2)$. If the two conditions hold, $T_1(x)$ and $T_2(x)$ orders the sample space the same.

4.3.2 Asymptotic distribution of $|D|$

We need a large sample distribution of $|D|(X_1, X_2)$ under H_0 if we want to calculate the A p-value with $|D|(X_1, X_2)$ as test statistic. We know that X_1 is a sum of n_1 independent Bernoulli random variables with success probability θ under H_0 and

Table 4.4: Comparing the orderings induced by $|D|$ and $|D|/10$ when $n_1 = 2$ and $n_2 = 2$. The possible outcomes are given in the leftmost column. For each row the second column gives the value of the test statistic $|D|$ evaluated in the outcome in the first column and the third column gives the value of $|D|/10$ evaluated in the same outcome. The rows are ordered in increasing value of $|D|$ from top to bottom .

$\boldsymbol{x}$	$\lvert D\rvert(\boldsymbol{x})$	$\frac{\lvert D\rvert}{10}(\boldsymbol{x})$
(0,0)	0	0
(1,1)	0	0
(2,2)	0	0
(1,0)	0.25	0.025
(0,1)	0.25	0.025
(1,2)	0.25	0.025
(2,1)	0.25	0.025
(2,0)	1	0.1
(0,2)	1	0.1

that X_2 is a sum of n_2 Bernoulli random variables with the same success probability under H_0. By the weak law of large numbers it follows that

$$\frac{X_1}{n_1} \xrightarrow{p} \theta \tag{4.17}$$

and

$$\frac{X_2}{n_2} \xrightarrow{p} \theta. \tag{4.18}$$

Due to Equation (4.17) and Equation (4.18) it seems natural that

$$\frac{X_1}{n_1} - \frac{X_2}{n_2} \xrightarrow{p} 0. \tag{4.19}$$

We show Equation (4.19) by using Chebychev's inequality (see for instance Casella & Berger (2002, p. 122)). We get

$$\begin{aligned}
\Pr_\theta\left(\left|\frac{X_1}{n_1} - \frac{X_2}{n_2} - 0\right| \geq \epsilon\right) &= \Pr_\theta\left(\left|\frac{X_1}{n_1} - \frac{X_2}{n_2}\right|^2 \geq \epsilon^2\right) = \Pr_\theta\left(\left(\frac{X_1}{n_1} - \frac{X_2}{n_2}\right)^2 \geq \epsilon^2\right) \\
&\overset{(\star)}{\leq} \frac{\mathrm{E}\left(\left(\frac{X_1}{n_1} - \frac{X_2}{n_2}\right)^2\right)}{\epsilon^2} = \frac{\mathrm{Var}\left(\frac{X_1}{n_1} - \frac{X_2}{n_2}\right)}{\epsilon^2} \\
&\overset{(\star\star)}{=} \frac{\frac{\mathrm{Var}(X_1)}{n_1^2} + \frac{\mathrm{Var}(X_2)}{n_2^2}}{\epsilon^2} = \frac{\frac{n_1\theta(1-\theta)}{n_1^2} + \frac{n_2\theta(1-\theta)}{n_2^2}}{\epsilon^2} \\
&= \frac{\theta(1-\theta)(\frac{1}{n_1} + \frac{1}{n_2})}{\epsilon^2} \to \frac{\theta(1-\theta)(0+0)}{\epsilon^2} = 0,
\end{aligned}$$

as $n_1, n_2 \to \infty$, where we use Chebychev's inequality in transition ($\star$) and equality in transition ($\star\star$) follows since X_1 and X_2 are independent.

We therefore know asymptotically that $\Pr(D(X_1, X_2) = 0) = 1$ and $\Pr(D(X_1, X_2) \neq 0) = 0$ under H_0 . It therefore follows (asymptotically) $\Pr(|D|(X_1, X_2) = 0) = 1$ and $\Pr(|D|(X_1, X_2) \neq 0) = 0$. So, if we use the asymptotic distribution of $|D|(X_1, X_2)$ we always get the realisation 1 of the p_A-value since the large sample distribution of $|D|$ is a constant under H_0. This means the power function is 0 for $(\theta_1, \theta_2) \in \Theta_0^c$. We want an asymptotic test with better power properties.

4.3.3 The statistic Z_p^2

One possibility to obtain a test statistic with better asymptotic properties than $|D|$ is to divide $D(X_1, X_2)$ by the standard deviation of $D(X_1, X_2)$ under H_0. Since each of X_1 and X_2 can be regarded as averages of Bernoulli random variables and have the same expected value θ, it seems natural by the central limit theorem that the resulting random variable converges to a standard normal variable. We therefore consider

$$\frac{\frac{X_1}{n_1} - \frac{X_2}{n_2}}{\sqrt{\mathrm{Var}(\frac{X_1}{n_1} - \frac{X_2}{n_2})}} = \frac{\frac{X_1}{n_1} - \frac{X_2}{n_2}}{\sqrt{\theta(1-\theta)(\frac{1}{n_1} + \frac{1}{n_2})}} \tag{4.20}$$

We do not know the true value of θ, else there would be no need to perform the hypothesis test and $\theta_1 \neq \theta_2$ under H_0. We therefore need to replace θ in Equation (4.20) with some estimate. One possibility is to replace it with the maximum likelihood estimator, which is given in Equation (4.12). Since this estimator corresponds to combining the data from the two binomial experiments it is called a pooled estimator. We get

$$Z_p(X) = \frac{\frac{X_1}{n_1} - \frac{X_2}{n_2}}{\sqrt{\frac{X_1+X_2}{n_1+n_2}(1 - \frac{X_1+X_2}{n_1+n_2})(\frac{1}{n_1} + \frac{1}{n_2})}}, \tag{4.21}$$

which is called the Z-pooled statistic since it has the form of a Z-statistic and we use the pooled estimate of θ (Lydersen et al. 2012). We want only large values of the test statistic to indicate that the null hypothesis is not true. Now both negative numbers with large magnitude and large positive numbers indicate that H_0 is not true. If we square the statistic then only large values of the resulting statistic indicate that H_1 is true. We get

$$Z_p^2(X_1, X_2) = \frac{(\frac{X_1}{n_1} - \frac{X_2}{n_2})^2}{\frac{X_1+X_2}{n_1+n_2}(1 - \frac{X_1+X_2}{n_1+n_2})(\frac{1}{n_1} + \frac{1}{n_2})}, \tag{4.22}$$

which we call the Z-pooled squared statistic. We note that $Z_p^2(n_1, n_2)$ and $Z_p^2(0, 0)$ are undefined since we get the undetermined expression $0/0$ in both cases. We set $Z_p^2(n_1, n_2) = Z_p^2(0, 0) = -99$, so that we make these two outcomes the weakest evidence against the null hypothesis.

Asymptotic distribution of Z_p^2

We want to show that Z_p^2 defined in Equation (4.22) is asymptotically distributed as a chi square distribution with one degree of freedom (Høyland 1986, p. 82). We start by showing that $Z_p(X_1, X_2)$ defined in Equation (4.21) has the standard normal distribution as n_1 and n_2 approach infinity.

We know n_1 and n_2 can approach infinity in numerous different ways. For instance, n_1 may approach infinity linearly, one example is $n_1 = 50 \cdot t$ for $t = 0, 1, 2, \ldots$, and n_2 may approach infinity exponentially, e.g $n_2 = 40 \cdot 2^t$. In deriving the asymptotic distribution of $Z_p(X_1, X_2)$ we set $N = n_1 + n_2$ and assume $\frac{n_1}{N} \to \rho$ as $n_1, n_2 \to \infty$ where $\rho > 0$. Since $\frac{n_1}{N} = \frac{n_1}{n_1+n_2} = \frac{1}{1+\frac{n_2}{n_1}}$ we have $\frac{n_2}{n_1} \to c > 0$, which means that by setting $\frac{n_1}{N} \to \rho$ we demand that n_1 and n_2 approach infinity at the same speed. This means we have limited the number of ways n_1 and n_2 can approach infinity. For instance if n_1 approaches infinity exponentially n_2 must also approach infinity exponentially and with the same base (i.e if $n_1 = c_1 2^t$ then $n_2 = c_2 2^t$ for some constants $c_1, c_2 > 0$). In Equation (4.21) X_1 is a sum of independent Bernoulli random variables with the same success probability θ, i.e a sum of 0's and 1's. This means that $\frac{X}{n_1}$ may be considered the average of the n_1 Bernoulli random variables. A similar interpretation holds for $\frac{X_2}{n_2}$. Since the variance of one Bernoulli random variable is $\theta(1-\theta)$ and the expectation is θ the central limit theorem gives

$$\sqrt{n_1}\frac{\frac{X_1}{n_1} - \theta}{\sqrt{\theta(1-\theta)}} \to N(0, 1) \tag{4.23}$$

and

$$\sqrt{n_2}\frac{\frac{X_2}{n_2} - \theta}{\sqrt{\theta(1-\theta)}} \to N(0, 1). \tag{4.24}$$

Since $\sqrt{\frac{N}{n_1}} \to \frac{1}{\sqrt{\rho}}$ we have by definition that $\sqrt{\frac{N}{n_1}} \xrightarrow{p} \frac{1}{\sqrt{\rho}}$. By Slutsky's Theorem we get

$$\sqrt{N}\frac{\frac{X_1}{n_1} - \theta}{\sqrt{\theta(1-\theta)}} = \sqrt{\frac{N}{n_1}}\sqrt{n_1}\frac{\frac{X_1}{n_1} - \theta}{\sqrt{\theta(1-\theta)}} \xrightarrow{d} \frac{1}{\sqrt{\rho}} \cdot N(0, 1) = N(0, \frac{1}{\rho}) \tag{4.25}$$

since $\sqrt{n_1}\frac{\frac{X_1}{n_1}-\theta}{\sqrt{\theta(1-\theta)}} \xrightarrow{d} N(0,1)$ by Equation (4.23) and $\sqrt{\frac{N}{n_1}} \xrightarrow{p} \frac{1}{\sqrt{\rho}}$. Furthermore

$$\begin{aligned}
\frac{N}{N} = 1 = \frac{n_1+n_2}{N} &= \frac{n_1}{N} + \frac{n_2}{N} \\
\frac{n_2}{N} &= 1 - \frac{n_1}{N} \\
\lim_{n_1,n_2\to\infty} \frac{n_2}{N} &= \lim_{n_1,n_2\to\infty}(1-\frac{n_1}{N}) \\
\lim_{n_1,n_2\to\infty} \frac{n_2}{N} &= 1-\rho.
\end{aligned}$$

By definition we then have $\frac{n_2}{N} \xrightarrow{p} 1-\rho$. By similar argumentation as in Equation (4.25) we get

$$\sqrt{N}\frac{\frac{X_2}{n_2}-\theta}{\sqrt{\theta(1-\theta)}} \xrightarrow{d} \frac{1}{\sqrt{1-\rho}}N(0,1) = N(0,\frac{1}{1-\rho}) \tag{4.26}$$

Since X_1 and X_2 are independent

$$\sqrt{N}\frac{\frac{X_1}{n_1}-\theta}{\sqrt{\theta(1-\theta)}} - \sqrt{N}\frac{\frac{X_2}{n_2}-\theta}{\sqrt{\theta(1-\theta)}} = \sqrt{N}\frac{\frac{X_1}{n_1}-\frac{X_2}{n_2}}{\sqrt{\theta(1-\theta)}} \xrightarrow{d} N(0,\frac{1}{\rho(1-\rho)}) \tag{4.27}$$

by the addition property of independent normal distributions and since $\frac{1}{\rho}+\frac{1}{1-\rho} = \frac{1}{\rho(1-\rho)}$. By expanding the expression for $Z_p(X_1,X_2)$ we get

$$\begin{aligned}
Z_p(X_1,X_2) &= \frac{\frac{X_1}{n_1}-\frac{X_2}{n_2}}{\sqrt{\frac{X_1+X_2}{n_1+n_2}(1-\frac{X_1+X_2}{n_1+n_2})(\frac{1}{n_1}+\frac{1}{n_2})}} \\
&= \frac{\sqrt{N\rho(1-\rho)\theta(1-\theta)}}{\sqrt{N\rho(1-\rho)\theta(1-\theta)}} \frac{\frac{X_1}{n_1}-\frac{X_2}{n_2}}{\sqrt{\frac{X_1+X_2}{n_1+n_2}(1-\frac{X_1+X_2}{n_1+n_2})(\frac{1}{n_1}+\frac{1}{n_2})}} \\
&= \frac{1}{\sqrt{N(\frac{1}{n_1}+\frac{1}{n_2})\rho(1-\rho)}} \frac{\sqrt{\theta(1-\theta)}}{\sqrt{\frac{X_1+X_2}{n_1+n_2}(1-\frac{X_1+X_2}{n_1+n_2})}} \sqrt{N}\frac{\frac{X_1}{n_1}-\frac{X_2}{n_2}}{\sqrt{\frac{\theta(1-\theta)}{\rho(1-\rho)}}}
\end{aligned} \tag{4.28}$$

It may be tempting to use Theorem 4.2.5 to conclude that

$$\frac{1}{\sqrt{N(\frac{1}{n_1}+\frac{1}{n_2})\rho(1-\rho)}} \frac{\sqrt{\theta(1-\theta)}}{\sqrt{\frac{X_1+X_2}{n_1+n_2}(1-\frac{X_1+X_2}{n_1+n_2})}} \xrightarrow{p} 1. \tag{4.29}$$

where

$$h(t) = \frac{\sqrt{\theta(1-\theta)}}{\sqrt{N(\frac{1}{n_1}+\frac{1}{n_2})\rho(1-\rho)t(1-t)}}$$

However, this would be incorrect since $h(t)$ changes as n_1 and n_2 approach infinity so that $h(t)$ is not continuous. (We note that $\sqrt{t}$ is continuous, but that $\sqrt{(1/n_1 + 1/n_2)t}$ changes as $n_1, n_2 \to \infty$ so that this function is not continuous.) Instead we use Slutsky's Theorem twice, or iteratively. The first time we use Slutsky, we start with considering $\frac{X_1+X_2}{n_1+n_2}$. We know that this estimator is the maximum likelihood estimator of θ under H_0. We also know that maximum likelihood estimators are consistent, which means that

$$\frac{X_1 + X_2}{n_1 + n_2} \xrightarrow{p} \theta. \tag{4.30}$$

This is easy to show directly in our case. We know from Section 4.2.1 that $Y = X_1 + X_2$ is binomially distributed with parameters θ and $n_1 + n_2$. Therefore $\frac{Y}{n_1+n_2}$ can be considered the average of $n_1 + n_2$ independent Bernoulli trials each with the same success probability θ. By the weak law of large numbers we then get Equation (4.30). By using Theorem 4.2.5 it follows that

$$\frac{\sqrt{\theta(1-\theta)}}{\sqrt{\frac{X_1+X_2}{n_1+n_2}(1-\frac{X_1+X_2}{n_1+n_2})}} \xrightarrow{p} 1 \tag{4.31}$$

since $f(x) = \frac{\sqrt{\theta(1-\theta}}{\sqrt{x(1-x)}}$ is continuous.

By Equation (4.27) we have

$$\sqrt{N}\frac{\frac{X_1}{n_1} - \frac{X_2}{n_2}}{\sqrt{\frac{\theta(1-\theta)}{\rho(1-\rho)}}} \xrightarrow{d} N(0,1) \tag{4.32}$$

By using Equation (4.31) and Equation (4.32) Slutsky's Theorem gives

$$\frac{\sqrt{\theta(1-\theta)}}{\sqrt{\frac{X_1+X_2}{n_1+n_2}(1-\frac{X_1+X_2}{n_1+n_2})}}\frac{\frac{X_1}{n_1} - \frac{X_2}{n_2}}{\sqrt{\frac{\theta(1-\theta)}{\rho(1-\rho)}}} \xrightarrow{d} 1 \cdot N(0,1) = N(0,1). \tag{4.33}$$

Since

$$\frac{1}{\sqrt{N(\frac{1}{n_1}+\frac{1}{n_2})\rho(1-\rho)}} \to \frac{1}{\sqrt{(\frac{1}{\rho}+\frac{1}{1-\rho})\rho(1-\rho)}} = 1$$

as $n_1, n_2 \to \infty$, we have that

$$\frac{1}{\sqrt{N(\frac{1}{n_1}+\frac{1}{n_2})\rho(1-\rho)}} \xrightarrow{p} 1. \tag{4.34}$$

Due to Equation (4.33) and Equation (4.34) Slutsky's Theorem gives us

$$Z_p(X_1, X_2) \xrightarrow{d} 1 \cdot N(0,1) = N(0,1),$$

This means

$$Z_p^2(X_1, X_2) \xrightarrow{d} \chi_1^2,$$

which is the desired result.

4.3.4 The Pearson chi-squared statistic

It is also possible to calculate the expected cell frequencies in the 2×2 table under the null hypothesis H_0 in Equation (4.1) and calculate the difference between the cell counts and the expected values. The larger any of the four differences is, the stronger the evidence against the null hypothesis. This is the logical reasoning behind the Pearson chi-square statistic, which is given by

$$\chi_2(X_1, X_2) = \sum_{i=1}^{2} \frac{(X_i - n_i\theta_i)^2}{n_i\theta_i} + \frac{(n_i - X_i - n_i(1-\theta_i))^2}{n_i(1-\theta_i)},$$

see for example Kateri (2014, p. 11, 24). However, since θ_1 and θ_2 are unknown, we replace them with the consistent maximum likelihood estimators given in Equation (4.13) and Equation (4.14) respectively, so that we get

$$\chi_2(X_1, X_2) = \sum_{i=1}^{2} \frac{(X_i - n_i\hat{\theta}_i)^2}{n_i\hat{\theta}_i} + \frac{(n_i - X_i - n_i(1-\hat{\theta}_i))^2}{n_i(1-\hat{\theta}_i)}. \tag{4.35}$$

Asymptotically (Kateri 2014, p. 11)

$$\chi_2(X_1, X_2) \xrightarrow{d} \chi_1^2,$$

i.e the asymptotic distribution of the Pearson chi squared statistic is a chi squared random variable with one degree of freedom.

Equality of the statistics $Z_p^2(X_1, X_2)$ and $\chi_2(X_1, X_2)$

In this section we show that $\chi_2(X_1, X_2)$ is the same as $Z_p^2(X_1, X_2)$. Firstly we consider each of the four terms in Equation (4.35) in turn. By expanding the first term we get

$$\frac{(x_1 - n_1\frac{(x_1+x_2)}{n_1+n_2})^2}{n_1\frac{(x_1+x_2)}{n_1+n_2}} = \frac{\frac{(x_1(n_1+n_2)-n_1x_1-n_1x_2)^2}{(n_1+n_2)^2}}{n_1\frac{(x_1+x_2)}{n_1+n_2}} = \frac{(x_1n_2 - n_1x_2)^2}{n_1(x_1+x_2)(n_1+n_2)} \tag{4.36}$$

And by similar expansions of the third we get

$$\frac{(x_2 - n_2\frac{(x_1+x_2)}{n_1+n_2})^2}{n_2\frac{(x_1+x_2)}{n_1+n_2}} = \frac{\frac{(x_2(n_1+n_2)-n_2x_1-n_2x_2)^2}{(n_1+n_2)^2}}{n_2\frac{(x_1+x_2)}{n_1+n_2}} = \frac{(x_2n_1 - n_2x_1)^2}{n_2(x_1+x_2)(n_1+n_2)} \tag{4.37}$$

We then consider the second term

$$\begin{aligned}\frac{(n_1 - x_1 - n_1\frac{n_1+n_2-(x_1+x_2)}{n_1+n_2})^2}{n_1\frac{n_1+n_2-(x_1+x_2)}{n_1+n_2}} &= \frac{(n_1-x_1)(n_1+n_2) - n_1^2 - n_1n_2 + n_1(x_1+x_2)}{n_1(n_1+n_2-(x_1+x_2))(n_1+n_2)} \\ &= \frac{(n_1^2 + n_1n_2 - n_1x_1 - n_2x_1 - n_1^2 - n_1n_2 + n_1x_1 + n_1x_2)^2}{n_1(n_1+n_2-(x_1+x_2))(n_1+n_2)} \\ &= \frac{(n_1x_2 - n_2x_1)^2}{n_1(n_1+n_2-(x_1+x_2))(n_1+n_2)}\end{aligned} \tag{4.38}$$

and finally the last term

$$\begin{aligned}\frac{(n_2 - x_2 - n_2\frac{n_1+n_2-(x_1+x_2)}{n_1+n_2})^2}{n_2\frac{n_1+n_2-(x_1+x_2)}{n_1+n_2}} &= \frac{(n_2-x_2)(n_1+n_2) - n_1n_2 - n_2^2 + n_2(x_1+x_2))}{n_2(n_1+n_2-(x_1+x_2))(n_1+n_2)} \\ &= \frac{(n_1n_2 + n_2^2 - n_1x_2 - n_2x_2 - n_1n_2 - n_2^2 + n_2x_1 + n_2x_2)^2}{n_2(n_1+n_2-(x_1+x_2))(n_1+n_2)} \\ &= \frac{(n_2x_1 - n_1x_2)^2}{n_2(n_1+n_2-(x_1+x_2))(n_1+n_2)}\end{aligned} \tag{4.39}$$

We see that the terms in Equation (4.36) to Equation (4.39) have the factors $(x_1n_2 - n_1x_2)^2$ and $\frac{1}{n_1+n_2}$ in common. When we insert for Equation (4.36) to

Equation (4.39) in Equation (4.35) we therefore get

$$\begin{aligned}\chi^2(x_1,x_2) =& \frac{(x_1n_2-n_1x_2)^2}{n_1+n_2}\Big(\frac{1}{n_1(x_1+x_2)}\frac{1}{n_2(x_1+x_2)}+\frac{1}{n_1(n_1+n_2-(x_1+x_2))}+\\ &\frac{1}{n_2(n_1+n_2-(x_1+x_2))}\Big)=\frac{(x_1n_2-n_1x_2)^2}{n_1+n_2}\Big(\frac{n_2(n_1+n_2-(x_1+x_2))}{n_1n_2(x_1+x_2)(n_1+n_2-(x_1+x_2))}+\\ &\frac{n_1(n_1+n_2-(x_1+x_2))+n_2(x_1+x_2)+n_1(x_1+x_2)}{n_1n_2(x_1+x_2)(n_1+n_2-(x_1+x_2))}\Big)\\ &=\frac{(x_1n_2-n_1x_2)^2}{n_1+n_2}\frac{(n_1+n_2)(n_1+n_2)}{n_1n_2(x_1+x_2)(n_1+n_2-(x_1+x_2))}\\ &=\frac{n_1^2n_2^2(\frac{x_1}{n_1}-\frac{x_2}{n_2})^2(n_1+n_2)}{n_1n_2(x_1+x_2)(n_1+n_2-(x_1+x_2))}=\frac{n_1n_2(\frac{x_1}{n_1}-\frac{x_2}{n_2})^2(n_1+n_2)}{(x_1+x_2)(n_1+n_2-(x_1+x_2))}\\ &=\frac{(\frac{x_1}{n_1}-\frac{x_2}{n_2})^2}{\frac{x_1+x_2}{n_1+n_2}(\frac{n_1+n_2}{n_1n_2})(1-\frac{x_1+x_2}{n_1+n_2})}=\frac{(\frac{x_1}{n_1}-\frac{x_2}{n_2})^2}{\frac{x_1+x_2}{n_1+n_2}(\frac{1}{n_2}+\frac{1}{n_1})(1-\frac{x_1+x_2}{n_1+n_2})}\\ &=Z_p^2(x_1,x_2),\end{aligned}$$

which is what we wanted to show.

4.4 Calculations of the different p values introduced in Section 4.1.2 and evaluation of power functions of tests based on the p-values.

We illustrate how to calculate the realisations of the A, E, C and M p-value, introduced in Section 4.1.2, using Z_p^2 as test statistic and when $n_1 = 5$, $n_2 = 5, x_1 = 5, x_2 = 2$. We use 0.05 as significance level, i.e we reject when $p(\boldsymbol{X}) \leq \alpha$. We could of course have used the severe near-sightedness example, but n_1 and n_2 is much larger in that example, so that the example we study in this section is more suitable when we want to illustrate how to calculate the realisations of the p-values and also illustrate how to calculate the power functions.

4.4.1 Calculation of p-values

The A-method The value of the A p-value is from Equation (4.5) given by

$$p_A(x) = 1 - F\Big(\frac{(\frac{x_1}{n_1}-\frac{x_2}{n_2})^2}{\frac{x_1+x_2}{n_1+n_2}(1-\frac{x_1+x_2}{n_1+n_2})(\frac{1}{n_1}+\frac{1}{n_2})}\Big) = 1 - F(4.2857) = 0.03843,$$

where F is the cumulative distribution function of the chi square distribution with one degree of freedom.

The E-method The maximum likelihood estimate of θ is from Equation (4.12) $\hat{\theta} = \frac{5+2}{10} = 0.7$. We use the enumeration procedure, given in Section 3.2, to calculate the realisation of the E p-value. The first steps are illustrated in Table 4.5. The next step is to calculate the probabilities of the outcomes marked with $\times$ and sum them to get $p_E(5,2)$. We therefore have by Equation (4.3)

$$\begin{aligned} p_E(\boldsymbol{x}) = \sum_{T(\boldsymbol{y}) \geq T(\boldsymbol{x})} \Pr_{\hat{\theta}}(\boldsymbol{X} = \boldsymbol{y}) &= \Pr_{0.7}(\boldsymbol{X} = (3,0)) + \Pr_{0.7}(\boldsymbol{X} = (4,0)) \\ &+ \Pr_{0.7}(\boldsymbol{X} = (5,0)) + \Pr_{0.7}(\boldsymbol{X} = (5,1)) + \Pr_{0.7}(\boldsymbol{X} = (5,2)) \\ &+ \Pr_{0.7}(\boldsymbol{X} = (0,3)) + \Pr_{0.7}(\boldsymbol{X} = (0,4)) + \Pr_{0.7}(\boldsymbol{X} = (0,5)) \\ &+ \Pr_{0.7}(\boldsymbol{X} = (1,5)) + \Pr_{0.7}(\boldsymbol{X} = (2,5)) \\ &= 0.0581. \end{aligned}$$

The M method When calculating $p_M(5,2)$ we consider the same sum of probabilities as when calculating $p_E(5,2)$. However, we use the θ that maximises this sum and not $\hat{\theta}$. We numerically find the maximum by evaluating the sum of probabilities in a equispaced grid from 0 to 1 and choose the value of θ that gives the maximum value among the values we have calculated. We discuss how to numerically find the maximum in Section 4.10.5. The procedure is illustrated in Figure 4.2, where we have used a grid consisting of 100 points. The plot of $\Pr(T(\boldsymbol{X}) \geq T(5,2))$ as a function of θ is called the *profile* of $T(5,2)$ by Lloyd (2008). We see that the profile is symmetric around $\theta = \frac{1}{2}$ and that the maximum occurs around 0.38 and 0.62. We then have from Equation (4.2)

$$\begin{aligned} p_M(5,2) = \sup_{\theta \in \Theta_0} \Pr(T(\boldsymbol{X}) \geq T(\boldsymbol{x})) = \sup_{\theta \in \Theta_0} &(\Pr_\theta(\boldsymbol{X} = (3,0)) + \Pr_\theta(\boldsymbol{X} = (4,0)) \\ &+ \Pr_\theta(\boldsymbol{X} = (5,0)) + \Pr_\theta(\boldsymbol{X} = (5,1)) + \Pr_\theta(\boldsymbol{X} = (5,2)) \\ &+ \Pr_\theta(\boldsymbol{X} = (0,3)) + \Pr_\theta(\boldsymbol{X} = (0,4)) + \Pr_\theta(\boldsymbol{X} = (0,5)) \\ &+ \Pr_\theta(\boldsymbol{X} = (1,5)) + \Pr_\theta(\boldsymbol{X} = (2,5))) \\ &= 0.0618. \end{aligned}$$

The C method When calculating $p_C(5,2)$ we also use the enumeration procedure, however, we only look at outcomes $\boldsymbol{y}$ that give the value $S(5,2) = 5+2 = 7$ of the sufficient statistic. The first steps of the enumeration procedure are shown

in Table 4.6. The next step in the enumeration procedure is to calculate the probabilities of the outcomes marked with $\star$, which are given by Equation (4.10) where $s = 7$. By summing these probabilities we get $p_C(5, 2)$. So from Equation (4.4) we have

$$\begin{aligned} p_C(5,2) &= \Pr(T(\boldsymbol{X}) \geq T(\boldsymbol{x}) \mid S(\boldsymbol{X}) = 7) = \sum_{T(\boldsymbol{y}) \geq T(\boldsymbol{x})} \Pr(\boldsymbol{X} = \boldsymbol{y} \mid S(\boldsymbol{X}) = 7) \\ &= \Pr(\boldsymbol{X} = (2,5) \mid S(\boldsymbol{X}) = 7) + \Pr(\boldsymbol{X} = (5,2) \mid S(\boldsymbol{X}) = 7) = 0.1667. \end{aligned}$$

Results of tests We reject the null hypothesis at the 0.05 significance level by using the A p-value but do not reject the null hypothesis if we use any of the remaining p-values. A natural question is then if the A p-value, which is only asymptotically valid, is valid when n_1 an n_2 are as low as 5, i.e is it safe to use asymptotic theory in our case? The realisation of the E p-value is the lowest one of the three p-values above 0.05. Since this p-value is also only asymptotically valid, it is also reasonable to question the validity of the p-value in our case. To answer these questions we need to look at the power functions of the level α-tests based on the p-values.

4.4.2 Evaluating the power functions

In this section we show how to calculate the power functions of the level 0.05 tests based on the p-values in Section 4.4.1. From Section 3.3 we know we first must calculate all the values of the p-value when we want to evaluate the power function. The realisations can be calculated in the same way we calculated the single realisations of the p-values in Section 4.4.1. After calculating all the values of each p-value, we consider for each p-value the realisations that are equal to or below $\alpha = 0.05$. In general, we denote by $\gamma_i(\theta_1, \theta_2; \alpha)$ the power function of the level α test based on the i p-value evaluated in (θ_1, θ_2), where i=M,E,C, A.

In Table 4.7 we show the values of the p-values C, M and E that are equal to or below 0.05 and in Table 4.8 we show the values of the A p-value that are below or equal to 0.05. We observe that the same outcomes have E, C and M p-value below or equal to 0.05. This means $\gamma_C(\theta_1, \theta_2; 0.05) = \gamma_E(\theta_1, \theta_2; 0.05) = \gamma_M(\theta_1\theta_2; 0.05)$ for all $(\theta_1, \theta_2) \in \Theta$. We also observe that the set $\{\boldsymbol{x} \mid p_C(\boldsymbol{x}) \leq 0.05\}$ is a subset of the set $\{\boldsymbol{x} \mid p_A(\boldsymbol{x}) \leq 0.05)\}$, which means $\gamma_A(\theta_1, \theta_2; 0.05)$ is least as great as either of the other considered power functions for all $(\theta_1, \theta_2) \in \Theta$. We initially evaluate the power functions in an equispaced grid on $[0, 1] \times [0, 1]$ with the same grid increment in both directions, where for instance the grid points are $0, 0.01, 0.02, 0.03, ..., 0.99, 1$ along both the θ_1- and the θ_2-axis.

Table 4.5: The first steps in the enumeration procedure when calculating $p_E(5,2)$. The possible outcomes are given in column 1 and 2 from the left. The values of the test statistic are given in column 3. Outcomes for which the test statistic is least as large as $Z_p^2(5,2)$ are marked with $\times$ in column 4.

y_1	y_2	$Z_p^2(y_1,y_2)$	$Z_p^2(y_1,y_2) \geq Z_p^2(5,2)$
0	0	−99.0000	
1	0	1.1111	
2	0	2.5000	
3	0	4.2857	×
4	0	6.6667	×
5	0	10.0000	×
0	1	1.1111	
1	1	0.0000	
2	1	0.4762	
3	1	1.6667	
4	1	3.6000	
5	1	6.6667	×
0	2	2.5000	
1	2	0.4762	
2	2	0.0000	
3	2	0.4000	
4	2	1.6667	
5	2	4.2857	×
0	3	4.2857	×
1	3	1.6667	
2	3	0.4000	
3	3	0.0000	
4	3	0.4762	
5	3	2.5000	
0	4	6.6667	×
1	4	3.6000	
2	4	1.6667	
3	4	0.4762	
4	4	0.0000	
5	4	1.1111	
0	5	10.0000	×
1	5	6.6667	×
2	5	4.2857	×
3	5	2.5000	
4	5	1.1111	
5	5	−99.0000	

Table 4.6: The first steps in the enumeration procedure when calculating $p_C(5,2)$. The possible outcomes are given in column 1 and 2 from the left. The values of $Z_p^2(y_1,y_2)$ are given in column 3. Outcomes $\boldsymbol{y}$ for which the sufficient statistic equals 7, i.e $y_1+y_2=7$, are marked with $\times$ in column 4 and the outcomes among these where $Z_p^2(y_1,y_2) \geq Z_p^2(5,2)$ are marked $\star$ in the last column.

y_1	y_2	$Z_p^2(y_1,y_2)$	$y_1+y_2=7$	$(Z_p^2(y_1,y_2) \geq Z_p^2(5,2)) \wedge y_1+y_2=7$
0	0	−99.0000		
1	0	1.1111		
2	0	2.5000		
3	0	4.2857		
4	0	6.6667		
5	0	10.0000		
0	1	1.1111		
1	1	0.0000		
2	1	0.4762		
3	1	1.6667		
4	1	3.6000		
5	1	6.6667		
0	2	2.5000		
1	2	0.4762		
2	2	0.0000		
3	2	0.4000		
4	2	1.6667		
5	2	4.2857	×	⋆
0	3	4.2857		
1	3	1.6667		
2	3	0.4000		
3	3	0.0000		
4	3	0.4762	×	
5	3	2.5000		
0	4	6.6667		
1	4	3.6000		
2	4	1.6667		
3	4	0.4762	×	
4	4	0.0000		
5	4	1.1111		
0	5	10.0000		
1	5	6.6667		
2	5	4.2857	×	⋆
3	5	2.5000		
4	5	1.1111		
5	5	−99.0000		

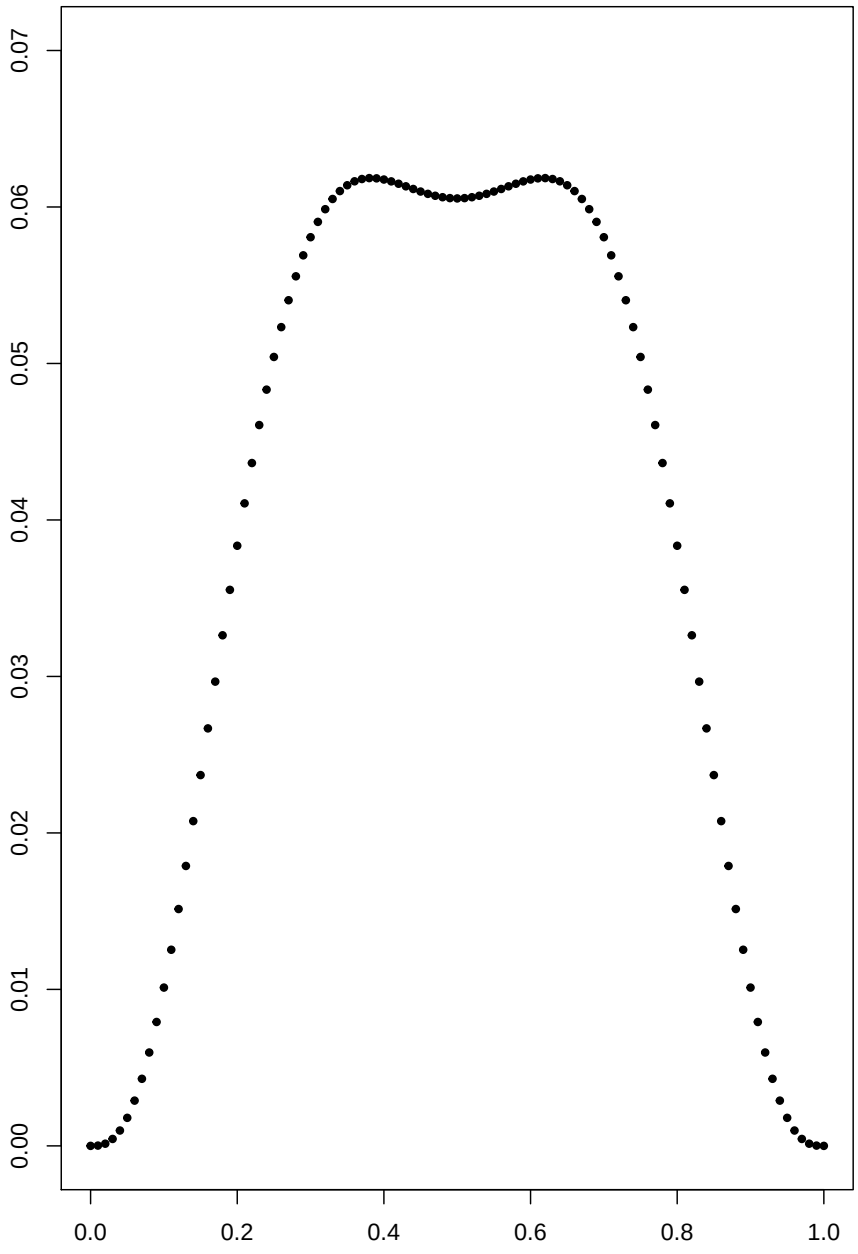


Figure 4.2: Plot of the profile $Z_p^2(5,2)$ when evalutated in the grid of θ-values $0, 01, 0.02, \cdots, 0.99, 1$ and $n_1 = n_2 = 5$.

Table 4.7: The different realisations of the C, E and M p-values below 0.05 when $n_1 = 5, n_2 = 5$. The outcomes in the experiment are given in the first two columns and respectively the C, E and M p-values are given in the third, fourth and fifth column.

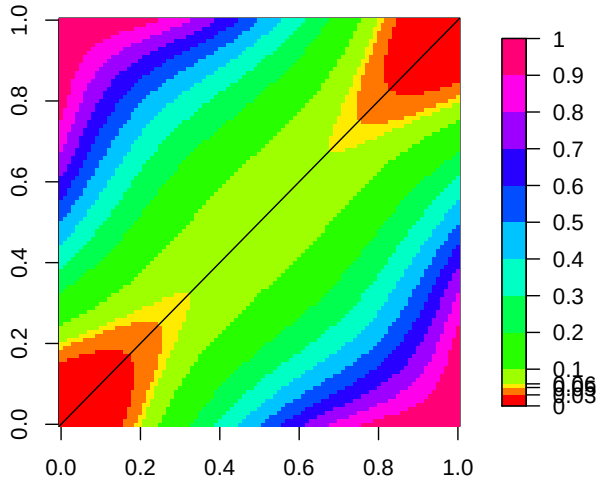
x_1	x_2	$p_C(x_1, x_2)$	$p_E(x_1, x_2)$	$p_M(x_1, x_2)$
4	0	0.04762	0.01884	0.02148
5	0	0.00794	0.00195	0.00195
5	1	0.04762	0.01884	0.02148
0	4	0.04762	0.01884	0.02148
0	5	0.00794	0.00195	0.00195
1	5	0.04762	0.01884	0.02148

Table 4.8: The values of the A p-value that are below 0.05 when $n_1 = 5, n_2 = 5$. The outcomes are given in the first two columns and the A p-value is in the third coloumn.

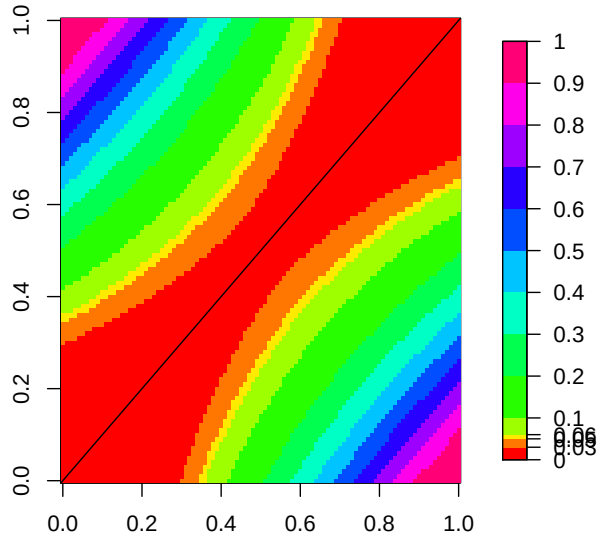
x_1	x_2	$p_A(x_1, x_2)$
3	0	0.03843
4	0	0.00982
5	0	0.00157
5	1	0.00982
5	2	0.03843
0	3	0.03843
0	4	0.00982
0	5	0.00157
1	5	0.00982
2	5	0.03843

By evaluating the power functions in the grid we get the two panels in Figure 4.3. In both figures we observe that the power increases as we approach either of the corners $(1,0)$ or $(0,1)$, which makes intuitively sense as it should be easiest to tell the success probabilities 1 and 0 from each other. We also observe that the power function $\gamma_A(\theta_1,\theta_2;0.05)$ takes at least as high values as the other power functions, as explained earlier. The type I error probabilities of each test for the different values of $\theta=\theta_1=\theta_2$ are given on the diagonal from $(0,0)$ to $(1,1)$ on each plot. We observe that the size of the level 0.05 test based on the A value is higher than 0.05, which means $p_A(x)$ is not valid. We also see that the size of the other tests are below the 0.05 level. In Figure 4.4 we have plotted the type I error probabilities, i.e $\Pr_\theta(p_i(\boldsymbol{X})\leq 0.05)$ where i is either A,E,C or M, for $\theta\in[0,1]$. It looks like $\Pr_\theta(p_i(\boldsymbol{X})\leq 0.5)$ is symmetric around $\theta=\frac{1}{2}$. We observe that the size of the tests based on either the E, C or M p-value is about 0.021, which is much lower than the nominal level, and that the test size of the test based on the A p-value is about 0.061. A test where the supremum of the type I error probabilities is below the test level is said to be *conservative* and when it is higher it is said to be *anti-conservative* or *liberal* (Krishnamoorthy 2015, p. 20). Therefore the level 0.05 tests based on the E, C and M p-values are conservative and the level 0.05 test based on the A p-value is liberal. In Section 4.6 we prove that the level α-tests based on the C or M p-value are guaranteed to not be liberal. And, we show that the p-values A and E are only asymptotically valid (as the sample sizes approach infinity), which means there is no guarantee that they are valid for finite sample sizes and therefore no guarantee that the level α tests based on these p-values not are liberal. For instance in Mehrotra et al. (2004) there are cases where the A method gives a liberal test procedure and in Krishnamoorthy & Thomson (2004) there are cases where the test procedure based on the E method is liberal.

By evaluating the power functions of the tests in the same grid while varying n_1 and n_2 but keeping α fixed at 0.05, it seems like the value of each of the power functions in the point $(\tilde{\theta}_1,\tilde{\theta}_2)$ is the same as the power function evaluated in the point resulting from reflecting the original point$(\tilde{\theta}_1,\tilde{\theta}_2)$ in $(\frac{1}{2},\frac{1}{2})$. This reflection operation corresponds to rotating the point $(\tilde{\theta}_1,\tilde{\theta}_2)$ 180 degrees around $(\frac{1}{2},\frac{1}{2})$. In Appendix B we show that the new coordinates is $(1-\tilde{\theta}_1,1-\tilde{\theta}_2)$. When we perform this rotation operation for all points, we can imagine that we have rotated the solid triangle with corners $(0,0),(1,0)$ and $(1,1)$, except from the line from $(0,0)$ to $(1,1)$, 180 degrees around $(\frac{1}{2},\frac{1}{2})$ and that the line from $(0,0)$ to $(\frac{1}{2},\frac{1}{2})$ has been rotated 180 degrees around $(\frac{1}{2},\frac{1}{2})$. This description of the rotation procedure describes the symmetry in the plots of the power functions. We have tried to illustrate the symmetry when $n_1=3,n_2=25$ and $n_1=30,n_2=2$, see respectively Figure 4.5 and Figure 4.6. In Section 4.5 we show that this is true in general for the power functions of the level α tests based on the considered p-values.



(a) Plot of the power function $\gamma_A(\theta_1, \theta_2; 0.05)$



(b) Plot of the power function $\gamma_i(\theta_1, \theta_2; 0.05)$ where i is either C, E or M.

Figure 4.3: Plots of the power functions $\gamma_A(\theta_1, \theta_2; 0.05)$ and $\gamma_i(\theta_1, \theta_2; 0.05)$ when $n_1 = n_2 = 5$ and i is either E, C or M. We evaluate the functions in an equispaced grid with the same grid increment in both directions and the θ_1-axis is along the abscissa and the θ_2-axis is along the ordinate. For instance the grid points along the θ_1-axis are $0, 0.01, 0.02, \cdots, 0.99, 1$. The line $\theta_1 = \theta_2$ is also drawn. To the right of each plot there is a legend that specifies the different colors for the different intervals of values used in making the plots. The four lowest cut-points are $0, 0.03, 0.05, 0.06$.

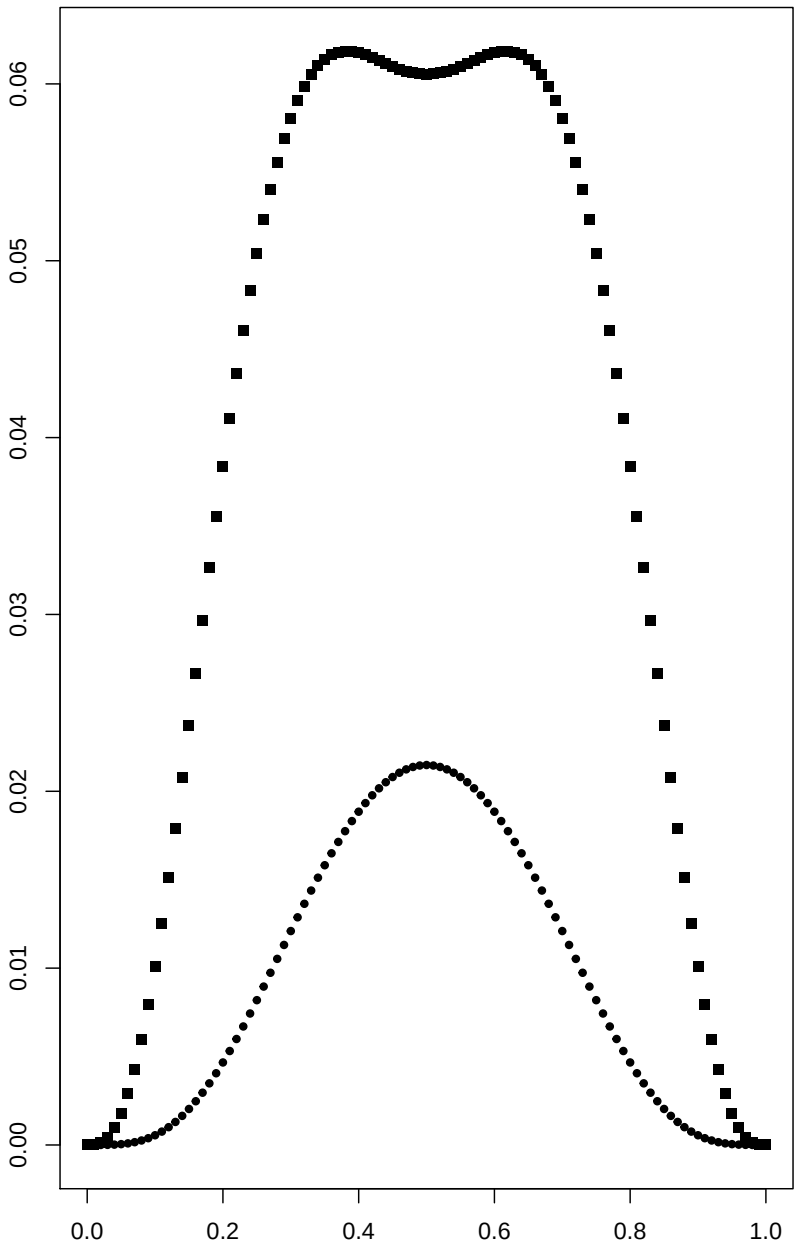
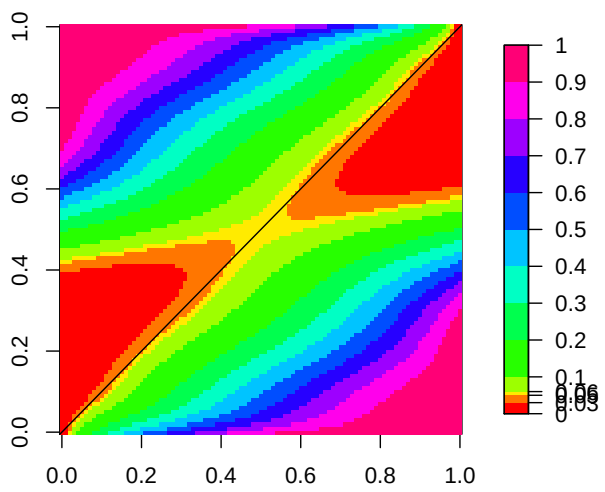
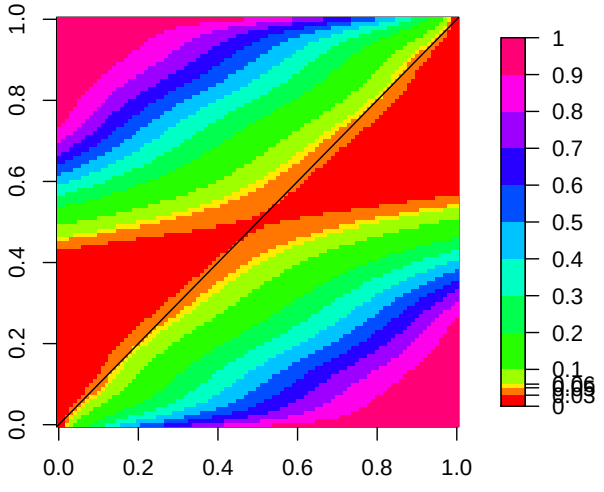


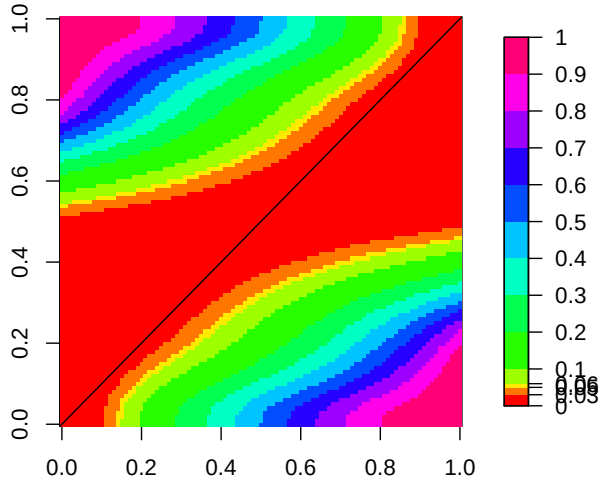
Figure 4.4: Plots of $\Pr_\theta(p_i(X) \leq 0.05)$ as a function of θ for i equal to M, A, C, E. We evalute θ in the equispaced grid $0, 0.01, \ldots, 0.99, 1$. The plot of the type I error probabilities are the same for the tests based on the M, C and E p-values and we have used a filled circle as plotting symbol. We have used a filled rectangle as plotting symbol when plotting the values of $\Pr_\theta(p_A(X) \leq 0.05)$.



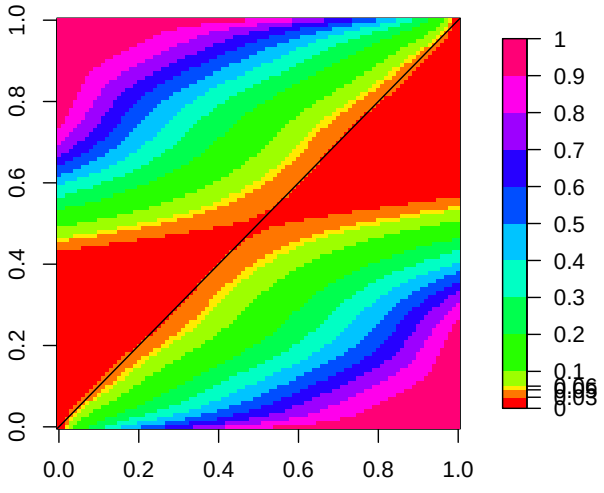
(a) Plot of the power function $\gamma_A(\theta_1, \theta_2; 0.05)$



(b) Plot of the power function $\gamma_E(\theta_1, \theta_2; 0.05)$

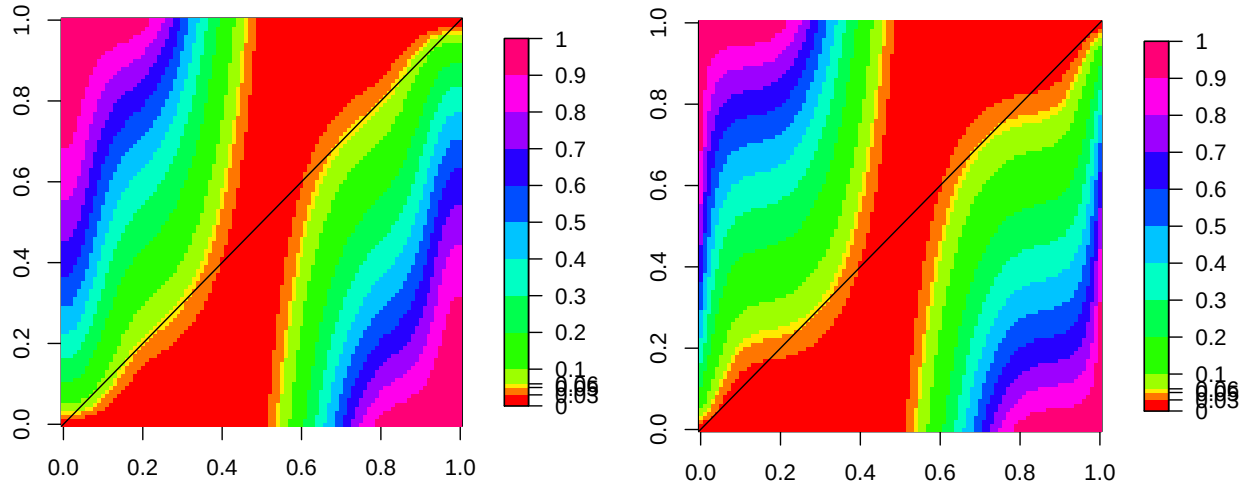


(c) Plot of the power function $\gamma_c(\theta_1, \theta_2; 0.05)$.



(d) Plot of the power function $\gamma_M(\theta_1, \theta_2; 0.05)$

Figure 4.5: Plots of the power functions $\gamma_i(\theta_1, \theta_2; 0.05)$ where i is either E, C, A or M and $n_1 = 3, n_2 = 25$. The θ_1-axis is along the abscissa and the θ_2-axis is along the ordinate. The line $\theta_1 = \theta_2$ is also drawn. To the right of each plot there is a legend that specifies the different colors for the different intervals of values used in making the plots. The four lowest cut-points are $0, 0.03, 0.05, 0.06$.



(a) Plot of the power function $\gamma_A(\theta_1, \theta_2; 0.05)$

(b) Plot of the power function $\gamma_E(\theta_1, \theta_2; 0.05)$

(c) Plot of the power function $\gamma_C(\theta_1, \theta_2; 0.05)$.

(d) Plot of the power function $\gamma_M(\theta_1, \theta_2; 0.05)$

Figure 4.6: Plots of the two power functions $\gamma_i(\theta_1, \theta_2; 0.05)$ where i is either E, C, A or M and $n_1 = 30, n_2 = 2$. The θ_1-axis is along the abscissa and the θ_2-axis is along the ordinate. The line $\theta_1 = \theta_2$ is also drawn. To the right of each plot there is a legend that specifies the different colors for the different intervals of values used in making the plots. The four lowest cut-points are $0, 0.03, 0.05, 0.06$.

4.5 Symmetry of the power functions

We want to show

$$\Pr_{(\theta_1,\theta_2)}(p_i(\boldsymbol{X}) \leq \alpha) = \Pr_{(1-\theta_1,1-\theta_2)}(p_i(\boldsymbol{X}) \leq \alpha)$$

holds when $i = A, C, E, M$. For convenience, we drop the subscript i and all statements made without the subscript are valid for all the mentioned p-values.

We consider an outcome $\boldsymbol{y}$ such that $p(\boldsymbol{y}) \leq \alpha$, which means we consider a realisation of the p-value that gives a that contributes to the power function $\Pr_{(\theta_1,\theta_2)}(p(\mathbf{X}) \leq \alpha)$. The contribution at (θ_1, θ_2) is from Equation (4.7)

$$\Pr_\theta(\boldsymbol{X} = \mathbf{y}) = \binom{n_1}{y_1}\binom{n_2}{y_2}{\theta_1}^{y_1}{\theta_2}^{y_2}(1-\theta_1)^{n_1-y_1}(1-\theta_2)^{n_2-y_2},$$

while the contribution to the power at $(1-\theta_1, 1-\theta_2)$ is

$$\Pr_\theta(\boldsymbol{X} = \mathbf{y}) = \binom{n_1}{y_1}\binom{n_2}{y_2}(1-\theta_1)^{y_1}(1-\theta_2)^{y_2}\theta_1^{n_1-y_1}\theta_2^{n_2-y_2},$$

so that the contributions of $p(\mathbf{y})$ to $\Pr_\theta(p(\boldsymbol{X}) \leq \alpha)$ at $\boldsymbol{\theta} = (\theta_1, \theta_2)$ and $\boldsymbol{\theta} = (1-\theta_1, 1-\theta_2)$ are not the same. However, if also $p(n_1 - y_1, n_2 - y_2) \leq \alpha$ then the contribution to the power function at (θ_1, θ_2) from $p(n_1 - y_1, n_2 - y_2)$ is

$$\begin{aligned}
\Pr_{(\theta_1,\theta_2)}(\boldsymbol{X} = (n_1 - y_1, n_2 - y_2)) &= \binom{n_1}{n_1 - y_1}\binom{n_2}{n_2 - y_2}\theta_1^{n_1-y_1}\theta_2^{n_2-y_2}(1-\theta_1)^{y_1}(1-\theta_2)^{y_2} \\
&= \binom{n_1}{y_1}\binom{n_2}{y_2}(1-\theta_1)^{y_1}(1-\theta_2)^{y_2}\theta_1^{n_1-y_1}\theta_2^{n_2-y_2} \\
&= \Pr_{(1-\theta_1,1-\theta_2)}(\boldsymbol{X} = \boldsymbol{y}),
\end{aligned}$$

since $\binom{n_1}{n_1-y_1} = \binom{n_1}{y_1}$ and $\binom{n_2}{n_2-y_2} = \binom{n_2}{y_2}$. It therefore follows that the contribution to the power at $(1-\theta_1, 1-\theta_2)$ from $p(n_1 - y_1, n_2 - y_2)$ is

$$\Pr_{(1-\theta_1,1-\theta_2)}(\boldsymbol{X} = (n_1 - y_2, n_2 - y_2)) = \Pr_{(\theta_1,\theta_2)}(\boldsymbol{X} = (y_1, y_2)).$$

So if $p(y_1, y_2) \leq \alpha$ if and only if $p(n_1 - y_1, n_2 - y_2) \leq \alpha$ then the two realisations of the p-value jointly give the same contributions to the power at $\boldsymbol{\theta} = (\theta_1, \theta_2)$ and $\boldsymbol{\theta} = (1-\theta_1, 1-\theta_2)$. The next step is to verify that $p(y_1, y_2) \leq \alpha$ if and only if $p(n_1 - y_1, n_2 - y_2) \leq \alpha$ holds for the different p-values. If this holds then the set $\{\boldsymbol{x} \mid p(\mathbf{x} \leq \alpha\}$ consist of pairs of outcomes that give the same contribution to the power function at (θ_1, θ_2) and at $(1-\theta_1, 1-\theta_2)$, so that in total the power

function is the same when evaluated in the two points. The first key observation when demonstrating that $p(y_1, y_2) \leq \alpha$ if and only if $p(n_1 - y_1, n_2 - y_2) \leq \alpha$ holds is

$$\begin{aligned} T(n_1 - x_2, n_2 - x_2) &= \frac{(\frac{n_1 - x_1}{n_1} - \frac{n_2 - x_2}{n_2})^2}{\frac{n_1 - x_1 + n_2 - x_2}{n_1 + n_2}(1 - \frac{n_1 - x_1 + n_2 - x_2}{n_1 + n_2})(\frac{1}{n_1} + \frac{1}{n_2})} \\ &= \frac{(1 - \frac{x_1}{n_1} - 1 + \frac{x_2}{n_2})^2}{(1 - \frac{x_1 + x_2}{n_1 + n_2})(1 - 1 + \frac{x_1 + x_2}{n_1 + n_2})(\frac{1}{n_1} + \frac{1}{n_2})} \\ &= \frac{(\frac{x_1}{n_1} - \frac{x_2}{n_2})^2}{\frac{x_1 + x_2}{n_1 + n_2}(1 - \frac{x_1 + x_2}{n_1 + n_2})(\frac{1}{n_1} + \frac{1}{n_2})} \\ &= T(x_1, x_2) \end{aligned} \tag{4.40}$$

We next consider the different p-values in turn and calculate $p(y_1, y_2)$ and $p(n_1 - y_1, n_2 - y_2)$.

A-method From Equation (4.5) we do the computations

$$p_A(y_1, y_2) = \int_{T(y_1, y_2)}^{\infty} f(t)\mathrm{d}t$$

and

$$p_A(n_1 - y_1, n_2 - y_2) = \int_{T(n_1 - y_1, n_2 - y_2)}^{\infty} f(t)\mathrm{d}t$$

Due to Equation (4.40) we have that $p_A(y_1, y_2) = p_A(n_1 - y_2, n_2 - y_2)$.

M-method By Equation (4.2)

$$p_M(y_1, y_2) = \sup_{\theta \in [0,1]} \sum_{T(\boldsymbol{x}) \geq T(y_1, y_2)} \Pr_{(\theta,\theta)}(\boldsymbol{X} = \boldsymbol{x}) \tag{4.41}$$

and

$$p_M(n_1 - y_1, n_2 - y_2) = \sup_{\theta \in [0,1]} \sum_{T(\boldsymbol{x}) \geq T(n_1 - y_1, n_2 - y_2)} \Pr_{(\theta,\theta)}(\boldsymbol{X} = \boldsymbol{x}) \tag{4.42}$$

Since $T(y_1, y_2) = T(n_1 - y_1, n_2 - y_2)$ due to Equation (4.40), we see that we maximise the same sum in Equation (4.41) as in Equation (4.42). Therefore $p_M(y_1, y_2) = p_M(n_1 - y_1, n_2 - y_2)$

E-method By Equation (4.3) we do the computations

$$p_E(y_1, y_2) = \sum_{T(\boldsymbol{x}) \geq T(y_1, y_2)} \Pr_{\hat{\boldsymbol{\theta}}_A}(\boldsymbol{X} = \boldsymbol{x}) \tag{4.43}$$

and

$$p_E(n_1 - y_2, n_2 - y_2) \sum_{T(\boldsymbol{x}) \geq T(n_1 - y_1, n_2 - y_2)} \Pr_{\hat{\boldsymbol{\theta}}_B}(\boldsymbol{X} = \boldsymbol{x}), \tag{4.44}$$

where $\hat{\boldsymbol{\theta}}_A = (\hat{\theta}_A, \hat{\theta}_A)$, $\hat{\boldsymbol{\theta}}_B = (\hat{\theta}_B, \hat{\theta}_B)$, $\hat{\theta}_A$ is maximum likelihood estimate of θ based on (y_1, y_2) and $\hat{\theta}_B$ is the maximum likelihood estimate of θ based on $(n_1 - y_1, n_2 - y_2)$. Due to Equation (4.40) we sum over the same $\boldsymbol{x}$ in Equation (4.43) as in Equation (4.44). From Equation (4.12) we know that

$$\hat{\theta}_A = \frac{y_1 + y_2}{n_1 + n_2}$$

and also

$$\hat{\theta}_B = \frac{n_1 - y_1 + n_2 - y_2}{n_1 + n_2} = 1 - \frac{y_1 + y_2}{n_1 + n_2} = 1 - \hat{\theta}_A$$

so that $\hat{\theta}_A$ is not equal to $\hat{\theta}_B$. We therefore use different values for θ when we compute the probabilities in the sums in Equation (4.43) and in Equation (4.44). It is therefore not obvious that $p_E(y_1, y_2) = p_E(n_1 - y_1, n_2 - y_2)$. We now show that this is still holds. We start by considering a point $\boldsymbol{x}$ which appears in the sum in Equation (4.43) and in Equation (4.44), i.e we consider a point $\boldsymbol{x} = (x_1, x_2)$ so that $T(\boldsymbol{x}) \geq T(y_1, y_2) = T(n_1 - y_1, n_2 - y_2)$. From Equation (4.40) we also have that the point $(x_1 - n_1, x_2 - n_2)$ is in the sum in Equation (4.43) and in Equation (4.44). From Equation (4.7) we have that

$$\begin{aligned} \Pr_{\hat{\boldsymbol{\theta}}_B}(\boldsymbol{X} = (x_1, x_2)) &= \hat{\theta}_B^{x_1 + x_2}(1 - \hat{\theta}_B)^{n_1 + n_2 - x_1 - x_2} \\ &= (1 - \hat{\theta}_A)^{x_1 + x_2} \hat{\theta}_A^{n_1 + n_2 - x_1 - x_2} \\ &= \Pr_{\hat{\boldsymbol{\theta}}_A}(\boldsymbol{X} = (n_1 - x_1, n_2 - x_2)), \end{aligned} \tag{4.45}$$

which also gives that

$$\Pr_{\hat{\boldsymbol{\theta}}_B}(\boldsymbol{X} = (n_1 - x_1, n_2 - x_2)) = \Pr_{\hat{\boldsymbol{\theta}}_A}(\boldsymbol{X} = (x_1, x_2)). \tag{4.46}$$

This means the joint contributions from (x_2, x_2) and $(n_1 - x_1, n_2 - x_2)$ are the same at $\hat{\theta}_A$ and $\hat{\theta}_B$. Since this hold for all points $\boldsymbol{x}$ so that $T(\boldsymbol{x}) \geq T(y_1, y_2) = T(n_1 - y_1, n_2 - y_2)$, we sum the same probabilities in Equation (4.43) as in Equation (4.44), which means that $p_E(y_1, y_2) = p_E(n_1 - y_1, n_2 - y_2)$.

C-method If we let $y_1 + y_2 = s$, then $n_1 - y_1 + n_2 - y_2 = n_1 + n_2 - (y_1 + y_2) = n_1 + n_2 - s$. When we calculate $p_C(y_1, y_2)$, we only look at tables $\boldsymbol{x}$ where $x_1 + x_2 = s$. Similarly, when we calculate $p_C(n_1 - y_1, n_2 - y_2)$ we only look at tables $\boldsymbol{x}$ such that $x_1 + x_2 = n_1 + n_2 - s$. For each table $\boldsymbol{x} = (x_1, x_2)$ we consider when we calculate $p_C(y_1, y_2)$ we consider the table $(n_1 - x_1, n_2 - x_2)$ when we calculate $p_C(n_1 - y_1, n_2 - y_2)$ since $n_1 - x_1 + n_2 - x_2 = n_1 + n_2 - s$. From Equation (4.40) we know that the tables $\boldsymbol{x} = (x_1, x_2)$ and $(n_1 - x_1, n_2 - x_2)$ give the same value of the test statistic. So if we order the tables $\boldsymbol{x}$ where $x_1 + x_2 = s$ in increasing value of test statistic and do the same for the tables $\boldsymbol{x}$ where $x_1 + x_2 = n_1 + n_2 - s$ and consider a table $\boldsymbol{z}$ such that $z_1 + z_2 = s$, then the order number of the the tables (z_1, z_2) and $(n_1 - z_1, n_2 - z_2)$ is the same. We have illustrated this in Table 4.9 when $n_1 = n_2 = 5$ and the outcome in the experiment is $(5, 2)$.

Table 4.9: The outcomes in the conditional experiments where $n_1 = 5$, $n_2 = 5$ and $x_1 + x_2 = 5 + 2$ or $x_1 + x_2 = n_1 + n_2 - 5 - 2 = 10 - 7 = 3$. The outcomes for the conditional experiment where $x_1 + x_2 = 7$ is given in the table on the left, while the outcomes in the other conditional experiment are given on the right. Both tables are ordered in increasing value of Z_p^2.

x_1	x_2	Z_p^2
4	3	0.4762
3	4	0.4762
5	2	4.2857
2	5	4.2857

x_1	x_2	Z_p^2
1	2	0.4762
2	1	0.4762
0	3	4.2857
3	0	4.2857

We know that $\binom{n_1}{n_1 - x_1} = \binom{n_1}{x_1}$ and that $\binom{n_2}{n_2 - x_2} = \binom{n_2}{x_2}$. Furthermore $\binom{n_1+n_2}{n_1+n_2-(x_1+x_2)} = \binom{n_1+n_2}{x_1+x_2}$. This gives

$$\begin{aligned}
\Pr(X_1 = n_1 - x_1 \mid X_1 + X_2 = n_1 + n_2 - (x_1 + x_2)) &= \frac{\binom{n_1}{n_1-x_1}\binom{n_2}{n_1+n_2-(x_1+x_2)-(n_1-x_1)}}{\binom{n_1+n_2}{n_1+n_2-(x_1+x_2)}} \\
&= \frac{\binom{n_1}{x_1}\binom{n_2}{x_2}}{\binom{n_1+n_2}{x_1+x_2}} \\
&= \Pr(X_1 = x_1 \mid X_1 + X_2 = x_1 + x_2),
\end{aligned}$$

which means the tables (x_1, x_2) and $(n_1 - x_1, n_2 - x_2)$ have the same conditional probability. So for each table we sum over when calculating $p_C(y_1, y_2)$ there is a table with the same conditional probability when calculating $p_C(n_1 - y_1, n_2 - y_2)$. This holds for all of the tables we consider when we calculate $p_C(n_1 - y_1, n_2 - y_2)$. In total, $p_C(y_1, y_2) = p_C(n_1 - y_1, n_2 - y_2)$.

We have therefore shown that $\Pr_{(\theta_1,\theta_2)}(p(\boldsymbol{X}) \leq \alpha) = \Pr_{(1-\theta_1,1-\theta_2)}(p(\boldsymbol{X}) \leq \alpha)$.

4.6 Proof of the different p-values being valid or asymptotically valid

We want to test

$$H_0 : \theta \in \Theta_0, H_1 : \boldsymbol{\theta} \in \Theta_1 \tag{4.47}$$

where Θ_0 is composite, $\boldsymbol{\theta}$ may be a vector of parameters and θ is the common value of the parameters in $\boldsymbol{\theta}$ under H_0. We use a test statistic $T(\boldsymbol{X})$ for which large values indicate that H_1 is correct and the larger the value the stronger the indication.

4.6.1 Proof of M p-value being valid

We want to show that $p_M(\boldsymbol{X})$ defined in Equation (4.2) is valid. We start by fixing $\theta_0 \in \Theta_0$. From Section 3.5 we know that

$$p^\star(\boldsymbol{x}) = \Pr_{\theta_0}(T(\boldsymbol{X}) \geq T(\boldsymbol{x})) \tag{4.48}$$

defines a valid p-value when testing

$$H_0 : \theta = \theta_0, H_1 : \theta \neq \theta_0,$$

where H_0 is simple. Due to Equation (4.2) $p^\star(\boldsymbol{x})$ is less than or equal to $p_M(\boldsymbol{x})$ for each outcome $\boldsymbol{x}$. In Table 4.10 we give a fictitious example to illustrate this property, where we show some of the calculated p-values using either Equation (4.48) or Equation (4.2). We see that $p^\star(\boldsymbol{x}) \leq p_M(\boldsymbol{x})$ for all outcomes shown. If we were to calculate the probability that either $p_M(\boldsymbol{X})$ or $p^\star(\boldsymbol{X})$ is less than or equal to 0.05 in the fictitious example when $\theta = \theta_0$, we know by using the enumeration procedure given in Section 3.3 that we should sum the probabilities of the outcomes $\boldsymbol{x}_1$ to $\boldsymbol{x}_4$ when calculating $\Pr_{\theta_0}(p(\boldsymbol{X}) \leq 0.05)$ and $\boldsymbol{x}_1$ to $\boldsymbol{x}_5$ when calculating $\Pr_{\theta_0}(p_M(\boldsymbol{X}) \leq 0.05)$. In this example we thus have $\Pr_{\theta_0}(p_M(\boldsymbol{X}) \leq 0.05) < \Pr_{\theta_0}(p^\star(\boldsymbol{X}) \leq 0.05)$ since we sum over a subset of the $\boldsymbol{x}_i$ used in calculating $\Pr_{\theta_0}(p^\star(\boldsymbol{X}) \leq 0.05)$ when we calculate $\Pr_{\theta_0}(p_M(\boldsymbol{X}) \leq 0.05)$. This observation holds in general, i.e we either use the same $\boldsymbol{x}_i$ or a subset of the $\boldsymbol{x}_i$ used in calculating $\Pr_{\theta_0}(p^\star(\boldsymbol{X}) \leq \alpha) = \sum_{p^\star(\boldsymbol{x}) \leq \alpha} \Pr_{\theta_0}(\boldsymbol{X} = \boldsymbol{x})$ when we calculate

Table 4.10: Fictitious example of the realisations of the p-values calculated using Equation (4.48) and (4.2) for outcomes $\boldsymbol{x}_1$ to $\boldsymbol{x}_6$. The results are not shown for the rest of the outcomes, but the p-values are larger than $p^\star(\boldsymbol{x}_5)$ and $p(\boldsymbol{x}_5)$ respectively.

$\boldsymbol{x}$	$p^\star(\boldsymbol{x})$	$p(\boldsymbol{x})$
$\boldsymbol{x}_1$	0.01	0.15
$\boldsymbol{x}_2$	0.03	0.03
$\boldsymbol{x}_3$	0.04	0.042
$\boldsymbol{x}_4$	0.045	0.051
$\boldsymbol{x}_5$	0.061	0.063

$\Pr_{\theta_0}(p_M(\boldsymbol{X}) \leq 0.05) = \sum_{\boldsymbol{p}_M(\boldsymbol{x}) \leq \alpha} \Pr_{\theta_0}(\boldsymbol{X} = \boldsymbol{x})$. We thus have (Casella & Berger 2002, p. 398)

$$\Pr_{\theta_0}(p_M(\boldsymbol{X}) \leq \alpha) \leq \Pr_{\theta_0}(p^\star(\boldsymbol{X}) \leq \alpha). \tag{4.49}$$

Since $p^\star(\boldsymbol{X})$ is a valid p-value we have

$$\Pr_{\theta_0}(p^\star(\boldsymbol{X}) \leq \alpha) \leq \alpha,$$

which holds for all $\alpha \in [0, 1]$. The above procedure can be carried out for all $\theta_0 \in \Theta_0$. This means that

$$\Pr_{\theta}(p_M(\boldsymbol{X}) \leq \alpha) \leq \alpha \tag{4.50}$$

holds for all $\alpha \in [0, 1]$ and all $\theta \in \Theta_0$, so that $p_M(\boldsymbol{X})$ defined by Equation (4.2) is a valid p-value.

4.6.2 Proof of E p-value being asymptotically valid

We want show that the E p-value defined in Equation (4.3) is asymptotically valid. In Equation (4.3) we regard $p_E(\boldsymbol{x})$ a function of $\hat{\theta}$ and not of $\boldsymbol{x}$. By using the enumeration procedure in Section 3.2 we get

$$p_E(\boldsymbol{x}) = \Pr_{\hat{\theta}}(T(\boldsymbol{X}) \geq T(\boldsymbol{x})) = \sum_{T(\boldsymbol{x}') \geq T(\boldsymbol{x})} \Pr_{\hat{\theta}}(\boldsymbol{X} = \boldsymbol{x}').$$

We assume that $\Pr_\theta(\boldsymbol{X} = \boldsymbol{x})$ is a continuous function of θ for each outcome $\boldsymbol{x}$. From Theorem 6 in Adams & Essex (2010, p. 80) we know that a sum of continuous functions is continuous. Therefore $p_E(\boldsymbol{x})$ is a continuous function of $\hat{\theta}$. Since maximum likelihood estimators are asymptotically consistent we know that

$$\hat{\theta} \xrightarrow{p} \tilde{\theta}, \tag{4.51}$$

where $\tilde{\theta}$ is the true value of θ under H_0. Theorem 4.2.5 then gives

$$p_E(\boldsymbol{x}) = \sum_{T(\boldsymbol{x}')\geq T(\boldsymbol{x})} \Pr_{\hat{\theta}}(\boldsymbol{X} = \boldsymbol{x}') \xrightarrow{p} \sum_{T(\boldsymbol{x}')\geq T(\boldsymbol{x})} \Pr_{\tilde{\theta}}(\boldsymbol{X} = \boldsymbol{x}'),$$

since $p(\boldsymbol{x}) = f(\hat{\theta})$ is a continuous function of $\hat{\theta}$ and $\hat{\theta} \xrightarrow{p} \tilde{\theta}$ by Equation (4.51). From Section 3.5

$$p_E(\boldsymbol{x}) = \Pr_{\tilde{\theta}}(T(\boldsymbol{X}) \geq T(\boldsymbol{x}))$$

defines a valid p-value when testing

$$H_0 : \theta = \tilde{\theta}, H_1 : \theta \neq \tilde{\theta},$$

and therefore $p_E(\boldsymbol{x})$ is asymptotically valid, which is what we wanted to show.

4.6.3 Proof of A p-values being asymptotically valid

The A p-value is defined in Equation (4.5), where we use the asymptotic distribution of $T(\boldsymbol{X})$ under H_0. This distribution does not depend on θ, so that the probability on the right hand side of Equation (4.5) does not depend on θ. Since the M p-value is valid, we know that the p-value defined in Equation (4.5) is valid since

$$\sup_{\theta\in\Theta_0} \Pr(Y \geq T(\boldsymbol{x})) = \Pr(Y \geq T(\boldsymbol{x})).$$

Therefore since the distribution of $T(\boldsymbol{X})$ converges to the asymptotic distribution of $T(\boldsymbol{X})$, Y, the A p-value is asymptotically valid.

4.6.4 Proof of C p-value being valid

The C p-value is defined in Equation (4.4). The probability on the right hand side of this equation does not depend on θ since $S(\boldsymbol{X})$ is a sufficient statistic for θ. It is important to realise that we do a unconditional experiment and condition on the value of $S(\boldsymbol{X})$ obtained in the experiment. We do not perform a conditional experiment where it is given that $S(\boldsymbol{X}) = S(\boldsymbol{x})$. For instance when testing equality of independent binomial proportions we know that the outcome $\boldsymbol{X}$ is given by $\boldsymbol{X} = (X_1, X_2)$ where $X_1 \sim \text{Binom}(\theta, n_1)$, $X_2 \sim \text{Binom}(\theta, n_2)$ and X_1 and X_2 are independent. However, given that $S(\boldsymbol{X}) = X_1 + X_2 = s$ the random variable $\boldsymbol{X}$ follows a hypergeometric distribution, or to be more exact $\boldsymbol{X}$ is determined by either of X_1 or X_2 (since the other one is s minus the value of the former). We know that X_1 follows a hypergeometric distribution with total size n_1+n_2, total number

of possible successes n_1 and s trials. We see that $\boldsymbol{X}$ is distributed differently in the conditional experiment than in the unconditional experiment.

Consider a specific conditional experiment where it is given that $S(\boldsymbol{X}) = s$. We then know

$$p_s(\boldsymbol{x}) = \Pr(T(\boldsymbol{X}) \geq T(\boldsymbol{X}) \mid S(\boldsymbol{X}) = s) \tag{4.52}$$

defines a valid p-value since

$$\sup_{\theta \in \Theta_0} \Pr(T(\boldsymbol{X}) \geq T(\boldsymbol{x}) \mid S(\boldsymbol{X}) = s) = \Pr(T(\boldsymbol{X}) \geq T(\boldsymbol{x}) \mid S(\boldsymbol{X}) = s)$$

and M p-values are valid from Section 4.6.1. For outcomes $\boldsymbol{x}$ such that $S(\boldsymbol{x}) = s$ we calculate the same p-values using Equation (4.52) and Equation (4.4). Therefore $p_C(\boldsymbol{x})$ defined in Equation (4.4) is valid given $S(\boldsymbol{X}) = s$. This observation holds for all the possible values of $S(\boldsymbol{X})$. By the law of total probability we then have (Casella & Berger 2002, p. 399)

$$\begin{aligned}
\Pr_\theta(p_C(\boldsymbol{X}) \leq \alpha) &= \sum_s \Pr_\theta(S(\boldsymbol{X}) = s) \Pr(p_C(\boldsymbol{X}) \leq \alpha \mid S(\boldsymbol{X}) = s) \\
&\overset{(\star)}{=} \sum_s \Pr_\theta(S(\boldsymbol{X}) = s) \Pr(p_s(\boldsymbol{X}) \leq \alpha \mid S(\boldsymbol{X}) = s) \\
&\overset{(\star\star)}{\leq} \sum_s \Pr_\theta(S(\boldsymbol{X}) = s) \cdot \alpha = 1 \cdot \alpha = \alpha,
\end{aligned}$$

where equality in transition $(\star)$ follows since $p_C(\boldsymbol{X})$ and $p_s(\boldsymbol{X})$ give the same p-value given $S(\boldsymbol{X}) = s$ and we have less than or equal to in transition $(\star\star)$ since $p_s(\boldsymbol{X})$ given $S(\boldsymbol{X}) = s$ is valid. This means the C p-value is valid.

In column 3 in Table 4.18 we have calculated all the C p-values using Z_p^2 as test statistic when $n_1 = n_2 = 5$. The rows are ordered in increasing value of the sufficient statistic $S(\boldsymbol{X}) = X_1 + X_2$. For each value of the sufficient statistic, the rows are ordered in increasing value of Z_p^2. From the above proof $p_C(\boldsymbol{X})$ is valid given every value of the sufficient statistic $S(X_1, X_2) = X_1 + X_2$, but is also unconditionally valid. The reason the C p-value is 1 when evaluated in other outcomes than $(0, 0)$ and (n_1, n_2) is that we calculate the realisations as conditional probabilities, conditional on the value of the sufficient statistic obtained in the unconditional experiment. In each conditional experiment the sum of the probabilities of the outcomes must be 1. Each possible outcome in the unconditional experiment is also the outcome in a conditional experiment. When we calculate all the realisations of $p_C(\boldsymbol{X})$ for the different outcomes in the unconditional experiment, we also consider all outcomes in the different conditional experiments. When we consider an outcome that is regarded as the weakest evidence against the null hypothesis

in the conditional experiment the outcome is part of, the conditional probability, which equals the realisation of the C p-value, must be 1.

As a note, we need to consider $p_s(\boldsymbol{X})$ given $S(\boldsymbol{X}) = s$ and not simply $p_s(\boldsymbol{X})$ since the distribution of $\boldsymbol{X}$ is given unconditionally and the unconditional distribution of $\boldsymbol{X}$ is different from the conditional distribution (which means the sample spaces are different). In the pure conditional setting, where it is given that $S(\boldsymbol{X}) = s$, it is possible to write

$$\Pr(p_s(\boldsymbol{X}) \leq \alpha).$$

We can of course also write

$$\Pr(p_s(\boldsymbol{X}) \leq \alpha \mid S(\boldsymbol{X}) = s)$$

but this is not common practise. In the unconditional setting we need to write

$$\Pr(p_s(\boldsymbol{X}) \leq \alpha \mid S(\boldsymbol{X}) = s)$$

when we want to consider the p-value $p_s(\boldsymbol{X})$ in the conditional experiment.

4.7 Comparing Z_p^2 and $|D|$, part I

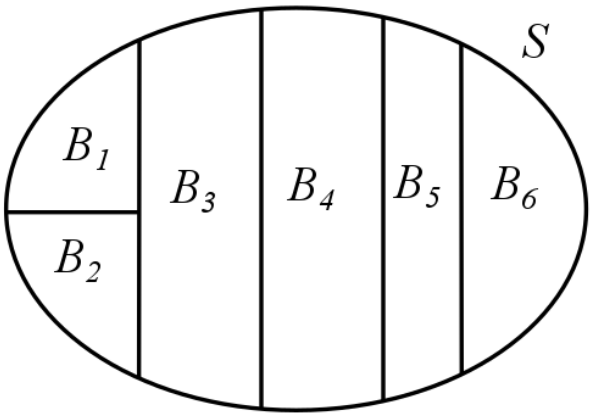
Even though we do not consider the test statistic $|D|$ (since the asymptotic distribution is a constant), we get a better understanding of the ordering property of the test statistic by comparing the ordering induced by this test statistic with the ordering induced by Z_p^2. We should also consider the realisations of the p-values generated by the same method using the two test statistics in turn as test statistic since only differences between the realisations of the p-values can give differences in the values of the power functions of the level α tests based on the p-values. We first consider the situation where $n_1 = 3$ and $n_2 = 3$. In columns 3 and 4 from the left in Table 4.11 we have evaluated the test statistics in the different outcomes and have ordered the rows in increasing value of $|D|(x_1, x_2)$. If we imagine creating boxes for the unique values of the test statistics and ordering the boxes for each test statistic, we see that the elements in the boxes (when we compare the boxes from one test statistic with the other) are either the same or we can think that we have split the box of the D-statistic in two to get the boxes for the Z_p^2-statistic.

To get a more precise statement we introduce the notion of *refinement* from set theory. Firstly, a *partition* of a sample space $\mathcal{S}$ is a set $\mathcal{P} = \{B_1, \ldots B_n\}$ of non-empty subsets B_i of $\mathcal{S}$ such that the subsets are disjoint ($B_i \cap B_j = \emptyset$ for $i \neq j$) and the union of them is the sample space ($B_1 \cup B_2 \cup \cdots \cup B_n = \mathcal{S}$) (Roman

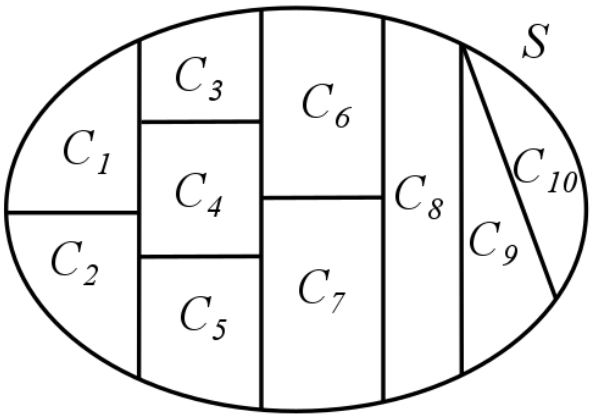
2012, p. 41). A partition $\mathcal{Q} = \{C_1, \ldots, C_m\}$ is called a refinement of a partition $\mathcal{P} = \{B_1, \ldots, B_n\}$ of the sample space $\mathcal{S}$ if each set C_i is completely contained in one of the B_i, or equivalently if B_i is the union of unique sets from $\mathcal{Q}$ (Roman 2012, p. 42). In Figure 4.7 we have illustrated a partition of a sample space and also a refinement of the partition. The figures are inspired from Figure 3.1 in Roman (2012, p. 42).

We can think that the test statistic creates a partition of the sample space where each subset consists of sample points with the same value of the test statistic, which is almost equivalent to the box analogy. Some informations is lost however when we use the concept of partition instead of the box analogy, since the sets in a partition are not ordered. More importantly, when $n_1 = 3$ and $n_2 = 3$ we can imagine the partition generated by Z_p^2 is a refinement of the partition generated by $|D|$. Does this fact have any practical consequences? To answer the question we study the p-value generated in this case by the M method. We postpone the treatment of the C and E p-values to Section 4.9.

In Table 4.11 we have also computed the M p-value where we have used $|D|$ as test statistic and the M p-value where we have used Z_p^2 as test statistic. We observe the M step reverses the ordering of the test statistic. This means the M p-value partitions the sample space the same as the original test statistic. However, if we make boxes and put outcomes with the same value of the p-value in the same box and order them in increasing value of the p-value, then the boxes appear in opposite order compared to if we order them using the original test statistic. This holds in general, i.e for any n_1 and n_2 when we use either $|D|$ or Z_p^2 as test statistic. When we consider the M p-value care must be taken. We maximise $\Pr_\theta(T(\boldsymbol{X}) \geq T(\boldsymbol{x}))$ over $\theta \in [0, 1]$, i.e including 0 and 1. From Equation (4.7) $\Pr_{(0,0)}(X_1 = 0, X_2 = 0) = 1$, $\Pr_{(1,1)}(X_1 = n_1, X_2 = n_2) = 1$ and both equals 0 if we change θ_1 or θ_2. From this equation we also have that $0 < \Pr_\theta(X_1 = x_1, X_2 = x_2) < 1$ for $0 < \theta_1, \theta_2 < 1$ and $\boldsymbol{x}$ not equal to either $(0, 0)$ or (n_1, n_2). We first consider Z_p^2. For a given outcome $x = (x_1, x_2)$, if $Z_p^2(x_1, x_2) < Z_p^2(n_1, n_2)$ or $Z_p^2(x_1, x_2) < Z_p^2(0, 0)$ the realisation of the p-value is 1 since we consider the outcomes $(0, 0)$ and (n_1, n_2) when calculating the realisation of the p-value, which is the same as the realisation when the outcome is either $(0, 0)$ or (n_1,n_2). If this is possible, the ordering of the M p-value is not the opposite of the ordering of the original test statistic. However, $Z_p^2(n_1, n_2)$ and $Z_p^2(0, 0)$ both give the lowest possible value of the test statistic. This also means that $\sup_{\theta \in [0,1]} \Pr_\theta(T(\boldsymbol{X}) \geq T(\boldsymbol{x}))$ will occur at $\theta \in (0, 1)$ when we calculate the realisation of the p-value for outcomes with a larger test statistic than $Z_p^2(n_1, n_2)$ and $Z_p^2(0, 0)$. If we consider $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ where $T(\boldsymbol{y}_1) > T(\boldsymbol{y}_2) > Z_p^2(n_1, n_2) = Z_p^2(0, 0)$ we sum over least as many outcomes $\boldsymbol{x}$ when calculating $p_M(\boldsymbol{y}_2)$ as when calculating $p_M(\boldsymbol{y}_1)$ and since



(a) The partition $\mathcal{P} = \{B_1, \ldots, B_n\}$ of the sample space S.



(b) The partition $\mathcal{Q} = \{C_1, \ldots, C_m\}$ of the sample space S.

Figure 4.7: Examples of two paritions $\mathcal{P}$ and $\mathcal{Q}$ of the sample space S, where the partition $\mathcal{Q}$ is a refinement of $\mathcal{P}$ since $B_1 = C_1, B_2 = C_2, B_3 = C_2 \cup C_3 \cup C_4 \cup C_5, B_4 = C_6 \cup C_7, B_5 = C_8, B_6 = C_9 \cup C_{10}$ (so that each B_i is a union of unique subsets from $\mathcal{Q}$).

$0 < \Pr_\theta(\boldsymbol{x}) < 1$ for $\theta \in (0,1)$ and $\boldsymbol{x} \notin \{(0,0),(n_1,n_2)\}$ for each $\boldsymbol{x}$, we must have $p_M(\boldsymbol{y}_2) > p_M(\boldsymbol{y}_1)$. This means the ordering of the negative of the M p-value is the same as the ordering of Z_p^2. If we replace Z_p^2 with $|D|$ almost all the previous statements will hold. The only difference is that there might be outcomes with the same value of the test statistic as either $(0,0)$ or (n_1,n_2). If we evaluate the M p-value in these outcomes, it will of course be 1. So, the M step reverses the ordering of Z_p^2 and $|D|$.

If we make a set of the all the different values of the p-value generated using the $|D|$-statistic that are less than or equal to α and make a similar set of the values of the p-value generated by the Z_p^2-statistic, then the first set is a subset of the other for all α. The reason is that the M step reverses the ordering of the original test statistic and that the partition given by Z_p^2 is a refinement of the partition given by $|D|$. Roughly speaking we get the same realisations of the p-value with Z_p^2 as test statistic as with $|D|$ as test statistic and some additional ones. When we calculate the power, we sum over at least as many outcomes for a given α when considering the p-value created by using the M step on Z_p^2 compared to when we use the p-value resulting from using the M method on $|D|$. This means the power function is least as high of the level α test based on M p-value generated by using Z_p^2 as test statistic as the power function of the level α test based on the M p-value generated by using $|D|$ as test statistic for all θ_1, θ_2 and α. Mathematically, we can write the last statement as $\gamma_1(\theta_1,\theta_2) \geq \gamma_2(\theta_1,\theta_2)$ for all α where $\gamma_1(\theta_1,\theta_2)$ is the power function of the level α test based on M p-value where Z_p^2 has been used as test statistic and $\gamma_1(\theta_1,\theta_2)$ is the power function of the level α test based on the M p-value where $|D|$ has been used as test statistic. So if a test statistic T_1 gives a partition of the sample space that is a refinement of the partition given by another test statistic T_2 and if we place the elements from the different sets in boxes and combine boxes for which the union of the elements give the sets in the partition given by T_2, the test statistic T_2 orders these boxes the same as its own boxes, then the power function of the level α test based on the M p-value where T_1 is used as test statistic is least as large as the power function of the level α test based on the M p-value where T_2 is used as test statistic.

Instead of introducing the concept of refinement, it is also possible to say that Z_p^2 is less discrete than $|D|$, meaning the former test statistic takes on more unique values than the latter (Mehrotra et al. 2004). Based on this, the M p value where Z_p^2 has been used as test statistic takes more unique values than the M p-value where $|D|$ has been used as test statistic. It is therefore expected that the power function of the level α-test based on the p-value generated by Z_p^2 takes least as high values as the power function of the level α-test based on the p-value generated by $|D|$ when we use the same method to create the p-value. However, this also

depends on the ordering induced by the test statistics. If a test statistic takes less unique values but induce a better ordering on the sample space than another test statistic that takes more unique values, it is not easy to compare the power function of the test based on the p-value where the first statistic is used as test statistic with the power function of the test based on the p-value where the second test statistic is used as test statistic for different α.

To illustrate this, we could compare the values of the p-values generated by the same enumeration method using $Z_{|D|}$ and $-Z_p^2$ as test statistics. If we use the M method then the only realisation of the p-value where $-Z_p^2$ is used as the test statistic would be 1 since $(0,0)$ and $(5,5)$ have the largest value of the test statistic and $\sup_{\theta\in[0,1]} \Pr(X_1 = 0, X_2 = 0) = 1 = \sup_{\theta\in[0,1]} \Pr(X_1 = 5, X_2 = 5)$. The p-value where one uses $|D|$ as test statistic takes more unique values. Note that $-Z_p^2$ takes more unique values than $|D|$. This means we also need to compare the orderings induced by the test statistics and not only the number of unique values of the test statistics when we want to compare the power functions of tests based on p-values created by the same method. We will see in Section 4.9 that discreteness of the test statistic is only certain to give discreteness in the possible values of the p-value when we use the M-method, meaning that the statement "$\ldots Z_p^2$ is less discrete than $|D| \ldots$ The p value generated from Z_p^2 takes more unique values than the p-value generated from $|D|$ when using the same method to create the p-value" is only true when we use the M-method to create the two p-values.

We also consider the M step when $n_1 = 3$ and $n_2 = 4$. By creating a new table in the same fashion as Table 4.11 was created, we get Table 4.12. Now we observe that the partition generated by Z_p^2 is not a refinement of the partition generated by $|D|$. The reason is that the outcomes $(2,1)$ and $(1,3)$ (which give the same value of each of the test statistics) are considered stronger evidence against H_0 than the outcomes $(1,0)$ and $(2,4)$ (both of which give the same value of each test statistic) by Z_p^2 and not as weaker evidence by $|D|$. Therefore the set $\{\boldsymbol{x} \mid p_{M1}(\boldsymbol{x}) \leq \alpha\}$ is not a subset of the set $\{\boldsymbol{x} \mid p_{M2}(\boldsymbol{x}) \leq \alpha\}$ for all $0 \leq \alpha \leq 1$. This means comparisons of the power functions based on the two p-values is not easy without doing computer simulations (of course, even if one of the sets is a subset of the other, you do not know how big difference there is between the power functions at different (θ_1, θ_2) without evaluating the power functions).

However, when $n_1 = 3$ and $n_2 = 4$ the orderings induced by the two test statistics are the same except for the order of the set of pairs of outcomes $B_1 = \{(2,1),(1,3)\}$ and $B_2 = \{(1,0),(2,4)\}$, see Table 4.11. If we use the box analogy, the difference between the orderings is that the box with the elements $\{(1,0),(2,4)\}$ comes just before the box with elements $\{(2,1),(1,3)\}$ when considering the boxes of Z_p^2 and that only the two mentioned boxes are reversed when we consider the boxes of

the other test statistic. This means only one realisation of the p-value generated by Z_p^2 is different from the realisations of the other p-value. The same holds when we compare the realisations of the other p-value with the realisations from the first p-value. This is verified by examining columns 4 and 5 of Table 4.11, where we see that only two realisations differ. This means the two power functions will be equal for some α. In fact, the power functions are equal except when $\alpha \in [30.54, 45.31)$.

Table 4.11: The M p-values where $|D|$ and Z_p^2 are used as test statistic when $n_1 = 3$ and $n_2 = 3$. For each row the first two columns from the left give the outcome, the third gives the value of $|D|$-statistic, the fourth column gives the value of the Z_p^2-statistic, the fifth and sixth cloumns give the M p-value where respectively $|D|$ and Z_p^2 is used as test statistic. We have ordered the rows in increasing value of $|D|$ from top to bottom.

x_1	x_2	$\lvert D\rvert(x_1, x_2)$	$Z_p^2(x_1, x_2)$	$p_{M1}(x_1, x_2)$	$p_{M2}(x_1, x_2)$
0	0	0.0000	−99.0000	1.0000	1.0000
1	1	0.0000	0.0000	1.0000	0.9687
2	2	0.0000	0.0000	1.0000	0.9687
3	3	0.0000	−99.0000	1.0000	1.0000
1	0	0.3333	1.2000	0.6875	0.5096
0	1	0.3333	1.2000	0.6875	0.5096
2	1	0.3333	0.6667	0.6875	0.6875
1	2	0.3333	0.6667	0.6875	0.6875
3	2	0.3333	1.2000	0.6875	0.5096
2	3	0.3333	1.2000	0.6875	0.5096
2	0	0.6667	3.0000	0.2187	0.2187
0	2	0.6667	3.0000	0.2187	0.2187
3	1	0.6667	3.0000	0.2187	0.2187
1	3	0.6667	3.0000	0.2187	0.2187
3	0	1.0000	6.0000	0.0312	0.0312
0	3	1.0000	6.0000	0.0312	0.0312

To sum up so far, a test statistic partitions the sample space. In each subset of the partition the elements have the same value of the test statistic. If we consider a test statistic T_1 that gives a refinement of the partition of the sample space given by a test statistic T_2, then the power function of the level α test based on the M p-value where T_1 is used as test statistic is least as high as the power function of the level α test where one instead of T_1 uses T_2 as test statistic. There is one necessary condition for the previous statement about the power functions to be true. In the box analogy, if we make new boxes for T_1 where we combine boxes so that the set of elements in each box is one of the subsets in the partition of T_2, then it is necessary that the boxes are ordered the same by the two test statistics. If

Table 4.12: The M p-values where $|D|$ and Z_p^2 are used as test statistic when $n_1 = 3$ and $n_2 = 4$. For each row the first two columns from the left give the outcome, the third gives the value of $|D|$-statistic, the fourth column gives the value of the Z_p^2-statistic, the fifth and sixth cloumns give the M p-value where respectively $|D|$ and Z_p^2 is used as test statistic. We have ordered the rows in increasing value of $|D|$ from top to bottom.

x_1	x_2	$\lvert D\rvert$	Z_p^2	p_{M1}	p_{M2}
0	0	0.0000	−99.0000	1.0000	1.0000
3	4	0.0000	−99.0000	1.0000	1.0000
1	1	0.0833	0.0583	0.9844	0.9844
2	3	0.0833	0.0583	0.9844	0.9844
2	2	0.1667	0.1944	0.7969	0.7969
1	2	0.1667	0.1944	0.7969	0.7969
0	1	0.2500	0.8750	0.5635	0.5635
3	3	0.2500	0.8750	0.5635	0.5635
1	0	0.3333	1.5556	0.4531	0.3054
2	4	0.3333	1.5556	0.4531	0.3054
2	1	0.4167	1.2153	0.4063	0.4531
1	3	0.4167	1.2153	0.4063	0.4531
0	2	0.5000	2.1000	0.2188	0.2188
3	2	0.5000	2.1000	0.2188	0.2188
2	0	0.6667	3.7333	0.1250	0.1250
1	4	0.6667	3.7333	0.1250	0.1250
3	1	0.7500	3.9375	0.0781	0.0781
0	3	0.7500	3.9375	0.0781	0.0781
3	0	1.0000	7.0000	0.0156	0.0156
0	4	1.0000	7.0000	0.0156	0.0156

the boxes are ordered the opposite, the statement made about the power functions is no longer valid. In Section 4.9 we consider the C and E method.

4.8 Interpretation of the p-value when the null hypothesis is composite

By definition a valid p-value satisfies

$$\sup_{\theta \in \Theta_0} \Pr_\theta(p(\boldsymbol{X}) \leq \alpha) \leq \alpha$$

for all $\alpha \in [0, 1]$. If we use a valid p-value to create a level α-test testing $H_0 : \theta \in \Theta_0$ against an alternative hypothesis, the realisation of the p-value we calculate after performing the experiment gives us the lowest significance level we could have specified *before* performing the experiment so that we would reject the null hypothesis. If we had specified a lower significance level, then we would not have rejected the null hypothesis. This is one possible interpretation of the p-value.

In Section 3.4 we gave a long run interpretation of the realisation of the p-value in an experiment, where we said that $p(x)$ (and x is the outcome in the experiment) is the long run proportion of experiments where we get a value of the test statistic at least as extreme as the original value of the test statistic. We call this interpretation for interpretation (a). When the null hypothesis is composite we do not know the true value of θ under the null hypothesis. Under H_0 only one value can be the true value of θ. When calculating $\Pr_\theta(T(X) \geq T(x))$ we get different numbers if we use different θ possible under H_0. The long run frequency converges to the number calculated using the true value of θ, which is unknown to the experimenter. The different four methods of calculating the realisations of a p-value when the null hypothesis is composite, given in Section 4.1.2, deal with this issue in different manners. In the E method we assume that the realisation of the maximum likelihood estimator for θ in the experiment is the true value of θ under H_0. If this holds, then interpretation (a) holds. In the M method we use the value of θ possible under H_0 that maximizes $\Pr_\theta(T(X) \geq T(x))$. If this value of θ is the true value of θ under H_0, interpretation (a) holds. In the A method we use the large sample distribution of the test statistic, which is free of θ. If this distribution is the true distribution of the test statistic, then interpretation (a) holds. When considering the C method we must be a little careful, since we calculate the realisations as conditional probabilities. Interpretation (a) holds if the repeated runs of the experiments are conditional experiments conditioned on the sufficient statistic being equal to the value in the original experiment. So interpretation (a) may also hold when the null hypothesis is composite.

In each of the four methods of calculating the realisations of a p-value when the null hypothesis is composite, we always calculate the realisation as a probability of obtaining a least as extreme test statistic that was obtained in the experiment. Exactly how we calculate the probability differers from method to method. This also means the lower the realisation we calculate the more evidence there is against the null hypothesis. From Section 3.3 we know we can view the p-value as a test statistic. This means we can view each of the four p-values, each created by one of the different methods in a specific experiment using one test statistic as a test statistic, as a test statistic for which low values of it indicate that the null hypothesis is false and the lower the realisation of the p-value we calculate for an outcome in the experiment the stronger the evidence the outcome provides against the null hypothesis (according to the p-value statistic). We want high values of a test statistic to indicate that a null hypothesis is wrong and the higher the value the stronger the evidence. So if we use the negative of the p-value the higher the value the stronger the evidence against the null hypothesis. So we can view the negative of a p-value as a test statistic. And we can possibly regard the negative of a p-value as a competing test statistic against the original test statistic which was used in the calculation of the p-value. We can also compare the different p-value statistics obtained from the four different methods of calculating the p-value. However, we need to investigate further if the ordering of the negative of any of the p-values are different from the original test statistic. If the orderings induced are the same, the power function of each of the new tests is the same as the power function of the corresponding original test.

When the null hypothesis is simple the ordering of the initial test statistic is the same as the ordering of the negative of the p-value. So we have that $T(\boldsymbol{x_1}) < T(\boldsymbol{x_2})$ if and only if $-p(\boldsymbol{x}_1) < -p(\boldsymbol{x}_2)$. The reason is that we sum over more outcomes when calculating $p(\boldsymbol{x}_1)$ than when calculating $p(\boldsymbol{x}_2)$ and use the same θ when calculating $\Pr_\theta(T(\boldsymbol{X}) \geq T(\boldsymbol{x}_1))$ as when computing $\Pr_\theta(T(\boldsymbol{X}) \geq T(\boldsymbol{x}_2))$ (which means we use the same probability distribution of the test statistic when we calculate the different realisations of the p-value).

Let us reason if the ordering of the negative of the different p-values are the same as that of the original test statistic when testing the hypotheses in Equation (4.1) using the Z_p^2 statistic. We start with the negative of the A p-value. Since we calculate all the realisations of the A p-value using the same distribution of the test statistic, the ordering of the negative of the A p-value must be the same as the ordering induced by the original test statistic (assuming that the asymptotic distribution is not a constant, which is not the case when we consider the asymptotic distribution of Z_p^2). The only two exceptions are for the outcomes $(0, 0)$ and (n_1, n_2). We have defined $Z_p^2(0, 0) = Z_p^2(n_1, n_2) = -99$, but a chi square random

variable cannot be less than 0, meaning we integrate the chi square distribution from 0 and not from -99 when we calculate $p_A(0,0)$ and $p_A(n_1,n_2)$. So the outcomes $(0,0)$ and (n_1,n_2) are regarded as equal evidence against the null hypothesis as the outcomes $\boldsymbol{x}$ where $Z_p^2(x_1,x_2)=0$ by $-p_A$.

When we calculate the realisations of the E p value we possibly use different values of $\hat{\theta}$ when we calculate $\Pr_{\hat{\theta}}(T(\boldsymbol{X}) \geq T(\boldsymbol{x}))$ for different outcomes $\boldsymbol{x}$. This means we possibly use different distributions of the test statistic when calculating the different realisations, so that the ordering of the negative of the E p-value does not need to be the same as the ordering of the original test statistic. When calculating the realisations of the C p-value, we possibly consider different conditional distributions of the test statistic when calculating $p_C(\boldsymbol{x}_1)$ and $p_C(\boldsymbol{x}_2)$ for two different outcomes $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. This means the ordering induced by the negative of the C p-value does not need to be equal to the ordering induced by the original test statistic.

The above discussion indicates that the ordering of the negative of the E and C p-values can be different from the ordering of Z_p^2 and that the ordering of the negative of the A p-value is the same as the ordering induced by Z_p^2 except for the outcomes $(0,0)$ and (n_1,n_2). From Section 4.7 we know the M step reverses the ordering of Z_p^2, so that the ordering of the negative of the M p-value is the same as the ordering induced by Z_p^2. We illustrate the mentioned properties in the example with $n_1=5, n_2=5$. We calculate all the realisations of the p-values using the four different methods. The results are shown in Table 4.13. As expected, we observe that the negative of the M p-value and the Z_p^2-statistic orders the sample space exactly the same. As previously noted, the ordering induced by the negative of the A p-value is almost the same as the ordering induced by the negative of the M p-value and the Z_p^2-statistic, where the difference is that not only the outcomes $(0,0)$ and (n_1,n_2) are regarded as the weakest evidence against the null hypothesis but also the outcomes (i,i) for $0 \leq i \leq 5$. We also observe that the ordering induced by the negative of the E p-value differs from the ordering induced by the other test statistics, where the outcome $(1,1)$ is regarded as stronger evidence against the null hypothesis than $(2,2)$ and $(3,3)$ and not the same evidence as by the other test statistics. We also observe that the negative of the C p-value statistic is -1 for many outcomes, which means that the ordering induced by this statistic differs from the order induced by the other test statistics. The reason we get so many -1's is given in Section 4.6.4.

We have illustrated that it is possible to use the negative of one of the four p values as a tests statistic and that the ordering of this test statistic may be different from the ordering induced by the original test statistic. We can then use the test statistic in one of the mentioned four procedures of calculating a p-value to create a new

Table 4.13: The negative of the four different p-values, $p_A(\boldsymbol{X})$, $p_M(\boldsymbol{X}), p_E(\boldsymbol{X})$ and $p_C(\boldsymbol{X})$, when $n_1 = n_2 = 5$. Column 1 and 2 give the outcomes in the experiment. Column 3 gives the value of the Z_p^2 statistic. Column 4 to 8 give the negative of respectively the A, M, C and E p-value statistics. The rows are ordered in increasing value of Z_p^2 starting from the top.

x_1	x_2	Z_p^2	$-p_A$	$-p_M$	$-p_C$	$-p_E$
0	0	−99.0000	−1.0000	−1.0000	−1.0000	−1.0000
5	5	−99.0000	−1.0000	−1.0000	−1.0000	−1.0000
1	1	0.0000	−1.0000	−0.9980	−1.0000	−0.8926
2	2	0.0000	−1.0000	−0.9980	−1.0000	−0.9938
3	3	0.0000	−1.0000	−0.9980	−1.0000	−0.9938
4	4	0.0000	−1.0000	−0.9980	−1.0000	−0.8926
3	2	0.4000	−0.5271	−0.7539	−1.0000	−0.7539
2	3	0.4000	−0.5271	−0.7539	−1.0000	−0.7539
2	1	0.4762	−0.4902	−0.6648	−1.0000	−0.6468
1	2	0.4762	−0.4902	−0.6648	−1.0000	−0.6468
4	3	0.4762	−0.4902	−0.6648	−1.0000	−0.6468
3	4	0.4762	−0.4902	−0.6648	−1.0000	−0.6468
5	4	1.1111	−0.2918	−0.5156	−1.0000	−0.4893
4	5	1.1111	−0.2918	−0.5156	−1.0000	−0.4893
1	0	1.1111	−0.2918	−0.5156	−1.0000	−0.4893
0	1	1.1111	−0.2918	−0.5156	−1.0000	−0.4893
3	1	1.6667	−0.1967	−0.3437	−0.5238	−0.3326
1	3	1.6667	−0.1967	−0.3437	−0.5238	−0.3326
4	2	1.6667	−0.1967	−0.3437	−0.5238	−0.3326
2	4	1.6667	−0.1967	−0.3437	−0.5238	−0.3326
2	0	2.5000	−0.1138	−0.1874	−0.4444	−0.1778
0	2	2.5000	−0.1138	−0.1874	−0.4444	−0.1778
5	3	2.5000	−0.1138	−0.1874	−0.4444	−0.1778
3	5	2.5000	−0.1138	−0.1874	−0.4444	−0.1778
4	1	3.6000	−0.0578	−0.1094	−0.2063	−0.1094
1	4	3.6000	−0.0578	−0.1094	−0.2063	−0.1094
5	2	4.2857	−0.0384	−0.0618	−0.1667	−0.0581
2	5	4.2857	−0.0384	−0.0618	−0.1667	−0.0581
3	0	4.2857	−0.0384	−0.0618	−0.1667	−0.0581
0	3	4.2857	−0.0384	−0.0618	−0.1667	-0.0581
4	0	6.6667	−0.0098	−0.0215	−0.0476	−0.0188
5	1	6.6667	−0.0098	−0.0215	−0.0476	−0.0188
0	4	6.6667	−0.0098	−0.0215	−0.0476	−0.0188
1	5	6.6667	−0.0098	−0.0215	−0.0476	−0.0188
5	0	10.0000	−0.0016	−0.0020	−0.0079	−0.0020
0	5	10.0000	−0.0016	−0.0020	−0.0079	−0.0020

Table 4.14: Excerpt of the realisations of $-p_E(\boldsymbol{X})$ where Z_p^2 is used as test statistic and $n_1 = 90, n_2 = 150$. For each row, the the third coloumn gives the value of the test statistic when evaluated in the outcome given in the two leftmost coloumns and the fourth coloumn gives the realisation of the negative of the E p-value statistic when evaluated in the outcome. The rows are ordered increasingly in Z_p^2 from top to bottom.

x_1	x_2	Z_p^2	$-p_E(\boldsymbol{x})$
44	73	0.001112	−0.9867
46	77	0.001112	−0.9867
43	72	0.001113	−0.9782
47	78	0.001113	−0.9782
41	68	0.001121	−0.9777
49	82	0.001121	−0.9777
40	67	0.001124	−0.9747
50	83	0.001124	−0.9747
38	63	0.001140	−0.9775
52	87	0.001140	−0.9775
37	62	0.001146	−0.9745
53	88	0.001146	−0.9745
55	92	0.001170	−0.9773

p-value. The question is then which of the four methods to use. We know that the A p-value and the E p-value can be liberal, so that they are not valid. We therefore do not consider these two methods. The remaining two p-values, the C or M $p-$value, are valid. This means that applying a C or M step to either the E or A p value will make it valid. The M method is preferred over the C method. The reason is given by the following theorem

Theorem 4.8.1 *Let $p(\boldsymbol{X})$ be any p-value statistic. We know that the lower the value of this statistic, the stronger the evidence against the null hypothesis. Let $p_M(\boldsymbol{X})$ denote the p-value when we use $-p(\boldsymbol{X})$ as test statistic in the M method, i.e $p_M(\boldsymbol{x}) = \sup_{\theta\in\Theta_0} \Pr_\theta(-p(\boldsymbol{X}) \geq -p(\boldsymbol{x}))$ for all possible outcomes $\boldsymbol{x}$. We then have (Bakke & Langaas n.d.)*

1. *$p_M(\boldsymbol{X})$ is a valid p-value*
2. *If $p(\boldsymbol{X})$ is a valid p-value, then $p_M(\boldsymbol{x}) \leq p(\boldsymbol{x})$ for all outcomes $\boldsymbol{x}$.*

The proof is as follows

1. We have that $p_{\mathrm{M}}(\boldsymbol{X})$ is valid since the original p-value can be considered an ordinary test statistic and the M-method gives a valid p-value. See Section 4.6.1 for a proof of the M method producing a valid p-value.
2. If $p(\boldsymbol{x})$ is valid, we know that $\Pr_\theta(p(\boldsymbol{X}) \leq \alpha) \leq \alpha$ holds for all $\theta \in \Theta_0$ and all $\alpha \in [0, 1]$. This must also hold if we set α equal to the original p-value statistic evaluated in a particular outcome $\boldsymbol{x}$, i.e $\Pr_\theta(p(\boldsymbol{X}) \leq p(\boldsymbol{x})) \leq p(\boldsymbol{x})$. This holds for all outcomes $\boldsymbol{x}$. Then we must also have (Bakke & Langaas n.d.) $\sup_{\theta\in\Theta_0} \Pr_\theta(p(\boldsymbol{X}) \leq p(\boldsymbol{x})) \leq p(\boldsymbol{x})$ for all outcomes $\boldsymbol{x}$, which means $p_M(\boldsymbol{x}) = \sup_{\theta\in\Theta_0} \Pr_\theta(-p(\boldsymbol{X}) \geq -p(\boldsymbol{x})) = \sup_{\theta\in\Theta_0} \Pr_\theta(p(\boldsymbol{X}) \leq p(\boldsymbol{x})) \leq p(\boldsymbol{x})$ so that $p_M(\boldsymbol{x}) \leq p(\boldsymbol{x})$ for all outcomes $\boldsymbol{x}$ when $p(\boldsymbol{x})$ is valid, which is what we wanted to show.

Care must be taken when we consider applying an M-step to the negative of any p-value. Since if the outcomes $(0, 0)$ and (n_1, n_2) are not considered the least evidence against the null hypothesis by the negative of the p-value, then applying a M step top the negative of the p-value might not reverse the ordering of the negative of the p-value. We know from Section 4.7 that $p_M(n_1, n_2) = p_M(0, 0) = 1$ and we have that $p_A(0, 0) = p_A(n_1, n_1) = 1$ since we integrate the chi square distribution from 0 to infinity. Also $p_C(0, 0) = p_C(n_1, n_2) = 1$ since $(0, 0)$ or (n_1, n_2) is the only outcome with the corresponding value of the sufficient statistic and the sum of the outcomes in the conditional experiment must be 1. When $(x_1, x_2) = (0, 0)$ then $\hat{\theta} = 0$ from Equation (4.12) and since $\Pr_{(0,0)}(X = (x_1, x_2)) = 1$ from Equation (4.7) it follows from Equation (4.3) that $p_E(0, 0) = 1$. Since $(n_1 - 0, n_2 - 0) = (n_1, n_2)$ (and from Section 4.5 we know the realisations of the p-values comes in the pairs

$p(x_1, x_2)$ and $p(n_1 - x_1, n_2 - x_2)$ with the same value of the p-value) $p_E(n_1, n_2) = 1$. We have shown that (n_1, n_2) and $(0, 0)$ are considered (among, there could also be other outcomes with the same value of the test statistic) the least evidence against the null hypothesis when we use the negative of either of the A, C, M or E p-value. This means the M-step reveres the ordering of the negative of the p-values. When we use a p-value as test statistic in any method, we always use the negative of the p-value as test statistic. We therefore say that the M step maintains the ordering.

We denote by $\mathrm{E} \circ \mathrm{M}$ the p-value obtained by first using the C method on the original test statistic and then using the M method on the negative of the resulting E p-value. We introduce this notation to avoid confusion with the EM-algorithm. This notation applies to any of the four introduced methods (A, E, C or M). For instance $\mathrm{C}^2 \circ \mathrm{M}$ means that we apply the C step two times before we apply the M step. As previously mentioned, the M step maintains the ordering of the original statistic, so applying the M step more than once in succession will produce the same p-value. If we try to apply the C step more than once in succession, the realisations of the p-value will also not change. We now explain why. Say that we get the outcome (x_1, x_2) in an experiment. When applying the C method we consider all tables (y_1, y_2) where $y_1 + y_2 = x_1 + x_2 = s$. From Equation (4.10) we know that the probability of such a table is given by

$$\Pr(X_1 = y_1 \mid S(\boldsymbol{X}) = s) = \frac{\binom{n_1}{y_1}\binom{n_2}{s-y_1}}{\binom{n_1+n_2}{s}}$$

And because $s - y_1 = y_2$

$$\Pr(X_1 = y_1 \mid S(\boldsymbol{X}) = s) = \frac{\binom{n_1}{y_1}\binom{n_2}{y_2}}{\binom{n_1+n_2}{s}} > 0$$

since $0 \leq x_1 \leq n_1, 0 \leq x_2 \leq n_2$. It therefore follows that when we calculate the different conditional tail probabilities, the conditional ordering is reversed. Since we use the negative of the C p-value and consider the same outcomes (y_1, y_2) such that $y_1 + y_2 = s$ when calculating the C p-value evaluated in (x_1, x_2) the second time, the resulting value is unchanged. This holds for all outcomes, meaning the C^2 p-value is the same as the C p-value.

When we compute the E p-value the ordering induced by the negative of the this p-value is likely to be different than from the original ordering, as previously explained. If we try to apply the E method a second time, we use the same value of θ when calculating the tail probability for the same outcome. However, there is still no reason to expect that the ordering is maintained with the E method.

So when we calculate the realisation of the E^2 p-value for a specific outcome, we possibly consider different outcomes than when we calculate the realisation of the E p-value for the same outcome. Therefore there is no reason the E^2 p-value should equal the E p-value.

We have illustrated that the ordering induced by the negative of the p-value may be different from the ordering induced by the original test statistic. We strongly recommend the reader to ponder the examples given until this becomes clear. For instance Günther et al. (2009) state that a test static $T(\boldsymbol{x})$ has the property that for all $\boldsymbol{x} \in \mathcal{S}$ and $\theta \in \Theta_0$, $\Pr_\theta(p(\boldsymbol{X}) \leq p(\boldsymbol{x})) = \Pr_\theta(T(\boldsymbol{X}) \geq T(\boldsymbol{x}))$. We know $\Pr_\theta(p(\boldsymbol{X}) \leq p(\boldsymbol{x}))$ is the profile of $-p(\boldsymbol{x})$ and that $\Pr_\theta(T(\boldsymbol{X}) \geq T(\boldsymbol{x}))$ is the profile of $T(\boldsymbol{x})$, which means the above statement says the two profiles are equal for all outcomes and all $\theta \in \Theta_0$. This would imply that the E $\circ$ M p-value would equal the M p-value, since if the above relation holds we must have $\sup_{\theta\in\Theta_0} \Pr_\theta(p_E(\boldsymbol{X}) \leq p_E(\boldsymbol{x})) = \sup_{\theta\in\Theta_0} \Pr_\theta(T(\boldsymbol{X}) \geq T(\boldsymbol{x}))$. This means $\Pr_\theta(p(\boldsymbol{X}) \leq p(\boldsymbol{x})) = \Pr_\theta(T(\boldsymbol{X}) \geq T(\boldsymbol{x}))$ does not hold in general. The general argument for this is that the partition given by the negative of the p-value does not need to be the same as the partition given by the test statistic (where the boxes are ordered the same) or give a refinement of the partition (where the "combined" boxes need to be ordered the same). For another example, see Table 4.19. We see for the given excerpt of outcomes the set of $\boldsymbol{x}$ for which $|D|(\boldsymbol{x}) \geq |D|(51, 41)$ is not the same as the set of outcomes $\boldsymbol{x}$ for which $p_{C_2}(\boldsymbol{x}) \leq p_{C_2}(51, 41)$.

By considering Table 4.13 it may look like the negative of the E p-value gives a refinement of the partition of the sample space given by Z_p^2. If this is generally true, i.e true for any n_1 and n_2, then the power function of the level α test based on the EM p-value will be least as great as the power function of the level α test based on the M p-value for any α, θ_1 and θ_2. However, this is not the case which we illustrate when $n_1 = 90$ and $n_2 = 150$. In Table 4.14 we have given a part of the values of the test statistic and the values of the negative of the E p-value. The rows are ordered in increasing value of Z_p^2 from top to bottom. We observe that the outcomes $(40, 67)$ and $(50, 83)$ are considered less evidence against the null hypothesis than $(38, 63)$ and $(52, 87)$ by the negative of the E p-value and not weaker as by the Z_p^2-statistic. This means the negative of the E p-value does not give a refinement of the partition given by Z_p^2.

The idea of using the negative of the p-value from one method as a test statistic in another method is not new. Boschloo (1970) applies the M step on the negative of the Fisher conditional p-value, which can be considered a C step, resulting in a C $\circ$ M p-value. In Fisher's conditional test we condition on the same sufficient statistic as in our C step, but the outcomes are ordered increasingly in the negative of the probability of the different outcomes. For instance if $(x_1, x_2) = (1, 1)$ when

$n_1 = 4, n_2 = 10$, the ordering of the different relevant outcomes in the Fisher test is given in Table 4.15. We have also included the Z_p^2-statistic evaluated in each of the outcomes in the same row as the probability for each outcome. We see that the ordering induced by the two test statistics are not the same. This means the p-values are not the same.

Table 4.15: Comparing the orderings induced by Z_p^2 and $-\Pr(X_1 = x_1 \mid T(\boldsymbol{X}) = 2)$ when $n_1 = 4$ $n_2 = 10$ and where we only consider outcomes such that $x_1 + x_2 = 2$. Column 1 and 2 give the outcomes x_1 and x_2 that satisfy $x_1 + x_2 = 2$. Column 3 gives the negative value of the conditional probability of the outcome, where we have conditioned on the sufficient statistic $T(X_1, X_2) = X_1 + X_2$ for θ. Column 4 gives the value of the Z_p^2 statistic evaluated in each outcome in the table. The rows are ordered in increasing value of the negative of the probabilities given in the third column from top to bottom.

x_1	x_2	$-\Pr(X_1 = x_1 \mid T(\boldsymbol{X}) = 2)$	$Z_p^2(x_1, x_2)$
0	2	−0.5238	0.8392
1	1	−0.4190	0.6425
2	0	−0.0571	6.3462

According to Theorem 4.8.1 we have that $p_{CM}(\boldsymbol{x}) \leq p_C(\boldsymbol{x})$ for all outcomes $\boldsymbol{x}$, i.e the C $\circ$ M p-value is less than or equal to the C p-value for each outcome $\boldsymbol{x}$. We have illustrated this in Figure 4.8 where we have plotted the C $\circ$ M p-value against the C p-value for each outcome $\boldsymbol{x}$ when $n_1 = 5, n_2 = 5$. We have also plotted the line $p_{CM} = p_M$. For a point below this line $p_{CM}(\boldsymbol{x})$ is less than $p_M(\boldsymbol{x})$. We note that more than one outcome may have the same value of both p_{CM} and p_M, so it is possible that a point in the plot corresponds to the pair of p-values for more than one outcome $\boldsymbol{x}$. Since $n_1 = 5, n_2 = 5$ we compute 36 pairs of p_{CM} and p_M (each outcome gives one unique pair). By considering the plot we see that there are only 7 unique values, so that more than one outcome must have the same pair of p-values. We also observe that all points on the plot in Figure 4.8 are below or on the line $p_M = p_{CM}$, which means $p_{CM}(\boldsymbol{x}) \leq p_C(\boldsymbol{x})$ for each outcome $\boldsymbol{x}$, and this is exactly what Theorem 4.8.1 predicts. It is important to realise that $p_{CM}(\boldsymbol{x}) \leq p_M(\boldsymbol{x})$ for all outcomes $\boldsymbol{x}$ means that $\{\boldsymbol{x} \mid p_M(\boldsymbol{x}) \leq \alpha\}$ is a subset of $\{\boldsymbol{x} \mid p_{CM}(\boldsymbol{x}) \leq \alpha\}$ for all $0 \leq \alpha \leq 1$. This means $\gamma_{CM}(\theta_1, \theta_2; \alpha) \geq \gamma_C(\theta_1, \theta_2; \alpha)$ for all (θ_1, θ_2) and all $0 \leq \alpha \leq 1$.

Lloyd (2008) also compares two independent binomial distributions where $n_1 = 47$ and $n_2 = 283$. The null hypothesis is the same as in Equation (4.1), but the alternative hypothesis studied is $\theta_1 > \theta_2$. As test statistic, Z_p is used. The author compares the E $\circ$ M p-value and the M p-value. When the outcome in the experiment is $(14, 48)$ the profile $\Pr_\theta(Z_p \geq Z_p(x_1, x_2))$ has a spike around $\theta = 0$.

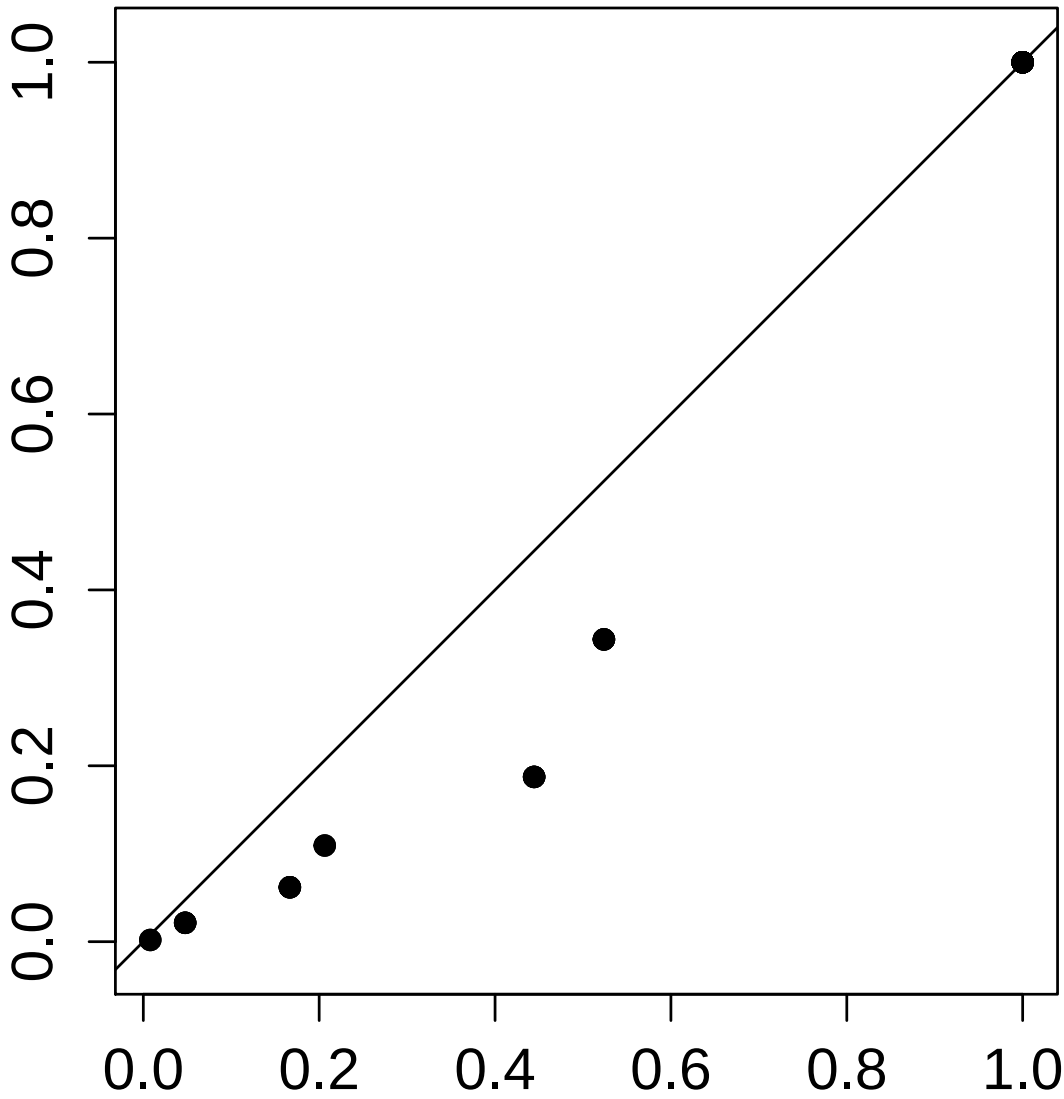


Figure 4.8: The C $\circ$ M p-value plotted against the C p-value for each outcome $\mathbf{x}$ when $n_1 = n_2 = 5$, i.e plot of $p_{CM}(\boldsymbol{x})$ against $p_C(\boldsymbol{x})$. The ordinate is the C $\circ$ M p-value and the abcissa is the C p-value.

When one instead considers the profile of the estimated p-value, i.e $\Pr_\theta(-p_E(\boldsymbol{X}) \geq -p_E(x_1, x_2))$, the profile is almost flat and the maximum value of the profile is lower than the spike in the profile of Z_p. Therefore the M p-value evaluated in the outcome is greater than the E $\circ$ M p-value evaluated in the outcome in the studied situation. We also observe the same phenomena when $n_1 = 90$, $n_2 = 150$, $x_1 = 60$, $x_2 = 109$ and the alternative hypothesis is two-sided. We have drawn the profile of $p_E(60, 109)$ in Figure 4.9 and the profile of $Z_p^2(60, 109)$ in Figure 4.10. When comparing the figures we see that $p_{EM}(60, 109)$ is smaller than $p_M(60, 109)$ since $\sup_{\theta\in[0,1]} \Pr_\theta(Z_p^2(\boldsymbol{X}) \geq Z_p^2(60, 109)) > \sup_{\theta\in[0,1]} \Pr_\theta(-p_E(\boldsymbol{X} \geq -p_E(60, 109))$.

One reasonable follow-up question to Lloyd's observation is if the E $\circ$ M p-value is less than the M p-value in general, i.e is $p_{EM}(\boldsymbol{x}) \leq p_M(\boldsymbol{x})$ for all outcomes $\boldsymbol{x}$? In Figure 4.11 the E $\circ$ M p-value is plotted against the M p-value for each outcome $\boldsymbol{x}$ where the realisations of either of the two p-values are below or equal to 0.10 in the example where $n_1 = 90$ and $n_2 = 150$. We observe that not all points are below the line $p_{EM} = p_M$, which means that $p_{EM}(\boldsymbol{x}) \leq p_M(\boldsymbol{x})$ does not hold in general. Since $p_{EM}(\boldsymbol{x}) \leq p_M(\boldsymbol{x})$ or vice versa does not hold in general for all outcomes $\boldsymbol{x}$, four situations may occur when comparing the power functions for a given α. Either (1) $\gamma_{EM}(\theta_1, \theta_2; \alpha) \geq \gamma_M(\theta_1, \theta_2; \alpha)$ for all (θ_1, θ_2), (2) $\gamma_{EM}(\theta_1, \theta_2; \alpha) \leq \gamma_M(\theta_1, \theta_2; \alpha)$ for all (θ_1, θ_2), (3) $\gamma_{EM}(\theta_1, \theta_2; \alpha) = \gamma_M(\theta_1, \theta_2; \alpha)$ for all (θ_1, θ_2) (4) combinations of situations (1) to (3) occur, for instance situation (1) might occur for some (θ_1, θ_2), situation (2) might occur for other (θ_1, θ_2) and for the remaining (θ_1, θ_2) situation 3 might occur.

When we want to calculate $\gamma_{EM}(\theta_1, \theta_2; \alpha)$ for $0 \leq \alpha \leq 0.1$ we consider the outcomes that give a point in Figure 4.11 below the horizontal line $p_{EM} = \alpha$ and when we want to calculate $\gamma_M(\theta_1, \theta_2; \alpha)$ we consider all the outcomes which have a point in Figure 4.11 below the vertical line $p_M = \alpha$. (Note: α can of course be above 0.10, but in our situation it needs to be less than or equal to this value since we have only plotted the realisations of the p_{EM}- or p_M-value where either one is below 0.10.) All the listed situations in the previous paragraph are possible when $n_1 = 90$ and $n_2 = 150$. We have tried to illustrate this in Figure 4.12. We see that situation (1) occurs when $\alpha = 0.08$ since the set of outcomes where $p_M(\boldsymbol{x}) \leq \alpha$ is a subset of the set of outcomes where $p_{EM}(\boldsymbol{x}) \leq \alpha$, situation (2) occurs when $\alpha = 0.058$ since then the set of outcomes where $p_{EM}(\boldsymbol{x}) \leq 0.058$ is a subset of the outcomes where $p_M(\boldsymbol{x}) \leq 0.058$, situation (3) occurs when $\alpha = 0.0273$ since then the set of outcomes where $p_{EM}(\boldsymbol{x}) \leq 0.0273$ is the same as the set of outcomes where $p_M(\boldsymbol{x}) \leq 0.0273$ and situation (4) occurs when $\alpha = 0.0276$ since then the set of outcomes where $p_{EM}(\boldsymbol{x}) \leq 0.0276$ and the set of outcomes where $p_M(\boldsymbol{x}) \leq 0.0276$ differ by two at least elements (there are at least two elements, at least one from

each set that is only in this set and not in the other. The reason that there is at least two points and not exactly two points, which the lower right plot in Figure 4.12 gives the impression of, is that more than one outcome can have the same E $\circ$ M and M p-value). We see that we get different results when comparing the power functions for fixed n_1, n_2 for different α. This means power comparisons in general only are valid for the studied α-values and are in fact heavily dependent on these values. Even if we find that one power function takes higher values than another for all (θ_1, θ_2) and does this for say $\alpha = 5 \cdot 10^{-k}$ for $k = 2, 3..., 8$ it is still possible that the other power function takes higher values than the first one (possibly for all $(\theta_1, \theta_2) \in \Theta$ or for some (θ_1, θ_2)) for other values of α.

4.9 Comparing Z_p^2 and $|D|$, part II

In this section we consider using the E and C method with Z_p^2 or $|D|$ as test statistic. We first consider the E method. In Table 4.16 we have calculated realisations of the p-values using the E step on either Z_p^2 (giving p_{E1}) or $|D|$ (giving p_{E2}) when $n_1 = 5, n_2 = 5$. We observe that the number of subsets in the partition of the sample space given by p_{E2} is much larger than the number of subsets in the partition given by $|D|$. If we exclude the realisations of the p-values that corresponds to outcomes where the test statistics are 0, p_{E1} and p_{E2} give the same partitions. If we consider the entire sample space (i.e also consider the outcomes where Z_p^2 or $|D|$ is 0), p_{E1} gives a refinement of the partition given by p_{E2}. We therefore observe that p_{E2} is much less discrete than $|D|$ and if we had applied a M step on p_{E1} and p_{E2} we would get almost the same values (only for the outcomes where Z_p^2 or $|D|$ is 0 the values of the resulting p-values differ). However $p_{E1}(\boldsymbol{x}) \leq p_{E2}(\boldsymbol{x})$ for all $\boldsymbol{x}$ meaning the power function of the level α test based on p_{E1} is least as great as the power function of the level α-test based on p_{E2} since for each α the set of $\boldsymbol{x}$ where $p_{E2}(\boldsymbol{x}) \leq \alpha$ is a subset of the set of $\boldsymbol{x}$ where $p_{E1}(\boldsymbol{x}) \leq \alpha$ for all α.

We also observe that p_{E2} is less discrete than $|D|$ for other n_1 and n_2 (data not shown). However since we use different $\hat{\theta}$ when calculating $\Pr_\theta(T(\boldsymbol{X}) \geq T(\boldsymbol{x}))$ for different outcomes $\boldsymbol{x}$ we cannot expect that the ordering induced by the negative of the E p-value is the same as the ordering induced by the original test statistic or gives a refinement. And since Z_p^2 is not in general a refinement of $|D|$ and the E method also does not give a refinement of the partition given by the original test statistic, we cannot expect in general that the p-value resulting from using E on Z_p^2 is a refinement of the p-value where one uses E on $|D|$. We have illustrated these observations in Table 4.17, where an excerpt of the E p-values using either

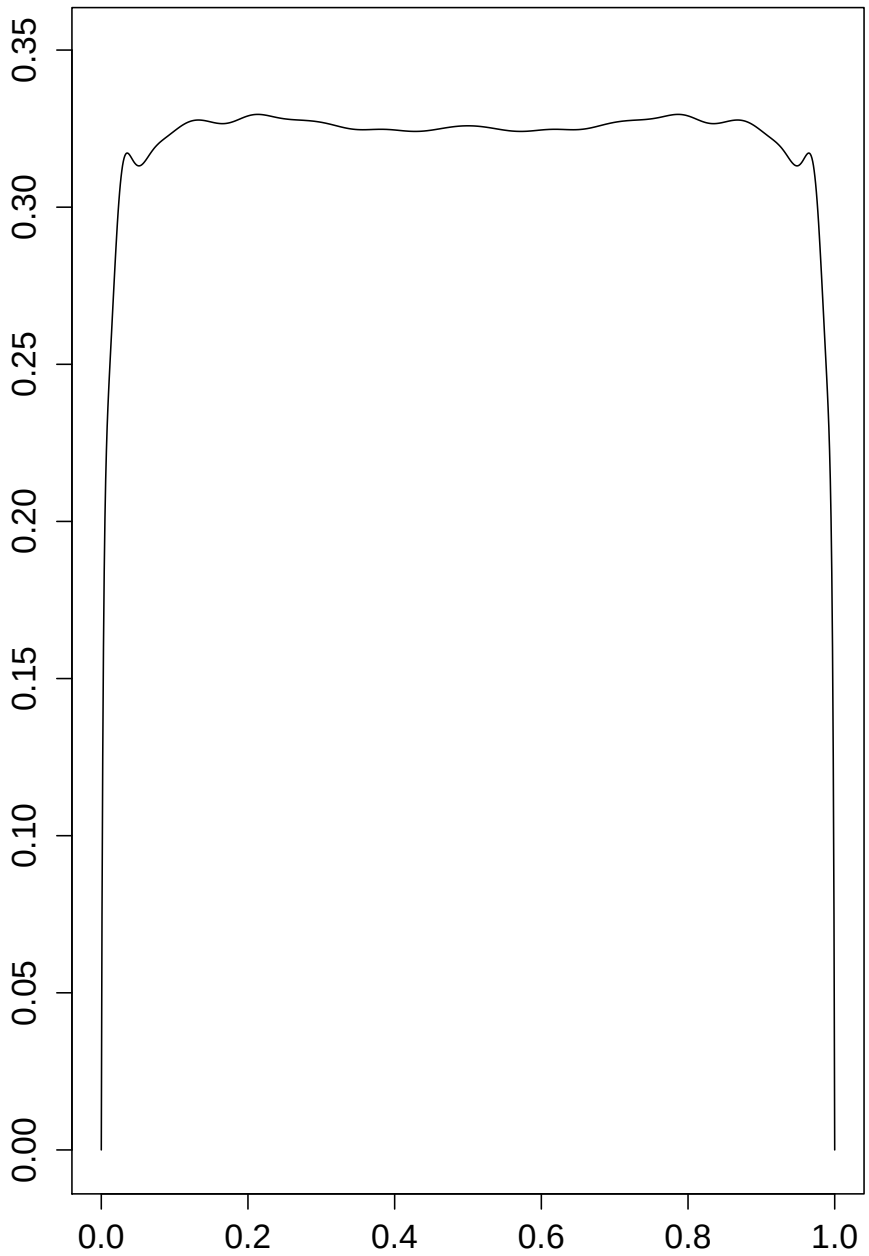


Figure 4.9: Profile of $-p_E(60, 109)$, i.e plot of $\Pr_\theta(-p_E(X_1, X_2) \geq -p_E(60, 109)$ as a function of θ, where p_E is the estimation p-value, $n_1 = 90$, $n_2 = 150$ and where we test equality of independent binomial proportions and the alternative hypothesis is two-sided.

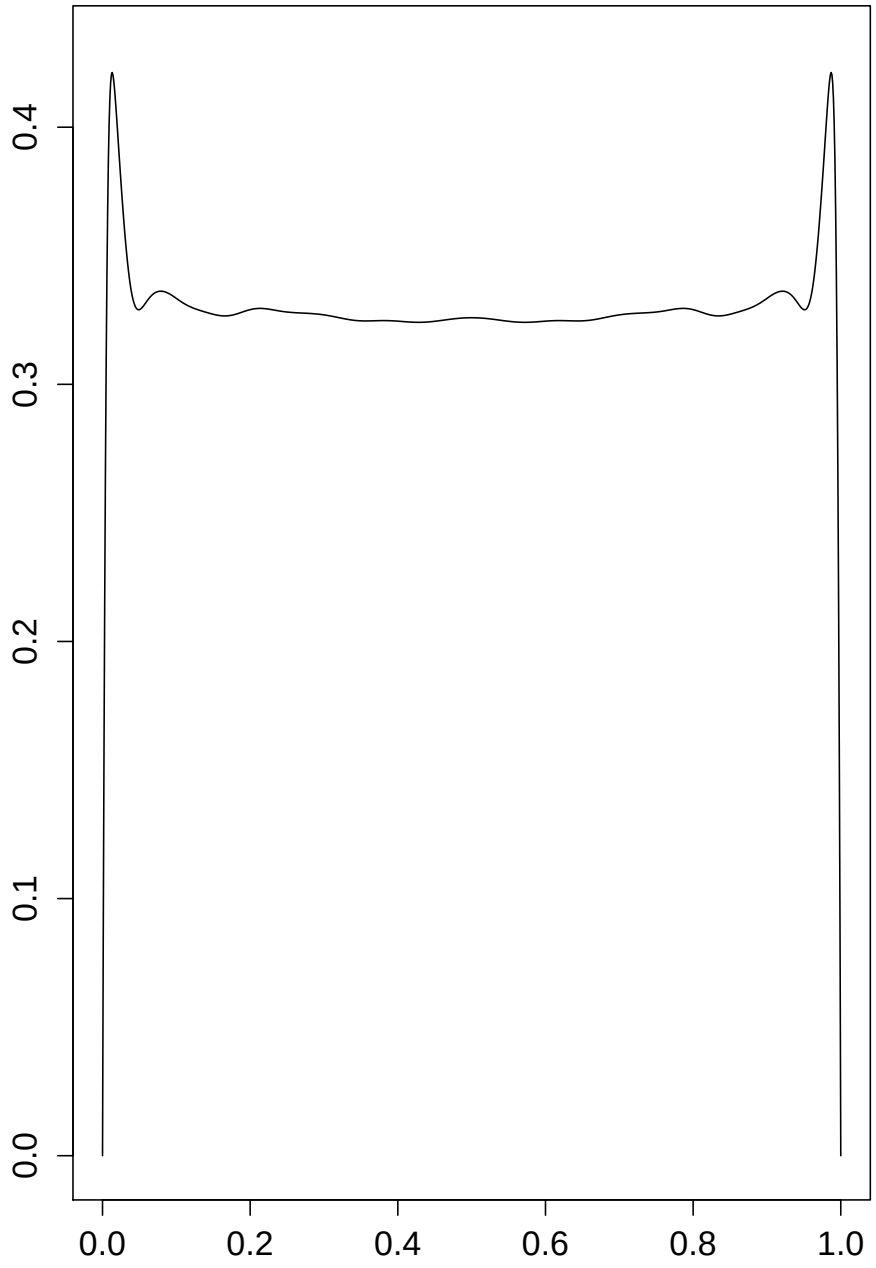


Figure 4.10: Profile of $Z_p^2(60, 109)$ when $n_1 = 90, n_2 = 150$.

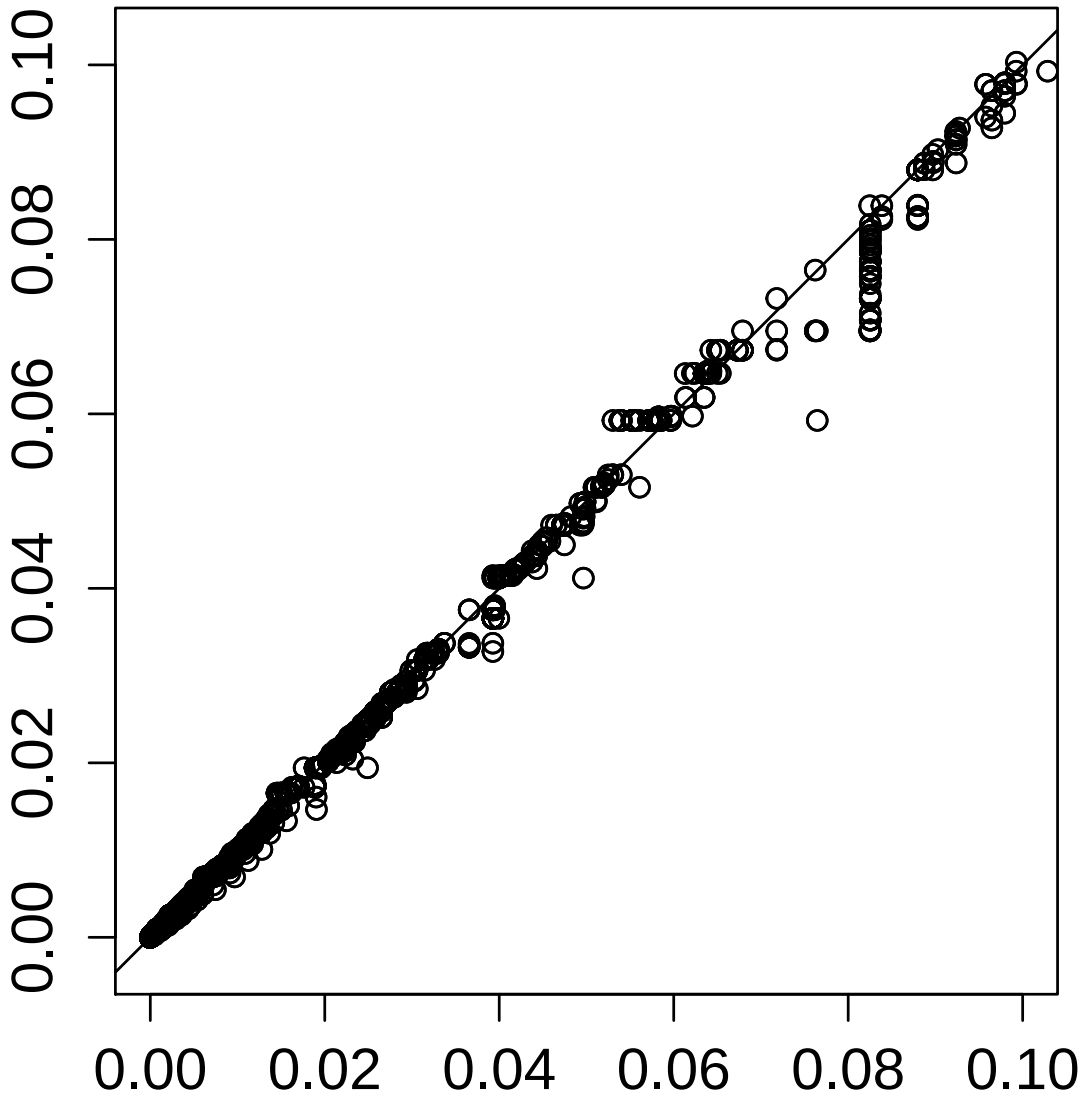


Figure 4.11: The E$\circ$M p-value plotted against the M p-value for each outcome $\boldsymbol{x}$ when $n_1 = 90$, $n_2 = 150$ and where we only consider realisations of the p-values were either one is below or equal to 0.10.

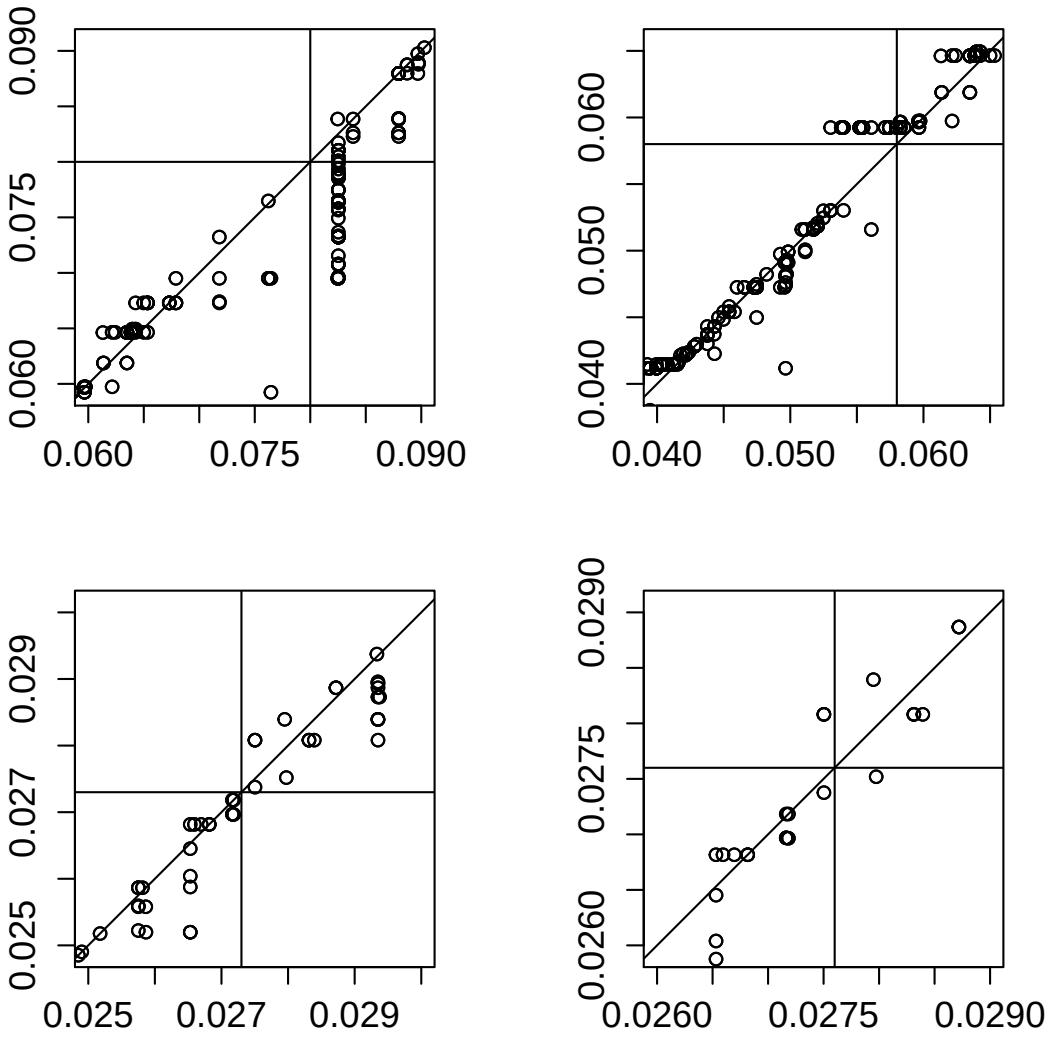


Figure 4.12: Comparing the sets $\{\boldsymbol{x} \mid p_{EM}(\boldsymbol{x}) \leq \alpha\}$ and $\{\boldsymbol{x} \mid p_M(\boldsymbol{x}) \leq \alpha\}$ for four different values of α, where p_{EM} is the E $\circ$ M p-value and p_M is the M p-value in the example where $n_1 = 90$ and $n_2 = 150$. We consider the values $\alpha_1 = 0.08, \alpha_2 = 0.058, \alpha_3 = 0.0273$ and $\alpha_4 = 0.0276$ of α. We have zoomed in on Figure 4.11 in four different regions to help with the comparisons, which give the four plots of the figure. Along the abscissa is the M p-value and along the ordinate is the E $\circ$ M p-value. The plots are ordered from left to right and the first plot starts on the top of the figure. In each plot we have drawn a vertical and horizontal line equal to α to help with the comparison of the sets. In the first plot the the lines equal α_1 and in the last plot they equal α_4.

Table 4.16: The realisations of the p-values when we use the E step on Z_p^2 or $|D|$ and $n_1 = 5, n_2 = 5$. The outcomes are given in the first two columns from the left. The value of Z_p^2 is given in the third column and the value of $|D|$ in the fourth column. The E p-value where Z_p^2 is used as test statistic is given in the fifth column and the p-value where $|D|$ is used as test statistic is given in the sixth column. The rows are ordered in increasing value of Z_p^2 from top to bottom.

x_1	x_2	Z_p^2	$\|D\|$	p_{E1}	p_{E2}
0	0	−99.0000	0.0000	1.0000	1.0000
5	5	−99.0000	0.0000	1.0000	1.0000
1	1	0.0000	0.0000	0.8926	1.0000
2	2	0.0000	0.0000	0.9938	1.0000
3	3	0.0000	0.0000	0.9938	1.0000
4	4	0.0000	0.0000	0.8926	1.0000
3	2	0.4000	0.2000	0.7539	0.7539
2	3	0.4000	0.2000	0.7539	0.7539
2	1	0.4762	0.2000	0.6468	0.7284
1	2	0.4762	0.2000	0.6468	0.7284
4	3	0.4762	0.2000	0.6468	0.7284
3	4	0.4762	0.2000	0.6468	0.7284
5	4	1.1111	0.2000	0.4893	0.5383
4	5	1.1111	0.2000	0.4893	0.5383
1	0	1.1111	0.2000	0.4893	0.5383
0	1	1.1111	0.2000	0.4893	0.5383
3	1	1.6667	0.4000	0.3326	0.3326
1	3	1.6667	0.4000	0.3326	0.3326
4	2	1.6667	0.4000	0.3326	0.3326
2	4	1.6667	0.4000	0.3326	0.3326
2	0	2.5000	0.4000	0.1778	0.2224
0	2	2.5000	0.4000	0.1778	0.2224
5	3	2.5000	0.4000	0.1778	0.2224
3	5	2.5000	0.4000	0.1778	0.2224
4	1	3.6000	0.6000	0.1094	0.1094
1	4	3.6000	0.6000	0.1094	0.1094
5	2	4.2857	0.6000	0.0581	0.0785
2	5	4.2857	0.6000	0.0581	0.0785
3	0	4.2857	0.6000	0.0581	0.0785
0	3	4.2857	0.6000	0.0581	0.0785
4	0	6.6667	0.8000	0.0188	0.0188
5	1	6.6667	0.8000	0.0188	0.0188
0	4	6.6667	0.8000	0.0188	0.0188
1	5	6.6667	0.8000	0.0188	0.0188
5	0	10.0000	1.0000	0.0020	0.0020
0	5	10.0000	1.0000	0.0020	0.0020

Z_p^2 or $|D|$ as test statistic are given when $n_1 = 148$ and $n_2 = 132$. We observe that p_{E1} and p_{E2} do not give partitions that are refinements of each other. We also observe that p_{E2} is much less discrete than $|D|$.

Table 4.17: Excerpt of the realisations of the p-values where one uses the E step on either Z_p^2 or $|D|$ and where $n_1 = 148$ an $n_2 = 132$. The outcomes are given in the two first columns on the left. The values of the Z_p^2 statistic when evaluated in the outcomes are given in column three and the corresponding values for $|D|$ are given in column four. The E p-value where Z_p^2 is used as test statistic is given in column five and the E p-value where $|D|$ is used as test statistic is given in column 6. The rows are ordered in increasing value of $|D|$ from top to bottom.

x_1	x_2	Z_p^2	$\lvert D\rvert$	p_{E1}	p_{E2}
41	41	0.3799	0.0336	0.5431	0.5406
70	58	0.3170	0.0336	0.5769	0.5763
78	74	0.3170	0.0336	0.5769	0.5763
107	91	0.3799	0.0336	0.5431	0.5406
115	107	0.4790	0.0336	0.4882	0.4922
144	124	1.9178	0.0336	0.1735	0.1667
32	33	0.4467	0.0338	0.5051	0.5060
42	33	0.4061	0.0338	0.5283	0.5268
69	66	0.3189	0.0338	0.5747	0.5747
79	66	0.3189	0.0338	0.5747	0.5747
106	99	0.4061	0.0338	0.5283	0.5268
116	99	0.4467	0.0338	0.5051	0.5060
143	132	4.5405	0.0338	0.0298	0.0328
5	0	4.5405	0.0338	0.0298	0.0328
51	41	0.3654	0.0340	0.5499	0.5486
60	58	0.3306	0.0340	0.5685	0.5682
14	8	1.1133	0.0340	0.2974	0.2959
23	25	0.5675	0.0340	0.4563	0.4541
88	74	0.3306	0.0340	0.5685	0.5682
97	91	0.3654	0.0340	0.5499	0.5486
125	107	0.5675	0.0340	0.4563	0.4541

We now consider the C method. In Table 4.18 we give the realisations of the C p-values obtained by using the C method with Z_p^2 and $|D|$ as test statistic. We observe that we calculate exactly the same p-value. This also holds when $n_1 = 148, n_2 = 132$. We give an excerpt of the realisations of the p-values in Table 4.19. So the discreteness of $|D|$ does not affect the power properties of the test when one applies the C step on this statistic when we compare with the power of the test resulting from using the same method on the less discrete test statistic

Z_p^2. The reason must be that conditional on any value of the sufficient statistic, Z_p^2 and $|D|$ order the outcomes the same.

To sum up the observations made in this section, we have seen that discreteness of the original test statistic does not necessarily mean the p-value where one has used this test statistic as the test statistic is less discrete than the p-value where one has used the same method on a less discrete test statistic. This is only guaranteed if one uses the M method to generate the p-value (which has been demonstrated in this section and in Section 4.7).

4.10 Notes

In this section we give various notes. We give a note on a property of the power functions, notes on properties of some of the power functions, notes on the E method and notes on how to numerically calculate the realisations of the M p-value.

4.10.1 Notes on symmetry of the power functions based on the i $\circ$ M p-values

AM-method, EM-method, CM-method From Section 4.5 we know that the realisations of the A, E and C p-values come in pairs $p(y_1, y_2)$ and $p(n_1-y_1, n_2-y_2)$ for all outcomes $\mathbf{y}$ where the values in each pair satisfies $p(\mathbf{y}) = p(n_1-y_1, n_2-y_2)$. When we calculate $p_{i\circ M}(y_1, y_2)$ and $p_{i\circ M}(n_1 - y_1, n_2 - y_2)$ we therefore maximise the same sum of probabilities, so that $p_{i\circ M}(y_1, y_2) = p_{i\circ M}(n_1 - y_1, n_2 - y_2)$. We also know that from Section 4.5 that the joint contribution of $p_{i\circ M}(y_1, y_2)$ and $p_{i\circ M}(n_1-y_1, n_2-y_2)$ to $\Pr_{\boldsymbol{\theta}}(p_{i\circ M}(\boldsymbol{X}) \leq \alpha)$ is the same at (θ_1, θ_2) and $(1-\theta_1, 1-\theta_2)$. Since this holds for all $\boldsymbol{y}$ such that $p_{i\circ M}(\boldsymbol{y}) \leq \alpha$, the power function is the same when evaluated at (θ_1, θ_2) and $(1 - \theta_1, 1 - \theta_2)$.

4.10.2 Notes on the type I error probabilities

Since the power function of the level α tests based on either the A, E, C, M, E$\circ$M, A$\circ$M or C$\circ$M p-values is the same when evaluated at (θ_1, θ_2) and $(1-\theta_1, 1-\theta_2)$, the power function is symmetric around $\theta = \frac{1}{2}$ when we intersect the graph of the power function with the plane $\theta_1 = \theta_2$, i.e when we consider the type I error probabilities. This can be formally shown as follows. We know that $\Pr_{(\theta_1,\theta_2)}(p(\boldsymbol{X}) \leq \alpha) =$

Table 4.18: The realisations of the C p-values where either Z_p^2 or $|D|$ is used as test statistic and $n_1 = 5, n_2 = 5$. The outcomes are given in column 1 and 2 from the left. The value of the sufficient statistic $S(X_1, X_2) = X_1 + X_2$ for $\theta = \theta_1 = \theta_2$ under H_0 is given in coloumn 3. The value of Z_p^2 and $|D|$ is given in respectively column 4 and column 5. The C p-value where Z_p^2 and $|D|$ is used as test statistic is given in respectively column 6 and column 7. The rows are ordered in increasing value of S from top to bottom. For each value of the sufficient statistic, the rows are ordered in increasing value of Z_p^2 from the top.

x_1	x_2	S	Z_p^2	$\lvert D\rvert$	p_{C1}	p_{C2}
0	0	0	−99.0000	0.0000	1.0000	1.0000
0	1	1	1.1111	0.2000	1.0000	1.0000
1	0	1	1.1111	0.2000	1.0000	1.0000
1	1	2	0.0000	0.0000	1.0000	1.0000
0	2	2	2.5000	0.4000	0.4444	0.4444
2	0	2	2.5000	0.4000	0.4444	0.4444
1	2	3	0.4762	0.2000	1.0000	1.0000
2	1	3	0.4762	0.2000	1.0000	1.0000
0	3	3	4.2857	0.6000	0.1667	0.1667
3	0	3	4.2857	0.6000	0.1667	0.1667
2	2	4	0.0000	0.0000	1.0000	1.0000
1	3	4	1.6667	0.4000	0.5238	0.5238
3	1	4	1.6667	0.4000	0.5238	0.5238
0	4	4	6.6667	0.8000	0.0476	0.0476
4	0	4	6.6667	0.8000	0.0476	0.0476
2	3	5	0.4000	0.2000	1.0000	1.0000
3	2	5	0.4000	0.2000	1.0000	1.0000
1	4	5	3.6000	0.6000	0.2063	0.2063
4	1	5	3.6000	0.6000	0.2063	0.2063
5	0	5	10.0000	1.0000	0.0079	0.0079
0	5	5	10.0000	1.0000	0.0079	0.0079
3	3	6	0.0000	0.0000	1.0000	1.0000
2	4	6	1.6667	0.4000	0.5238	0.5238
4	2	6	1.6667	0.4000	0.5238	0.5238
5	1	6	6.6667	0.8000	0.0476	0.0476
1	5	6	6.6667	0.8000	0.0476	0.0476
2	5	7	4.2857	0.6000	0.1667	0.1667
5	2	7	4.2857	0.6000	0.1667	0.1667
3	4	7	0.4762	0.2000	1.0000	1.0000
4	3	7	0.4762	0.2000	1.0000	1.0000
4	4	8	0.0000	0.0000	1.0000	1.0000
5	3	8	2.5000	0.4000	0.4444	0.4444
3	5	8	2.5000	0.4000	0.4444	0.4444
4	5	9	1.1111	0.2000	1.0000	1.0000
5	4	9	1.1111	0.2000	1.0000	1.0000
5	5	10	−99.0000	0.0000	1.0000	1.0000

Table 4.19: Excerpt of the realisations C p-values where either Z_p^2 or $|D|$ is used as test statistic and $n_1 = 148, n_2 = 132$. The possible outcomes in the experiment are given in the two first columns on the left, the values of the Z_p^2 when evaluated in the outcomes are given in column three and the values of $|D|$ are given in column four. The C p-value where Z_p^2 has been used as test statistic is given in column five. The values we get by applying the C method on $|D|$ are given in column six. The rows are ordered in increasing value of Z_p^2 from top to bottom.

x_1	x_2	Z_p^2	$\lvert D\rvert$	p_{C1}	p_{C2}
41	41	0.3799	0.0336	0.5992	0.5992
70	58	0.3170	0.0336	0.6311	0.6311
78	74	0.3170	0.0336	0.6311	0.6311
107	91	0.3799	0.0336	0.5992	0.5992
115	107	0.4790	0.0336	0.5554	0.5554
144	124	1.9178	0.0336	0.2379	0.2379
32	33	0.4467	0.0338	0.5711	0.5711
42	33	0.4061	0.0338	0.5892	0.5892
69	66	0.3189	0.0338	0.6321	0.6321
79	66	0.3189	0.0338	0.6321	0.6321
106	99	0.4061	0.0338	0.5892	0.5892
116	99	0.4467	0.0338	0.5711	0.5711
143	132	4.5405	0.0338	0.0623	0.0623
5	0	4.5405	0.0338	0.0623	0.0623
51	41	0.3654	0.0340	0.6106	0.6106
60	58	0.3306	0.0340	0.6280	0.6280
14	8	1.1133	0.0340	0.3750	0.3750
23	25	0.5675	0.0340	0.5258	0.5258
88	74	0.3306	0.0340	0.6280	0.6280
97	91	0.3654	0.0340	0.6106	0.6106
125	107	0.5675	0.0340	0.5258	0.5258

$f(\theta_1, \theta_2)$ and that $f(\theta_1, \theta_2) = f(1-\theta_1, 1-\theta_2)$. We consider $\theta_1 = \theta_2 = \theta$ and introduce the shift of coordinates $\tilde{\theta} = \theta - \frac{1}{2}$, which means $\theta = \tilde{\theta} + \frac{1}{2}$. Then

$$f(\theta, \theta) = g(\theta) = g(\tilde{\theta} + \frac{1}{2}) \overset{(\star)}{=} g(1-\theta) = g(1 - \tilde{\theta} - \frac{1}{2}),$$

where we have used $g(\theta) = g(1-\theta)$ in transition $(\star)$. This means

$$g(\tilde{\theta} + \frac{1}{2}) = g(\frac{1}{2} - \tilde{\theta}),$$

where we have shifted the coordinates of θ in the old coordinate system to $\tilde{\theta}$ in the new coordinate system. Since the new coordinate system is centred in $\theta = \frac{1}{2}$, g is symmetric around $\frac{1}{2}$.

From Figure 4.4 it looks like the power at (0,0) and (1,1) is 0. In fact, this holds for all n_1, n_2, which we now show. We know from Section 4.8 that $p(n_1, n_2) = p(0,0) = 1$ for the considered p-values. At $(0,0)$ we have $\Pr_\theta(\mathbf{X} = \mathbf{x}) > 0$ only for $\mathbf{x} = (0,0)$ and $\Pr_\theta(\mathbf{X} = (0,0)) = 1$, which means the power function evaluated at $(0,0)$ it is 0 when $\alpha < 1$ and 1 if $\alpha = 1$. Since $\Pr_{(\theta,\theta)}(p(\boldsymbol{X} \leq \alpha)$ is symmetric in $\alpha = \frac{1}{2}$, we also know that $\gamma(1,1;\alpha) = \Pr_{(1,1)}(p(\boldsymbol{X} \leq \alpha)$ is 0 when $\alpha < 1$ and 1 if $\alpha = 1$.

We note that the observation of symmetry of the power functions does not appear to be new. For instance in Mehrotra et al. (2004) they only calculate type I error probabilities for $0 \leq \theta \leq \frac{1}{2}$ and when they evaluate the power functions they only evaluate them at (θ_1, θ_2) above the line $\theta_1 = \theta_2$ (without making any comments about the power symmetry).

4.10.3 Comparing γ_{EM} with γ_E and γ_{AM} with γ_A

From Section 4.8 we know the ordering induced by negative of the E $\circ$ M or A $\circ$ M p-value is the same as the ordering induced by respectively the negative of the E or A p-value. This means the sets $\{\boldsymbol{x} \mid p_A(\boldsymbol{x}) \leq \alpha\}$ and $\{\boldsymbol{x} \mid p_{AM}(\boldsymbol{x}) \leq \alpha\}$ are either the same or one of them is a proper subset of the other. The same holds for the sets $\{\boldsymbol{x} \mid p_E(\boldsymbol{x}) \leq \alpha\}$ and $\{\boldsymbol{x} \mid p_{EM}(\boldsymbol{x}) \leq \alpha\}$. When $\sup_{\theta \in [0,1]} \gamma_A(\theta, \theta; \alpha) > \alpha$, then $\gamma_A(\theta_1, \theta_2) \geq \gamma_{AM}(\theta_1, \theta_2)$ for all $(\theta_1, \theta_2) \in \Theta$ since $\{\boldsymbol{x} \mid p_{AM}(\boldsymbol{x}) \leq \alpha\}$ must be a proper subset of $\{\boldsymbol{x} \mid p_A(\boldsymbol{x}) \leq \alpha\}$. The reason the set $\{\boldsymbol{x} \mid p_{AM}(\boldsymbol{x}) \leq \alpha\}$ must be a proper subset of $\{\boldsymbol{x} \mid p_A(\boldsymbol{x}) \leq \alpha\}$ is that the condition $\sup_{\theta \in [0,1]} \gamma_{AM}(\theta, \theta; \alpha) \leq \alpha$ must be fulfilled. Similarly, $\gamma_E(\theta_1, \theta_2; \alpha) \geq \gamma_{EM}(\theta_1, \theta_2; \alpha)$ for all $(\theta_1, \theta_2) \in \Theta$ when $\sup_{\theta \in [0,1]} \gamma_E(\theta, \theta; \alpha) > \alpha$. Since the A $\circ$ M p-value is the same as the M p-value, $\gamma_A(\theta_1, \theta_2; \alpha) \geq \gamma_M(\theta_1, \theta_2; \alpha)$ when $\sup_{\theta \in [0,1]} \gamma_A(\theta, \theta; \alpha) > \alpha$.

We know from Theorem 4.8.1 that $p_{CM}(\boldsymbol{x}) \leq p_C(\boldsymbol{x})$ for all outcomes $\boldsymbol{x}$. One reasonable question is then if similar results holds for the E $\circ$ M and E p-values and for the A $\circ$ M and A p-values. In Figure 4.13 we plot $p_{EM}(\boldsymbol{x})$ against $p_E(\boldsymbol{x})$ for each outcome $\boldsymbol{x}$ when $n_1 = n_2 = 10$. We see that not every point is below or on the line $p_{EM} = p_E$, so that $p_{EM}(\boldsymbol{x}) \leq p_E(\boldsymbol{x})$ does not hold in general. In Figure 4.14 we plot $p_M = p_{AM}$ against p_A for each outcome $\boldsymbol{x}$ also when $n_1 = n_2 = 10$, but we only plot pairs of realisations of the p-values for which both are below 0.005. Not every point is below or on the line $p_M = p_A$, which means $p_A(\boldsymbol{x}) \leq p_{AM}(\boldsymbol{x}) = p_M(\boldsymbol{x})$ does not hold in general. So it is not possible in general to establish similar results for either the E $\circ$ M and E p-values or the A $\circ$ M and A p-values as for the C $\circ$ M and C p-values.

4.10.4 The E method

As noted in the proof of the asymptotic validity of the E method, see Section 4.6.2, the null hypothesis in Equation (4.47) changes asymptotically when we use the E method to

$$H_0 : \theta = \tilde{\theta}, \tag{4.53}$$

where $\tilde{\theta}$ is the true value of θ under H_0. The reason is that we use the mle of θ which converges in probability to the true value of θ under H_0. For finite sample sizes, however, the null hypothesis is still given by Equation (4.47) since the mle has most likely not converged to the true value of θ under H_0. In fact we use different values of θ when we calculate the different realisations of $p_E(\boldsymbol{X})$. There is therefore no reason the E p-value should be valid for finite sample sizes.

A sufficient condition for validity of a p-value is

$$p(\boldsymbol{x}) \geq p_M(\boldsymbol{x}), \tag{4.54}$$

for all outcomes $\boldsymbol{x}$ where $p_M(\boldsymbol{X})$ is the M p-value, since then we can replace $p_M(\boldsymbol{x})$ in Equation (4.49) and Equation (4.50) with $p(\boldsymbol{x})$ so that $p(\boldsymbol{X})$ is valid. This condition is not satisfied for the E p-value since

$$p_E(\boldsymbol{x}) \leq p_M(\boldsymbol{x}), \tag{4.55}$$

which follows from the fact that when calculating $p_E(\boldsymbol{x})$ we use the mle of θ based on $\boldsymbol{x}$ but when calculating $p_M(\boldsymbol{x})$ we pick the value of θ that maximizes the tail probability $\Pr_\theta(T(\boldsymbol{X}) \geq T(\boldsymbol{x}))$. The condition in Equation (4.54) is not a necessary condition for validity, so that $p_E(\boldsymbol{x}) \leq p_M(\boldsymbol{x})$ for all outcomes $\boldsymbol{x}$ does not imply anything about the validity of $p_E(\boldsymbol{X})$. If this condition was necessary, $\gamma_{EM}(\theta_1, \theta_2; \alpha) \leq \gamma_M(\theta_1, \theta_2; \alpha)$ for all $(\theta_1, \theta_2) \in \Theta$ for each fixed $\alpha \in [0, 1]$. As we

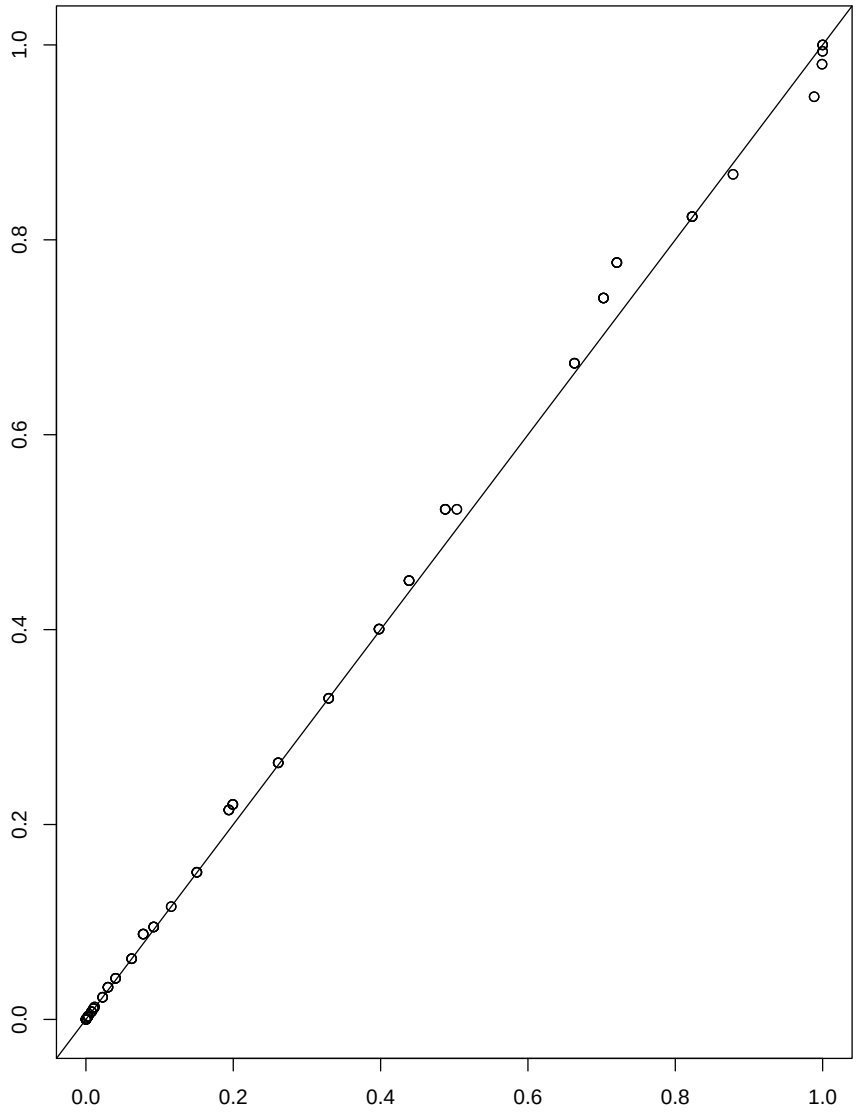


Figure 4.13: The E ∘ M p-value plotted against the E p-value for each outcome $\mathbf{x}$ when $n_1 = n_2 = 10$, i.e plot of $p_{EM}(\boldsymbol{x})$ against $p_E(\boldsymbol{x})$. The ordinate is the E ∘ M p-value and the abcissa is the M p-value.

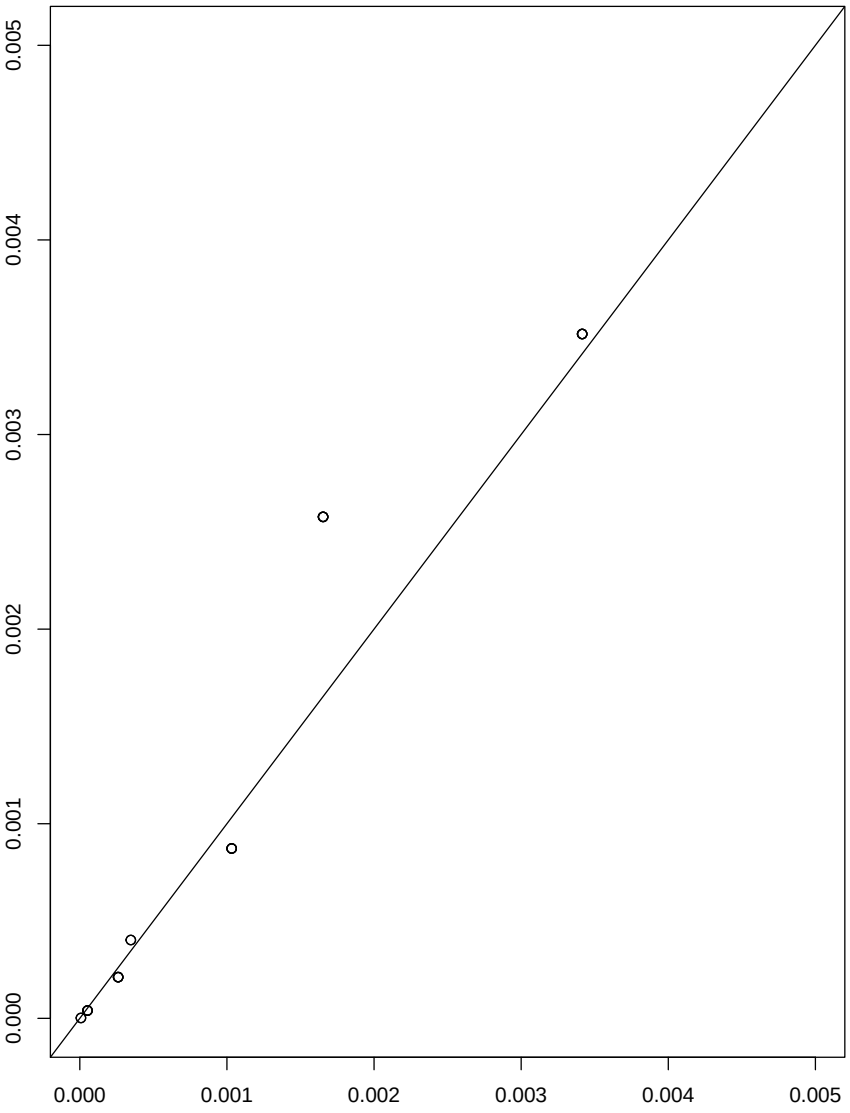


Figure 4.14: The A $\circ$ M p-value plotted against the A p-value for each outcome $\mathbf{x}$ when $n_1 = n_2 = 10$ and only realisations between 0 and 0.005 for each p-value are shown, i.e plot of $p_{AM}(\boldsymbol{x})$ against $p_A(\boldsymbol{x})$ when $0 \leq p_{AM}(\boldsymbol{x}), p_A(\boldsymbol{x}) \leq 0.005$. The ordinate is the A $\circ$ M p-value and the abcissa is the A p-value.

will see in Chapter 3, this does not hold in general. However, as previously noted the E p-vale is in general not valid.

Due to Equation (4.55) the set $\{\boldsymbol{x} \mid p_M(\boldsymbol{x}) \leq \alpha\}$ is either a proper subset of or equal to the the set $\{\boldsymbol{x} \mid p_E(\boldsymbol{x}) \leq \alpha\}$ for all α, which means $\gamma_E(\theta_1, \theta_2; \alpha) \geq \gamma_M(\theta_1, \theta_2 : \alpha)$ for all $(\theta_1, \theta_2) \in \Theta$ and for each fixed $\alpha \in [0, 1]$.

4.10.5 How to numerically perform the M-step

When calculating the realisation of the M p-value for an outcome $\boldsymbol{x}$ with $Z_p^2(\boldsymbol{X})$ as test statistic we need to find the supremum of $\Pr_{(\theta,\theta)}(T(\boldsymbol{X}) \geq T(\boldsymbol{x}))$ over $\theta \in [0, 1]$. There are at least two ways this can be done. The first way is to use some numerical optimisation procedure. If we use R (R Core Team 2014) we can use the function `optim`. Since we want to do restricted maximisation, we should for instance use the method `L-BFGS-B` which allows for box constraints, i.e that the maximum is attained at parameter value that satisfies $0 \leq \theta \leq 1$. The method is a limited memory quasi-Newton optimisation procedure. However, since the function to be optimised as a function of θ is not concave down, there is no reason to expect that the maximum found by the procedure is the maximum on $[0, 1]$ (i.e the function may return a local maximum). We illustrate this when $n_1 = 90, n_2 = 150, x_1 = 60$ and $x_2 = 109$. The profile of $Z_p^2(60, 109)$ is drawn in Figure 4.10. If we set the start value to $\theta = 0.10$, then the function returns the local maximum 0.336 attained at the parameter value 0.0789. However, the true maximum is 0.421 attained at either 0.01299 or 0.987. If we use the start value 0.6, the function returns an error message. We therefore see that even if the function returns a candidate for a maximum value, there is no guarantee that this is the maximum on $[0, 1]$. We have investigated several cases and the above results apply when the profile is not concave down (data not shown).

The second way to compute $\sup_{\theta \in [0,1]} \Pr_\theta(T(\boldsymbol{X}) \geq T(\boldsymbol{x}))$ numerically is to calculate $\Pr_\theta(T(\boldsymbol{X}) \geq T(\boldsymbol{x}))$ at a grid of values of θ over [0,1] and choose the largest of the values calculated. We will use this method when calculating the M p-value.

When investigating the mentioned other cases the profile of $T(\boldsymbol{x})$ appears to be symmetric around $\theta = \frac{1}{2}$. If this holds in general we can restrict the search area for a maximum to either [0,0.5] or to [0.5,1]. Since the M step maintains the ordering of the initial test statistic we know that $\Pr_{(\theta,\theta)}(Z_p^2(\boldsymbol{X}) \geq Z_p^2(\boldsymbol{x})) = \Pr_{(\theta,\theta)}(p_M(\boldsymbol{X}) \leq p_M(\boldsymbol{x}))$ for all outcomes $\boldsymbol{x}$. Since $\Pr_{(\theta,\theta)}(p_M(\boldsymbol{X}) \leq p_M(\boldsymbol{x}))$ is symmetric around $\theta = \frac{1}{2}$, $\Pr_{(\theta,\theta)}(Z_p^2(\boldsymbol{X}) \geq Z_p^2(\boldsymbol{x}))$ is symmetric around $\theta = \frac{1}{2}$. We can therefore restrict the search for the maximum to $[0, \frac{1}{2}]$ when using Z_p^2 as test statistic.

We know from the previous subsection that $\Pr_{(\theta,\theta)}(p(\boldsymbol{X}) \leq \alpha)$ is symmetric around

$\theta = \frac{1}{2}$, where $p(\boldsymbol{X})$ is either the M, E, A or C p-value. This means we can restrict the search area for a maximum to $[0, \frac{1}{2}]$ when calculating $\sup_\theta \Pr_{(\theta,\theta)}(p(\boldsymbol{X}) \leq p(\boldsymbol{x})) = \sup_\theta \Pr_{(\theta,\theta)}(-p(\boldsymbol{X}) \geq -p(\boldsymbol{x}))$, i.e when calculating the M, E $\circ$ M, A $\circ$ M or C $\circ$ M p-value.

Chapter 5

Power study: Testing equality of two binomial proportions

In this chapter we study the power functions of level α tests testing the null hypothesis specified in Equation (4.1) against the alternative hypothesis specified in the same equation. We base the tests on p-values. We know from Section 4.8 that the ordering induced by the A p-value equals the ordering induced by the M p-value except for the largest realisations of the p-values. This means we do not consider the p_{AM}-value since it will be equal to p_M except for the largest realisations of the p-values (which are not important when evaluating the power functions). We also know that the M^2 p-value is equal to the M p-value, so that the M^2 p-value is not considered. We consider the p-values $p_A, p_M, p_{EM}, p_C, p_{CM}$.

We know 0.05 is a commonly used significance level. In genome wide association studies multiple hypotheses are tested and to control the familywise error rate a commonly used significance level for a single test is $5 \cdot 10^{-8}$, see for instance Dudbridge & Gusnanto (2008) or Panagiotou & Ioannidis (2012). We consider the significance levels $5 \cdot 10^{-k}, k = 2, \ldots, 8$ to get a spectrum of levels. We also need to consider specific values of n_1 and n_2. We study both *balanced designs*, where $n_1 = n_2$, and unbalanced designs , where $n_1 \neq n_2$, and use the values considered in Mehrotra et al. (2004, p. 445) in addition to $(97, 103)$, where the latter are the sample sizes in example (d) in Section 4.1.1. The studied configurations are given in Table 5.1. We see that all unbalanced designs are of the form $n_2 = 4n_1$ (the configuration $(97, 103)$ is considered a balanced design since $n_1 \approx n_2$). This is the maximum number of controls per case recommended in the literature, see for instance Wacholder et al. (1992).

We cannot evaluate the power functions in every point of the continuous param-

Table 5.1: The configurations of n_1 and n_2 used in the power study. The first row gives the balanced configurations and the second row gives the unbalanced configurations.

Balanced	$(10, 10)$	$(25, 25)$	$(50, 50)$	$(97, 103)$	$(150, 150)$
Unbalanced	$(4, 16)$	$(10, 40)$	$(20, 80)$	$(60, 240)$	

eter space. We therefore only evaluate the functions in a grid of values. From Section 4.10.2 we know we only need to consider the type I error probabilities, or equivalently the power functions under H_0, for $\theta \in [0, 1/2]$ and from Section 4.5 and Section 4.10.1 we know we only need to consider the power functions under H_1 for $\theta_2 > \theta_1$. We use a grid with different grid increments under H_0 and H_1. Under H_0 we consider the values $0, 0.05, \ldots, 0.45, 0.5$ of θ and under H_1 we use $\theta_2 = \theta_1 + \delta$, where $\delta = 0.01, 0.02, \ldots, 0.99, 1$ and for each fixed value of δ, $\theta_1 = 0, 0.01, 0.02, \ldots, 1 - \delta$. We illustrate the grid in Figure 5.1.

When calculating the realisations of a p-value by using a M step, we know from Section 4.10.5 that we only need to do the maximization over $[0, 0.5]$ and not over the entire interval $[0, 1]$. We use a equispaced grid over $[0, 5]$ with grid increment $5 \cdot 10^{-6}$. The code used in the power study for generating the C and A p-values are given in Appendix C and have been programmed in R (R Core Team 2014). The R implementations of the M and E-values have been replaced with already existing implementations in C++ by Øyvind Bakke that are much faster. For instance the program for calculating the M p-values use parallel computing. We have evaluated the power functions in the mentioned grid using a C++-program made by Øyvind Bakke. We use enumeration both when when calculating the realisations of each p-value and also when evaluating the power functions, see Section 3.2 and Section 3.3.

In the power study we can change n_1, n_2, k, θ_1 and θ_2 (under H_0 we change $\theta = \theta_1 = \theta_2$). We call (θ_1, θ_2) either a point in the parameter space or a parameter point. In the rest of the chapter we discuss the results of the power study, first the results under H_0 and then under H_1.

5.1 Assessing validity of the p-values and simple comparison of the power functions under H_0

In this section we check validity of the p-values and also take a first look at the size of the tests and do a simple comparison of the power functions. We start

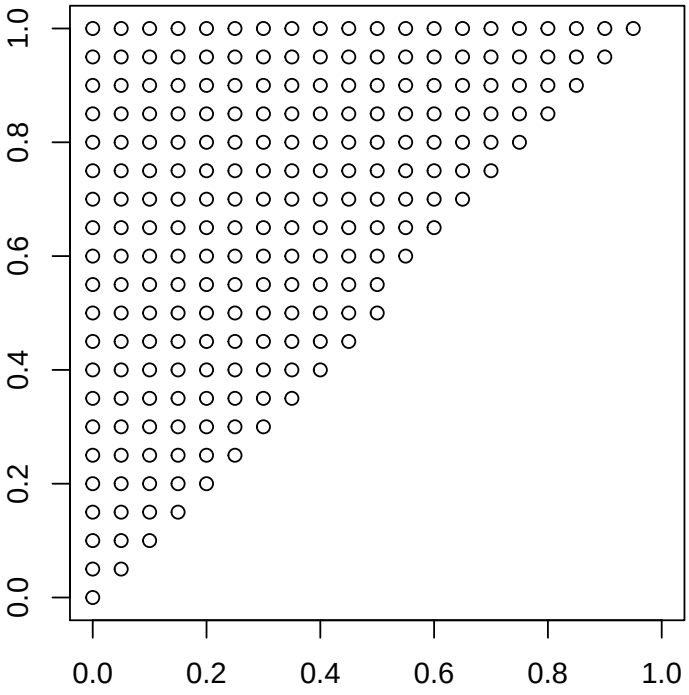


Figure 5.1: Illustration of the grid used in the power calculations, where θ_2 is along the ordinate and θ_1 is along the abcissa. There are extra points under H_1 not shown.

with considering validity of the different p-values, which is the main topic of this section. In Table D.1 in Appendix D we show the type I error probabilities at selected values of θ and n_1, n_2, k. From this table we observe that the A and E p-values in general are not valid, as noted in Section 4.4.2. The maximum observed type I error probabilities are highest for unbalanced designs. The A p-value is in general found to be more liberal than the E p-value, but there are cases where E is liberal and A is not, for instance when $n_1 = n_2 = 150$ and $k = 5, 6$. The highest observed value of the type I error probabilities for the A p-value is 24 times higher the significance level and is observed when $n_1 = 60, n_2 = 240, k = 8$ and $\theta = 0.05$. It is important to notice that the type I error probability in this case is $120.54 \cdot 10^{-8} = 1.2054 \cdot 10^{-6}$ and not 120.54%, which of course would be impossible. The type I error probabilities of the E p-value are never found to be higher than 1.12 times the significance level (which occurs when $n_1 = 97, n_2 = 103, k = 6, \theta = 0.15$).

From Table D.1 we also observe that $\gamma_E(\theta) \geq \gamma_M(\theta)$, as explained in Section 4.10.4. As pointed out in Section 4.10.3, either (1) $\gamma_A(\theta; \alpha) \geq \gamma_M(\theta; \alpha)$, (2) $\gamma_A(\theta; \alpha) \leq \gamma_M(\theta; \alpha)$ or (3) $\gamma_A(\theta; \alpha) = \gamma_M(\theta; \alpha)$ for all $\theta \in [0, 1]$ for a specific choice of α, n_1 and n_2. All three situations occur in Table D.1. For instance situation (1) occurs when $n_1 = n_2 = 25, k = 2$, situation (2) occurs when $n_1 = n_2 = 50, k = 8$ and situation (3) occurs when $n_1 = 4, n_2 = 16, k = 5$. We also know from Theorem 4.8.1 that $\gamma_C(\theta; \alpha) \leq \gamma_{CM}(\theta; \alpha)$ for all θ and all choices of n_1, n_2, α. For instance we see that $\gamma_{CM}(\theta; \alpha)$ is several times $\gamma_C(\theta; \alpha)$ for many values of θ when $n_1 = n_2 = 10$ for the studied values of k in the table and also when $n_1 = 4, n_2 = 16, k = 5$. In many of the remaining cases we also observe a power increase. In several of these cases the power increment is roughly about $1 \cdot 10^{-k}$, but the power increase can be both lower and higher than this number.

As expected, we observe from Table D.1 that $p_C(\boldsymbol{X}), p_{CM}(\boldsymbol{X}), p_{EM}(\boldsymbol{X})$ and $p_M(\boldsymbol{X})$ are valid. We know this holds in general from Section 4.6 and Theorem 4.8.1. The reason we observe this from Table D.1 is that this table shows every situation where at least one of the power functions exceeds the significance level on the specified grid and no case where either of $\gamma_C, \gamma_M, \gamma_{CM}$ or γ_{EM} exceeds the significance level is shown.

From Table D.1 we observe that the test sizes are in general going to be closer to α when n_1, n_2 increases for the valid p-values. (Note: the entries in the table are based on a sparse grid of θ on $[0, 1]$, so that the largest value we calculate for a specific α, n_1 and n_2 is only a lower bound on the test size.) We now try to explain why. We consider a single binomial experiment Y, i.e $Y \sim \text{Binom}(\theta, n)$ and $\theta \in (0, 1)$. The results will generalize to our experiment with two independent binomial experiments. We know the mode(s) of the pmf of Y is close to the expected value

of Y for all n. The probability of the mode(s) must become smaller and smaller as n increases. The reasons are as n increases we (1) get more outcomes with positive probability (when $\theta \in (0,1)$), (2) must still have $\sum \Pr_\theta(Y = y) = 1$ and (3) $\text{Var}(Y) = n\theta(1-\theta)$ increases. If for instance the probability of the mode was increasing and the pmf became more and more concentrated around the mode(s), then the variance would need to be decreasing with increasing n to get a behaviour similar to that of a random variable whose limiting distribution is a constant in the mode(s) and is zero for rest of the outcomes. The shape of the pmf cannot remain unchanged as we get more outcomes with positive probability. In total, the probability of the mode(s) is decreasing with increasing n for $\theta \in (0,1)$.

As n increases, (1) we get more outcomes y where $\Pr_\theta(Y = y) > 0$ ($\theta \in (0,1)$), $\sum \Pr_\theta(Y = y) = 1$, (2) the probability of the mode decreases and (3) $\text{Var}(Y)$ increases. Due to the mentioned facts, the values of the probabilities of the outcomes must in general become smaller as n increases. Note that we do not compare the probabilities outcome-wise, i.e we do not compare the probabilities of each fixed outcome when n increases. This would be much harder as more outcomes are possible with increasing n and as the mode and expected value changes with increasing n. Since the values of the probabilities of the different outcomes in general become lower with increasing n and the test statistic takes more unique values with increasing n, there are greater chances of (1) observing lower values of the p-value, (2) observing that the distances between the realisations of the p-value decrease and (3) observing that the p-value takes more unique values. This also holds when testing equality of two independent binomial proportions and n_1 and n_2 increase. The mentioned facts also imply that that there are greater chances of getting a test size closer to α as n_1, n_2 increases when considering the level α test based on the p-value. This is perhaps easiest to understand when considering the M p-value since $Z_p^2(\boldsymbol{X})$ and $-p_M(\boldsymbol{X})$ where Z_p^2 is used as test statistic order the sample space the same. Since $-p_M(\boldsymbol{X})$ and $Z_p^2(\boldsymbol{X})$ order the sample space the same we must have that $\sup_{\theta\in[0,1]} \Pr_\theta(p(\boldsymbol{X}) \leq p(\boldsymbol{x})) = \sup_{\theta\in[0,1]} \Pr_\theta(Z_p^2(\boldsymbol{X}) \geq Z_p^2(\boldsymbol{x}))$. Since $\sup_{\theta\in[0,1]} \Pr_\theta(Z_p^2(\boldsymbol{X}) \geq Z_p^2(\boldsymbol{x})) = p(\boldsymbol{x})$, $\sup_{\theta\in[0,1]} \Pr_\theta(p(\boldsymbol{X}) \leq p(\boldsymbol{x})) = p(\boldsymbol{x})$. We know the chances of $p_M(\boldsymbol{X})$ taking more unique values on the whole of $[0,1]$ are increasing with increasing n_1, n_2, meaning the chances of $p_M(\boldsymbol{X})$ taking values closer and closer to a given α increase as n_1, n_2 increase, so that the chances of the test size of the level α test based on $p_M(\boldsymbol{X})$ being closer to a given significance level α increase.

Comparing the power functions of the level α tests based on the remaining three valid p-values, i.e γ_{CM}, γ_{EM} and γ_M, is much harder than comparing γ_C with γ_{CM}. They are in general more similar than γ_C and γ_{CM}. None of the power functions takes the largest in all parameter points under H_0 when considering all of the

studied sample sizes and significance levels in Table D.1. Since γ_{CM}, γ_{EM} and γ_M are similar under H_0, we get similar results when comparing either of γ_{EM} or γ_M with γ_C as when comparing γ_{CM} with γ_C.

It is important to realize that we compare the power functions in the previous explanations and each entry of Table D.1 in the same parameter point. One could for instance only compare test sizes, i.e compare $\sup_{\theta\in[0,1]} \gamma_i(\theta;\alpha)$ and $\sup_{\theta\in[0,1]} \gamma_j(\theta;\alpha)$ for different i and j. However, this can be problematic as the maximum value of $\gamma_i(\theta;\alpha)$ may occur in another point than the maximum value of $\gamma_j(\theta_1;\alpha)$. Only one value of θ on $[0,1]$ is possible under H_0, which means we should compare power functions in the same point. Only comparing test sizes also means we only do one comparison of the two power functions in possibly two different points (we do one comparison, where we compare $\gamma_i(\theta_1;\alpha)$ with $\gamma_j(\theta_2;\alpha)$ where possibly $\theta_1 \neq \theta_2$).

5.2 Comparing the power functions under the alternative hypothesis

As mentioned in Section 2.7 it is common practise to only consider sample size(s) large enough so that the power of the test is at least 80 % at $\boldsymbol{\theta}$ that are scientifically meaningful under the alternative hypothesis. When comparing the power functions under the alternative hypothesis we compare two power functions at the same (θ_1, θ_2). We consider all possible pairwise comparisons and compare power functions at points at which at least one of the tests has power 80 % or greater. This also means we compare the power functions in all possible pairs of the studied power functions in the same set of points, even if both tests in one pair have power lower than 80 % at some of the studied points. The reason is simply that performing all the pairwise comparisons is equivalent to comparing all of the studied power functions at once. Since we consider all points at which one power function is at least 80, we should compare all of the power functions in these points.

We compare the power functions in different ways. In Section 5.2.1 we divide the differences into four intervals for each value of n_1, n_2, k and calculate the proportion of points giving the differences in each interval. In Section 5.2.2 we study plots of the power functions and also study plots telling where in parameter space the different differences occur. Finally in Section 5.2.3 we look at the distribution of the differences between the power functions in each pair of power functions.

5.2.1 Considering the differences of the power functions in four intervals

One reasonable question when comparing two power functions in a point is when the two power functions should be considered equal. If the two power functions are exactly the same, the two power functions should of course be considered equal in that point. If the difference between the two power functions is small we should also regard the two power functions as equal. The next question is then what "small" means. Since we only study points at which one power functions is above 80, one possible choice of small is 2 %. This means we regard two power functions whose absolute difference is smaller than 2 pp. in a point as equal in that point.

When comparing two power functions the differences must be on the interval $(-\infty, \infty)$. We divide this interval into the four disjoint intervals $(-\infty, -2]$, $(-2, 0)$, $[0, 2)$, $[2, \infty)$. We have chosen to split the interval $(-2, 2)$, where the two power functions are regarded as equal, into two intervals. The reason is that $\gamma_{CM}(\theta_1, \theta_2; \alpha) - \gamma_C(\theta_1, \theta_2; \alpha) \geq 0$, so that no points should give a negative difference. Afterwards, after classifying which of four intervals the difference of the two power functions belongs to for each considered point, we calculate the percentage of points classified into each of the intervals. We do this for all pairs of power functions and each considered value of n_1, n_2 and k. The results are presented in Table 5.2. We also show the number of points considered in each case in this table. If the number of points considered for a specific choice of n_1, n_2 and k is less than 50 we do not consider comparing the tests since only a few points give power that is 80 % or greater. This will be points close to $(\theta_1, \theta_2) = (0, 1)$. The percentage numbers in these cases are included for completeness.

EM, CM and M vs. C We observe that $\gamma_{CM}, \gamma_{EM}, \gamma_M$ are much larger than γ_C for the small sample sizes $(10, 10)$ and $(4, 16)$, i.e. both for balanced and unbalanced designs. The power increase is also large for the sample sizes $(25, 25)$ and $(50, 50)$, but not as large as for the smallest sample sizes. For unbalanced designs γ_{CM} and γ_{EM} are also much larger than γ_C when $(n_1, n_2) = (10, 40)$ and $(n_1, n_2) = (20, 80)$, but again not as great as for the smallest sample sizes. Actually γ_C is larger than γ_M in a relatively large percentage of points when $(n_1, n_2) = (10, 40), k = 6$, $(n_1, n_2) = (20, 80), k = 3, \ldots, 8$ and for $(n_1, n_2) = (60, 240), k = 4, \ldots, 8$. We observe that all studied power functions are very similar in the large balanced designs $(n_1, n_2) = (97, 103)$ and $(n_1, n_2) = (150, 150)$. We have that γ_{EM}, γ_{CM} and γ_C are very similar and larger than γ_M in the large unbalanced designs $(n_{1,2}) = (20, 80), (n_1, n_2) = (60, 240)$, with one exception for

the last mentioned design when $k = 2$. In the mentioned exception all studied power functions are very similar.

EM and CM vs. M For balanced designs γ_{CM}, γ_{EM} and γ_M are very similar. Three notable exceptions are $(n_1, n_2) = (25, 25), k = 6$ and $(n_1, n_2) = (50, 50), k = 7, 8$, where γ_{CM}, γ_{EM} are larger than γ_M in a relatively large percentage of points. For unbalanced designs γ_{CM} and γ_{EM} are larger than γ_M except when $(n_1, n_2) = (10, 40), k = 2, 8$, $(n_1, n_2) = (20, 80), k = 2$ and $(n_1, n_2) = (60, 240), k = 2, 3$. When $(n_1, n_2) = (10, 40), k = 8$ the proportion of points where γ_M is larger than γ_{CM} is larger than the proportion of points where γ_{CM} is larger than γ_M. In this case γ_M and γ_{EM} are very similar.

EM vs. CM For balanced designs γ_{CM} and γ_{EM} are very similar. For unbalanced designs they are also very similar, but there are some exceptions. When $(n_1, n_2) = (10, 40)$ and $k = 2, 3, 8$ γ_{EM} is larger than γ_{CM} (when $k = 8$ there are also some points where γ_{CM} wins over γ_{EM}, but γ_{EM} wins in a greater proportion of points). This also holds for $(n_1, n_2) = (20, 80), k = 3$. When $(n_1, n_2) = (10, 40), k = 7$, γ_{CM} is larger than γ_{EM} in 14 % of the points.

Table 5.2: Pairwise comparisons of the power functions. In column 4 to column 9 we compare all possible combinations of the power functions. For instance "M vs. C" means we consider all of the differences $\gamma_M(\theta_1, \theta_2; \alpha) - \gamma_C(\theta_1, \theta_2; \alpha)$ on the studied grid of (θ_1, θ_2) under the alternative hypothesis. We divide the differences in the intervals $(-\infty, -2], (-2, 0), [0, 2), [2, \infty)$. For each column there are four values specifying the percentage of points (θ_1, θ_2) under H_1 that give power differences in the four mentioned intervals. The *leftmost* entry gives the percentage of points that give power differences in the interval $[2, \infty)$ and the *rightmost* entry gives the percentage of points where the power differences are in the interval $(-\infty, -2]$. We only consider differences in power for points where the power of at least one power function is at least 80 %. For each row column 1 gives the sample sizes in the studied situation, column 2 gives the significance level specified as $5 \cdot 10^{-k}$ and column 4 gives the number of points we compare the power functions, i.e the number of points where at least one power function is above or equal to 80 %. We compare all of the power functions in the same points. The symbol $-$ means the value in the entry is exactly equal to 0 and the symbol $\times$ means the entry has not been calculated.

(n_1, n_2)	k	No. points	EM vs. C				CM vs C				M vs. C				EM vs. M				CM vs. M				EM vs. CM			
			$(\infty, 2]$	$(2, 0]$	$(0, -2)$	$[-2, -\infty)$	$(\infty, 2]$	$(2, 0]$	$(0, -2)$	$[-2, -\infty)$	$(\infty, 2]$	$(2, 0]$	$(0, -2)$	$[-2, -\infty)$	$(\infty, 2]$	$(2, 0]$	$(0, -2)$	$[-2, -\infty)$	$(\infty, 2]$	$(2, 0]$	$(0, -2)$	$[-2, -\infty)$	$(\infty, 2]$	$(2, 0]$	$(0, -2)$	$[-2, -\infty)$
(10,10)	2	977	0.778	0.222	–	–	0.778	0.222	–	–	0.778	0.222	–	–	–	1.000	–	–	–	1.000	–	–	–	1.000	–	–
	3	384	0.880	0.120	–	–	0.880	0.120	–	–	0.880	0.120	–	–	–	1.000	–	–	–	1.000	–	–	–	1.000	–	–
	4	136	0.956	0.044	–	–	0.956	0.044	–	–	0.956	0.044	–	–	–	1.000	–	–	–	1.000	–	–	–	1.000	–	–
	5	45	0.978	0.022	–	–	0.978	0.022	–	–	0.978	0.022	–	–	–	1.000	–	–	–	1.000	–	–	–	1.000	–	–
	6	6	1.000	–	–	–	1.000	–	–	–	1.000	–	–	–	–	1.000	–	–	–	1.000	–	–	–	1.000	–	–
	7	0	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
	8	0	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
(25,25)	2	2160	0.278	0.722	–	–	0.275	0.725	–	–	0.278	0.722	–	–	–	1.000	–	–	–	0.658	0.315	0.027	0.027	0.973	–	–
	3	1531	0.400	0.600	–	–	0.384	0.616	–	–	0.384	0.616	–	–	0.052	0.948	–	–	–	1.000	–	–	0.052	0.948	–	–
	4	1081	0.498	0.502	–	–	0.496	0.504	–	–	0.496	0.504	–	–	0.050	0.950	–	–	–	1.000	–	–	0.050	0.950	–	–
	5	745	0.370	0.630	–	–	0.370	0.630	–	–	0.325	0.675	–	–	0.083	0.917	–	–	0.083	0.917	–	–	–	1.000	–	–
	6	512	0.518	0.482	–	–	0.518	0.482	–	–	0.346	0.654	–	–	0.223	0.777	–	–	0.223	0.777	–	–	–	1.000	–	–
	7	362	0.494	0.506	–	–	0.494	0.506	–	–	0.494	0.506	–	–	–	1.000	–	–	–	1.000	–	–	–	1.000	–	–
	8	242	0.698	0.302	–	–	0.698	0.302	–	–	0.698	0.302	–	–	–	1.000	–	–	–	1.000	–	–	–	1.000	–	–
(50,50)	2	2966	0.177	0.823	–	–	0.152	0.848	–	–	0.169	0.831	–	–	0.005	0.995	–	–	–	0.463	0.526	0.011	0.034	0.966	–	–
	3	2394	0.170	0.830	–	–	0.161	0.839	–	–	0.161	0.839	–	–	0.009	0.991	–	–	–	1.000	–	–	0.009	0.991	–	–
	4	1987	0.249	0.751	–	–	0.246	0.754	–	–	0.246	0.754	–	–	0.011	0.989	–	–	–	1.000	–	–	0.011	0.989	–	–
	5	1650	0.207	0.793	–	–	0.207	0.793	–	–	0.161	0.839	–	–	0.038	0.962	–	–	0.038	0.962	–	–	–	1.000	–	–
	6	1400	0.301	0.699	–	–	0.300	0.700	–	–	0.276	0.724	–	–	0.024	0.976	–	–	0.021	0.979	–	–	0.016	0.984	–	–
	7	1178	0.277	0.723	–	–	0.209	0.791	–	–	0.121	0.705	0.115	0.059	0.166	0.834	–	–	0.109	0.891	–	–	0.063	0.937	–	–
	8	988	0.308	0.692	–	–	0.273	0.727	–	–	0.184	0.713	0.071	0.032	0.128	0.872	–	–	0.085	0.915	–	–	0.036	0.964	–	–
(97, 103)	2	3571	0.017	0.907	0.076	–	0.010	0.990	–	–	0.025	0.975	–	–	0.001	0.606	0.392	0.001	–	0.715	0.279	0.005	0.003	0.707	0.290	–
	3	3135	0.025	0.975	–	–	0.015	0.985	–	–	0.015	0.985	–	–	0.003	0.997	–	–	0.001	0.728	0.270	–	0.004	0.830	0.166	–
	4	2800	0.036	0.964	–	–	0.012	0.988	–	–	0.009	0.956	0.027	0.007	0.018	0.983	–	–	0.012	0.871	0.116	–	0.005	0.838	0.157	–
	5	2520	0.014	0.986	–	–	0.010	0.990	–	–	0.012	0.919	0.058	0.012	0.023	0.813	0.164	–	0.014	0.878	0.108	–	0.007	0.704	0.289	–
	6	2289	0.031	0.969	–	–	0.005	0.995	–	–	0.009	0.929	0.047	0.015	0.027	0.818	0.155	–	0.020	0.652	0.328	–	0.006	0.948	0.047	–
	7	2091	0.050	0.950	–	–	0.022	0.978	–	–	0.022	0.891	0.068	0.019	0.034	0.829	0.137	–	0.030	0.682	0.287	–	0.005	0.995	–	–
	8	1907	0.033	0.967	–	–	0.016	0.984	–	–	0.010	0.830	0.133	0.027	0.044	0.793	0.163	–	0.042	0.826	0.131	–	0.008	0.714	0.278	–
(150, 150)	2	3836	0.044	0.956	–	–	0.034	0.966	–	–	0.046	0.954	–	–	–	0.865	0.135	–	–	0.776	0.217	0.007	0.007	0.993	–	–
	3	3470	0.029	0.971	–	–	0.054	0.946	–	–	0.061	0.939	–	–	–	0.620	0.380	–	–	0.823	0.175	0.002	0.002	0.655	0.342	–
	4	3192	0.049	0.951	–	–	0.042	0.958	–	–	0.046	0.954	–	–	0.004	0.726	0.270	–	–	0.806	0.194	–	0.004	0.837	0.159	–
	5	2952	0.058	0.942	–	–	0.042	0.958	–	–	0.035	0.946	0.016	0.003	0.010	0.990	–	–	0.005	0.995	–	–	0.007	0.993	–	–
	6	2755	0.070	0.930	–	–	0.061	0.939	–	–	0.050	0.926	0.020	0.004	0.015	0.985	–	–	0.012	0.988	–	–	0.003	0.997	–	–
	7	2568	0.058	0.942	–	–	0.065	0.935	–	–	0.042	0.912	0.037	0.009	0.022	0.741	0.237	–	0.019	0.981	–	–	0.005	0.627	0.367	–
	8	2415	0.069	0.931	–	–	0.056	0.944	–	–	0.027	0.917	0.046	0.011	0.028	0.972	–	–	0.022	0.978	–	–	0.005	0.995	–	–
(4, 16)	2	642	0.782	0.218	–	–	0.782	0.218	–	–	0.782	0.218	–	–	–	1.000	–	–	–	1.000	–	–	–	1.000	–	–
	3	194	0.845	0.155	–	–	0.845	0.155	–	–	0.845	0.155	–	–	–	1.000	–	–	–	1.000	–	–	–	1.000	–	–
	4	25	0.760	0.240	–	–	0.760	0.240	–	–	0.760	0.240	–	–	–	1.000	–	–	–	1.000	–	–	–	1.000	–	–
	5	8	1.000	–	–	–	1.000	–	–	–	1.000	–	–	–	–	1.000	–	–	–	1.000	–	–	–	1.000	–	–
	6	0	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
	7	0	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
	8	0	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
(10, 40)	2	1731	0.357	0.643	–	–	0.270	0.730	–	–	0.325	0.675	–	–	0.030	0.970	–	–	0.007	0.500	0.431	0.062	0.111	0.588	0.301	–
	3	1064	0.444	0.556	–	–	0.233	0.767	–	–	0.197	0.803	–	–	0.220	0.780	–	–	0.075	0.595	0.268	0.062	0.208	0.792	–	–
	4	653	0.613	0.387	–	–	0.613	0.387	–	–	0.358	0.642	–	–	0.254	0.746	–	–	0.254	0.746	–	–	–	1.000	–	–
	5	414	0.676	0.324	–	–	0.676	0.324	–	–	0.483	0.338	0.080	0.099	0.222	0.778	–	–	0.222	0.778	–	–	–	1.000	–	–
	6	241	0.763	0.237	–	–	0.755	0.245	–	–	0.602	0.232	0.058	0.108	0.195	0.805	–	–	0.187	0.813	–	–	0.075	0.925	–	–
	7	129	0.845	0.155	–	–	0.845	0.155	–	–	0.744	0.256	–	–	0.209	0.791	–	–	0.233	0.767	–	–	–	0.333	0.527	0.140
	8	60	0.783	0.217	–	–	0.833	0.167	–	–	0.783	0.217	–	–	–	1.000	–	–	0.150	0.317	0.200	0.333	0.333	0.467	0.050	0.150
(20,80)	2	2535	0.188	0.812	–	–	0.116	0.884	–	–	0.136	0.864	–	–	0.025	0.975	–	–	0.004	0.472	0.512	0.012	0.038	0.882	0.080	–
	3	1916	0.261	0.739	–	–	0.125	0.875	–	–	0.072	0.826	0.092	0.009	0.185	0.815	–	–	0.103	0.673	0.174	0.050	0.108	0.829	0.063	–
	4	1484	0.296	0.704	–	–	0.199	0.801	–	–	0.028	0.212	0.383	0.377	0.436	0.564	–	–	0.401	0.521	0.063	0.015	0.061	0.883	0.047	0.009
	5	1167	0.332	0.668	–	–	0.238	0.762	–	–	0.034	0.323	0.335	0.308	0.410	0.590	–	–	0.362	0.529	0.092	0.018	0.061	0.841	0.069	0.028
	6	910	0.364	0.636	–	–	0.320	0.680	–	–	0.086	0.340	0.281	0.293	0.385	0.615	–	–	0.367	0.482	0.140	0.011	0.052	0.892	0.038	0.018
	7	718	0.368	0.632	–	–	0.389	0.611	–	–	0.102	0.354	0.276	0.269	0.373	0.627	–	–	0.373	0.627	–	–	–	0.801	0.177	0.022

Continued on next page

(n_1, n_2)	k	No. points	EM vs. C				CM vs C				M vs. C				EM vs. M				CM vs. M				EM vs. CM			
			$(\infty, 2]$	$(2, 0]$	$(0, -2)$	$[-2, -\infty)$	$(\infty, 2]$	$(2, 0]$	$(0, -2)$	$[-2, -\infty)$	$(\infty, 2]$	$(2, 0]$	$(0, -2)$	$[-2, -\infty)$	$(\infty, 2]$	$(2, 0]$	$(0, -2)$	$[-2, -\infty)$	$(\infty, 2]$	$(2, 0]$	$(0, -2)$	$[-2, -\infty)$	$(\infty, 2]$	$(2, 0]$	$(0, -2)$	$[-2, -\infty)$
	8	555	0.342	0.658	–	–	0.416	0.584	–	–	0.162	0.317	0.213	0.308	0.351	0.400	0.236	0.013	0.360	0.400	0.231	0.009	–	0.247	0.668	0.085
(60, 240)	2	3538	0.037	0.963	–	–	0.018	0.982	–	–	0.017	0.983	–	–	0.001	0.999	–	–	–	0.730	0.269	0.001	0.003	0.934	0.062	–
	3	3110	0.069	0.931	–	–	0.022	0.978	–	–	0.009	0.718	0.271	0.001	0.075	0.925	–	–	0.039	0.911	0.046	0.004	0.007	0.964	0.029	–
	4	2767	0.057	0.943	–	–	0.031	0.969	–	–	0.005	0.449	0.349	0.197	0.217	0.783	–	–	0.211	0.770	0.015	0.004	0.007	0.908	0.083	0.002
	5	2483	0.055	0.945	–	–	0.043	0.957	–	–	0.009	0.450	0.340	0.200	0.219	0.781	–	–	0.214	0.767	0.019	–	0.001	0.903	0.092	0.004
	6	2246	0.073	0.927	–	–	0.021	0.979	–	–	0.005	0.396	0.341	0.257	0.273	0.727	–	–	0.261	0.710	0.024	0.005	0.020	0.889	0.089	0.003
	7	2046	0.077	0.923	–	–	0.073	0.927	–	–	0.011	0.435	0.319	0.236	0.253	0.734	0.014	–	0.251	0.709	0.035	0.005	0.004	0.885	0.104	0.008
	8	1869	0.061	0.932	0.007	–	0.094	0.906	–	–	0.012	0.409	0.329	0.250	0.266	0.734	–	–	0.263	0.719	0.017	0.002	0.002	0.597	0.378	0.023

5.2.2 Plots of four different intervals in parameter space and plots of the power functions

When finding differences between two power functions in a large proportion of points it is also important to establish where in the parameter space the power functions are different. If they are different in points that are unlikely values of the success probabilities (θ_1, θ_2) under the alternative hypothesis, the differences are most likely of little importance. In this subsection we take a closer look at some of the plots of the power functions and what we call difference plots. In the difference plots we plot which points gives power differences in each of the four intervals for each possible pair of power functions considered. In the difference plots we divide the interval $(-\infty, \infty)$ into almost the same intervals as in Section 5.2.1, the only difference is the intervals $(-\infty, 2], [2, \infty)$. For instance if the minimum of the differences is below -2, we replace $(-\infty, 2]$ with $(\min(\gamma_i - \gamma_j), -2]$ and if the minimum is above -2 we replace $(-\infty, -2]$ with $(-4, -2]$. The reason is that we want to use the same colouring scheme for the points in the different plots. When comparing the difference plots across n_1, n_2, k is is important to realize that 10 % of the points in one figure is not equal to 10 % in another figure if the number of points where the power of the tests is 80 % or greater differ. The number corresponding to each figure can be found in Table 5.2

$\boldsymbol{(n_1, n_2) = (10, 10), k = 2}$ In Figure 5.2 we have plotted the power functions when $(n_1, n_2) = (10, 10)$ and $k = 2$, and in Figure 5.3 we have plotted the differences between the power functions in each possible combination of power functions. We see from Figure 5.2 that the power is highest in the region close to $(0, 1)$, which makes intuitively sense as it should be easiest to tell these success probabilities from one another. We see that the power increase of γ_{CM}, γ_{EM} and γ_M compared with γ_C is in the region of the studied points that most likely is of most practical importance.

$\boldsymbol{(n_1, n_2) = (10, 40), k = 3}$ Figure 5.4 shows the plots of the power functions, and Figure 5.5 shows the difference plots. We observe from Figure 5.5 that γ_{EM} is larger than γ_C in the region of the studied points that most likely is of most practical significance. The same also holds when comparing γ_M with γ_C and γ_{EM} with γ_M, but the regions are smaller compared to the region first considered. When comparing γ_{CM} with γ_C and γ_{EM} with γ_{CM} we observe that the points where the first mentioned power functions are larger than the last mentioned power function are divided into several regions. The regions are still in the part of interesting points that is most likely of most practical significance. The points where γ_{CM} is

larger than γ_M are also divided into two regions. The total mass of the regions is smaller than in the other plots. They are in the interesting part of the points considered. We also observe that there is a region where γ_M is larger than γ_{CM}, which is of size almost equal to the regions where γ_{CM} is larger than γ_M and which is also in the interesting region of points considered.

$\boldsymbol{(n_1, n_2) = (97, 103), k = 2}$ The power functions are plotted in Figure 5.6, and in Figure 5.7 the difference plots are plotted. We see that the power is greater or equal to 80 % in a larger proportion of the points for the different power functions compared with the same plots with smaller sample sizes. This makes intuitively sense. With larger sample sizes it should be easier to tell smaller success probabilities from one another compared to when the sample sizes are smaller. In Figure 5.7 we observe that there are only small differences between the power functions.

$\boldsymbol{(n_1, n_2) = (60, 240), k = 8}$ Figure 5.8 shows the plots of the power functions and Figure 5.9 shows the difference plots. We observe in Figure 5.9 that the regions where γ_C is larger than γ_M, γ_{EM} is larger than γ_M and γ_{CM} is larger than γ_M are in the region of studied of points that most likely is of most scientific importance. The differences between the remaining power functions are relatively small.

In total, we observe that there are differences between some of the power functions in regions which can be of practical importance. It therefore makes sense to try to obtain more precise statements about the differences between the power functions, which we try to do in the subsequent two subsections.

5.2.3 Plots of the empirical cumulative difference functions

One drawback with only knowing the proportions of points in each of the intervals $(-\infty, -2], (-2, 0), [0, 2), [2, \infty)$ is that we do not know the distributions of the differences. For instance when $n_1 = n_1 = 10, k = 2$ and we consider $\gamma_{EM} - \gamma_{CM}$ the fraction of points in the interval $[0, 2)$ is 1 from Table 5.2 , but we do not know for instance if the two power functions are exactly equal for most of the points or if most differences are close to 2%. The empirical cumulative function may then be of use. This function returns the proportion of points equal to or below the point in which it is evaluated. For instance if we evaluate the function in 0 when considering the differences of two power functions, the function returns the proportion of points where the differences between the two power functions are equal to or smaller than 0.

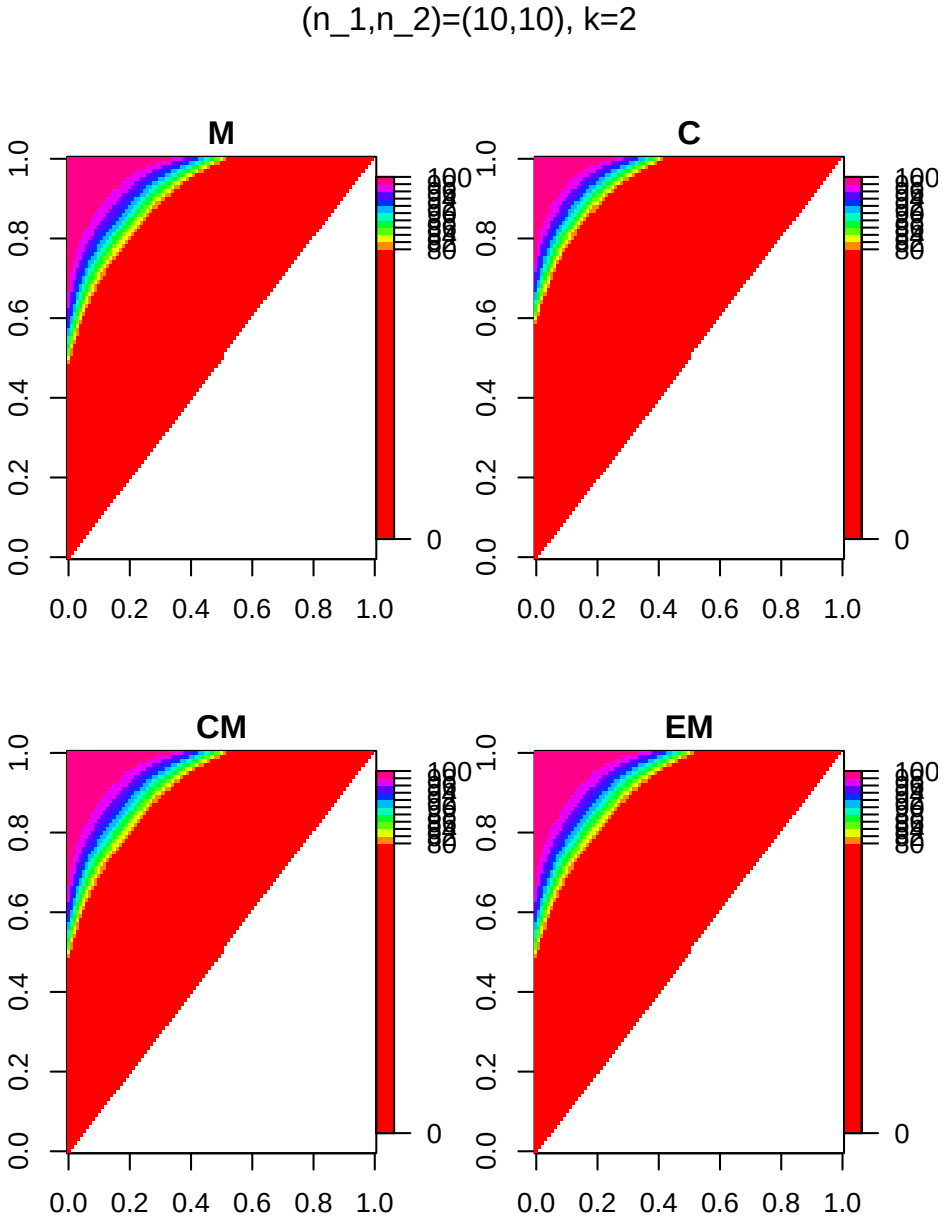


Figure 5.2: Plots of the different power functions when $n_1 = n_2 = 10$ and $\alpha = 5 \cdot 10^{-2}$ ($k = 2$). In the title of each panel is the name of the power function shown. For instance "EM" means $\gamma_{EM}(\theta_1, \theta_2)$ is illustrated. The power function is only evaluated in the grid specified at the beginning of this chapter. To the right of each panel there is a legend specifying the different colors used for the different intervals of values in making the plot. The different cut points are $80, 82, 94, \ldots, 98, 100$. The θ_1-axis is along the abscissa and the θ_2-axis is along the ordinate of each panel.

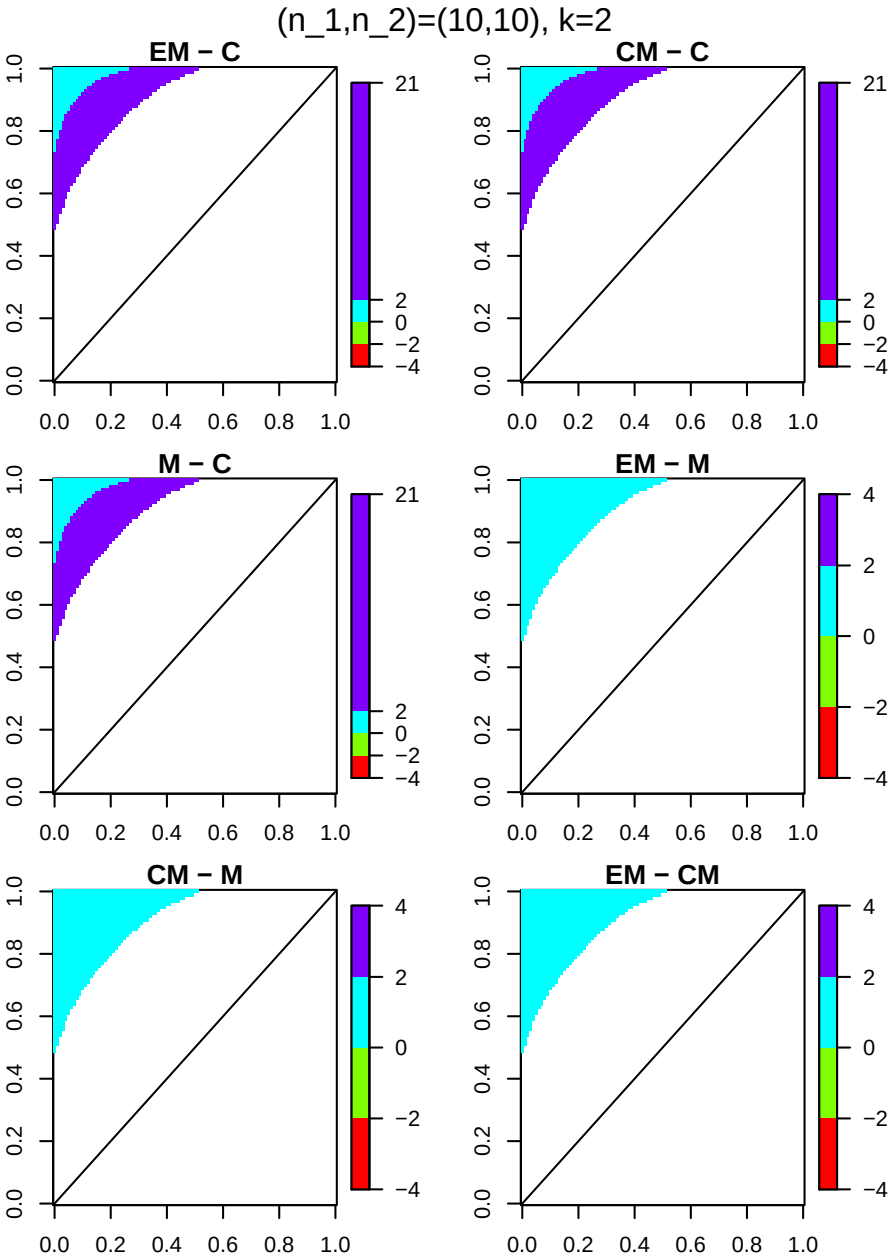


Figure 5.3: Plots of points in the parameter space that gives differences between to power functions in four considered intervals when $n_1 = n_2 = 10$ and $k = 2$. In each panel we consider the differences between two power functions. For instance "EM vs C" means we consider the differences $\gamma_{CM} - \gamma_C$ and consider which points give differences in the four intervals considered. We always consider four intervals, even if there are no points in some of them. This is to ensure that the same color is used for the equivalent intervals for different pairs of power functions. The θ_1-axis is along the abscissa and the θ_2-axis is along the ordinate. We only consider grid points in which the power of at least one of all power functions is 80 % or greater and where the grid is specified at the beginning of this chapter. To the right of each panel there is a legend specifying which colors have been used for points that give differences in the four intervals of differences considered when creating the plot.

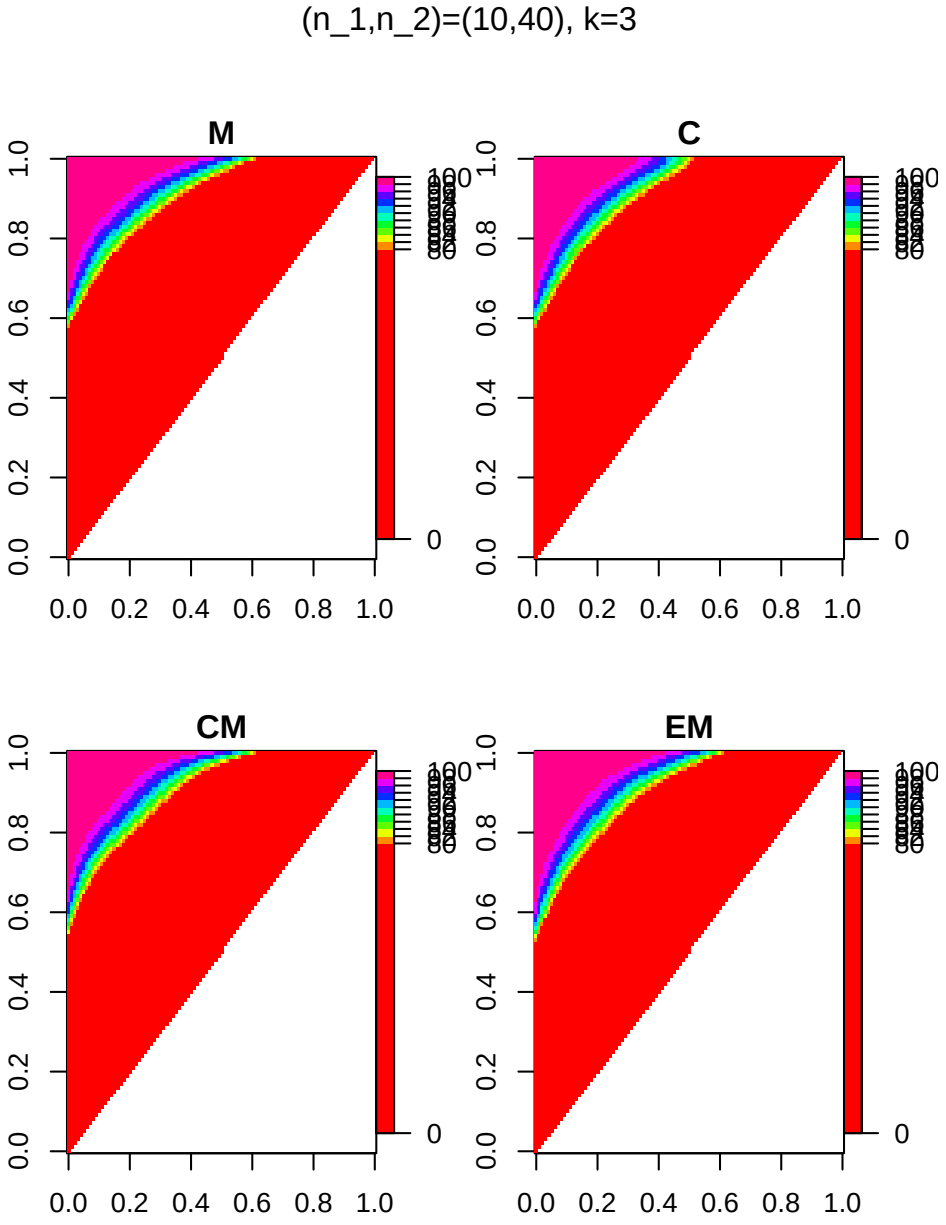


Figure 5.4: Plots of the different power functions when $n_1 = 10, n_2 = 40$ and $\alpha = 5 \cdot 10^{-3}$ ($k = 3$). In the title of each panel is the name of the power function shown. For instance "EM" means $\gamma_{EM}(\theta_1, \theta_2)$ is illustrated. The power function is only evaluated in the grid specified at the beginning of this chapter. To the right of each panel there is a legend specifying the different colors used for the different intervals of values in making the plot. The different cut points are $80, 82, 94, \ldots, 98, 100$. The θ_1-axis is along the abscissa and the θ_2-axis is along the ordinate of each suplot.

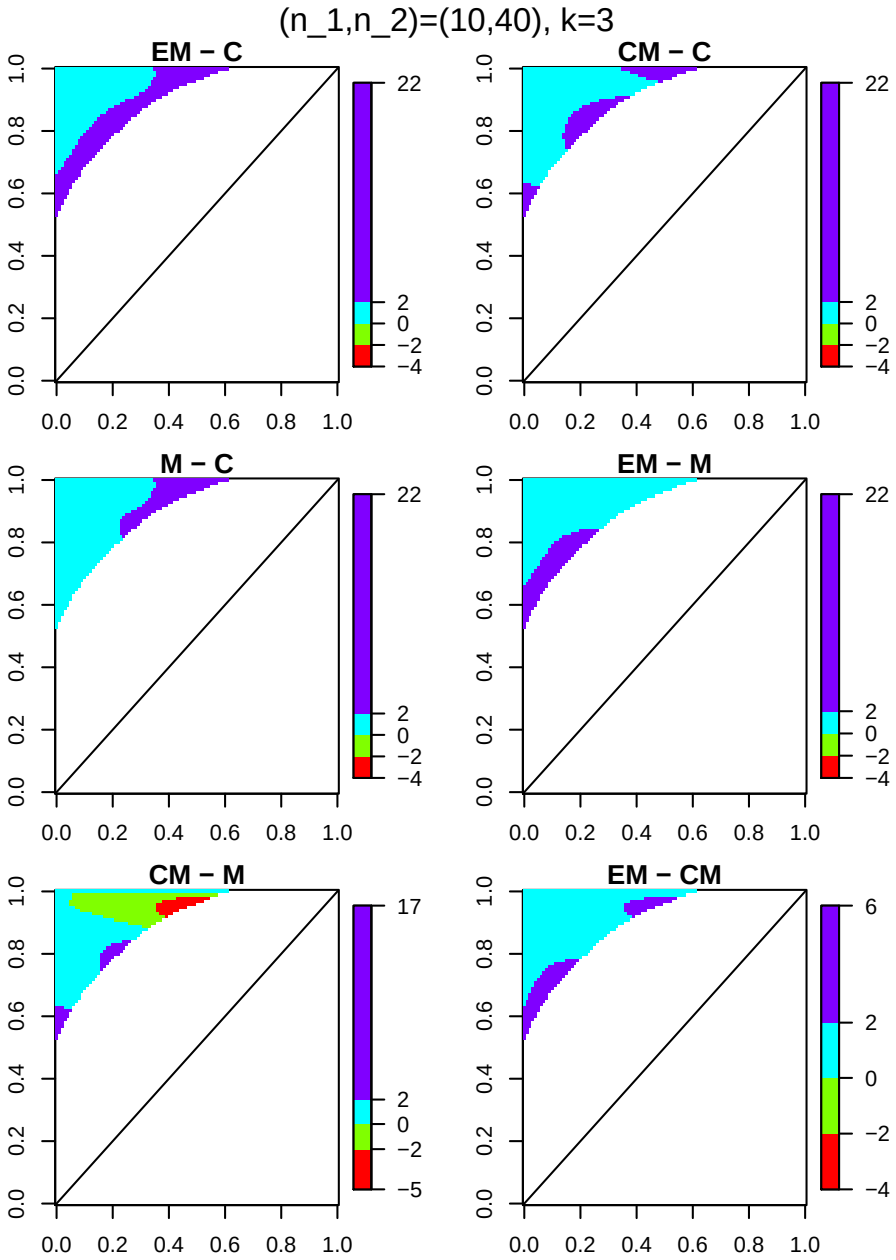


Figure 5.5: Plots of points in the parameter space that gives differences between to power functions in four considered intervals when $n_1 = 10, n_2 = 40$ and $k = 3$. In each panel we consider the differences between two power functions. For instance "EM vs C" means we consider the differences $\gamma_{CM} - \gamma_C$ and consider which points give differences in the four intervals considered. We always consider four intervals, even if there are no points in some of them. This is to ensure that the same color is used for the equivalent intervals for different pairs of power functions. The θ_1-axis is along the abscissa and the θ_2-axis is along the ordinate. We only consider grid points in which the power of at least one of all power functions is 80 % or greater and where the grid is specified at the beginning of this chapter. To the right of each panel there is a legend specifying which colors have been used for points that give differences in the four intervals of differences considered when creating the plot.

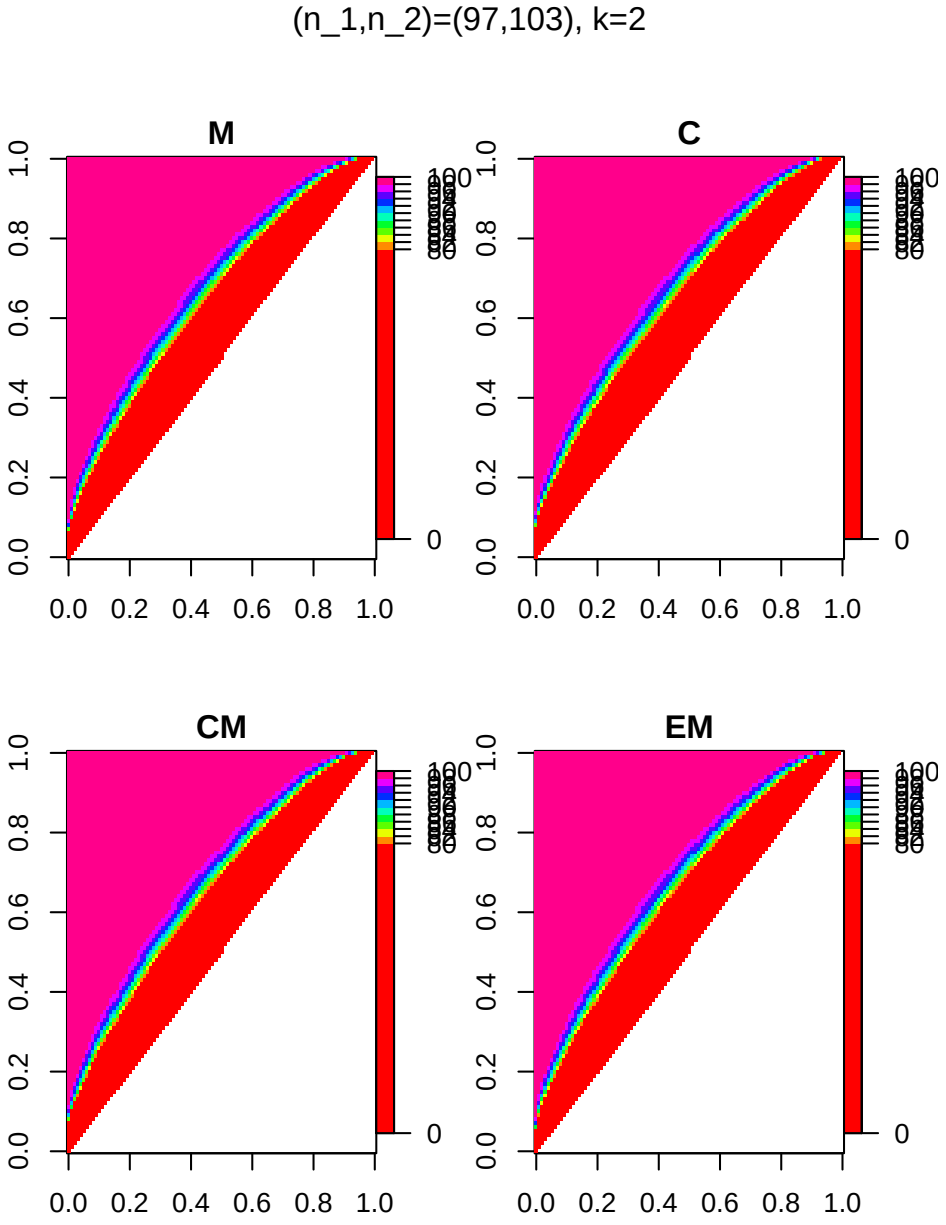


Figure 5.6: Plots of the different power functions when $n_1 = 97, n_2 = 103$ and $\alpha = 3 \cdot 10^{-2}$ ($k = 2$). In the title of each panel is the name of the power function shown. For instance "EM" means $\gamma_{EM}(\theta_1, \theta_2)$ is illustrated. The power function is only evaluated in the grid specified at the beginning of this chapter. To the right of each panel there is a legend specifying the different colors used for the different intervals of values in making the plot. The different cut points are $80, 82, 94, \ldots, 98, 100$. The θ_1-axis is along the abscissa and the θ_2-axis is along the ordinate of each panel.

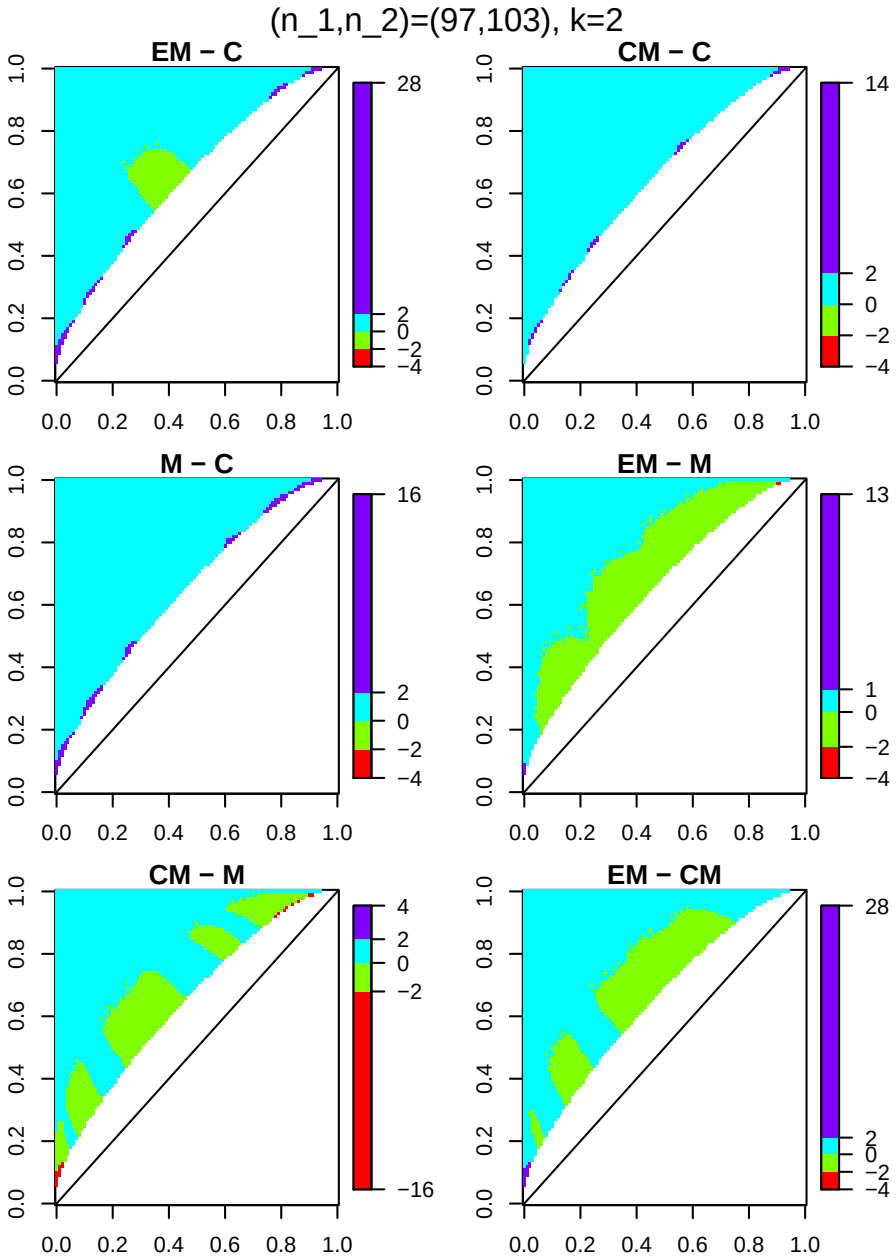


Figure 5.7: Plots of points in the parameter space that gives differences between to power functions in four considered intervals when $n_1 = 97, n_2 = 103$ and $k = 2$. In each panel we consider the differences between two power functions. For instance "EM vs C" means we consider the differences $\gamma_{CM} - \gamma_C$ and consider which points give differences in the four intervals considered. We always consider four intervals, even if there are no points in some of them. This is to ensure that the same color is used for the equivalent intervals for different pairs of power functions. The θ_1-axis is along the abscissa and the θ_2-axis is along the ordinate. We only consider grid points in which the power of at least one of all power functions is 80 % or greater and where the grid is specified at the beginning of this chapter. To the right of each panel there is a legend specifying which colors have been used for points that give differences in the four intervals of differences considered when creating the plot.

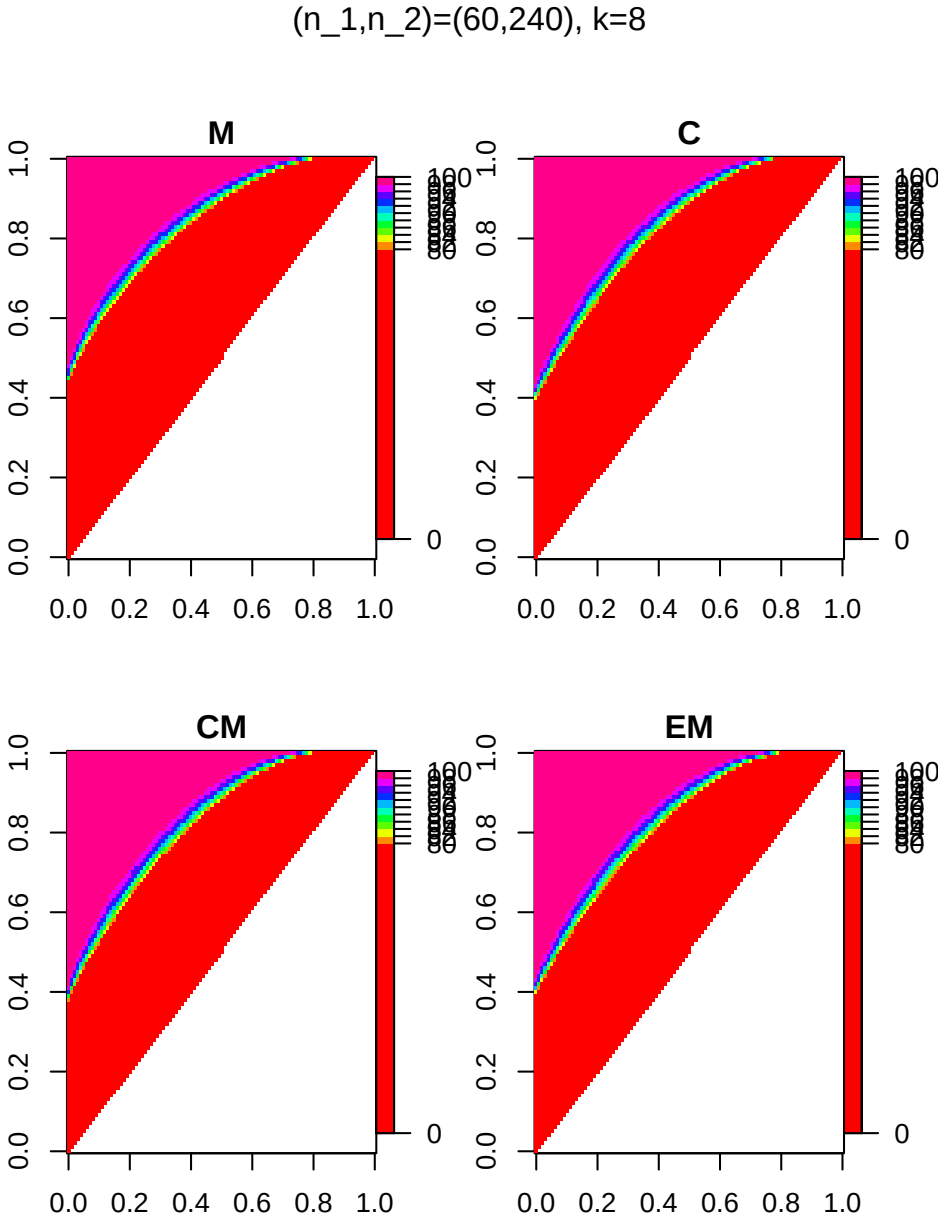


Figure 5.8: Plots of the different power functions when $n_1 = 60, n_2 = 240$ and $\alpha = 3 \cdot 10^{-8}$ ($k = 8$). In the title of each panel is the name of the power function shown. For instance "EM" means $\gamma_{EM}(\theta_1, \theta_2)$ is illustrated. The power function is only evaluated in the grid specified at the beginning of this chapter. To the right of each panel there is a legend specifying the different colors used for the different intervals of values in making the plot. The different cut points are $80, 82, 94, \ldots, 98, 100$. The θ_1-axis is along the abscissa and the θ_2-axis is along the ordinate of each panel.

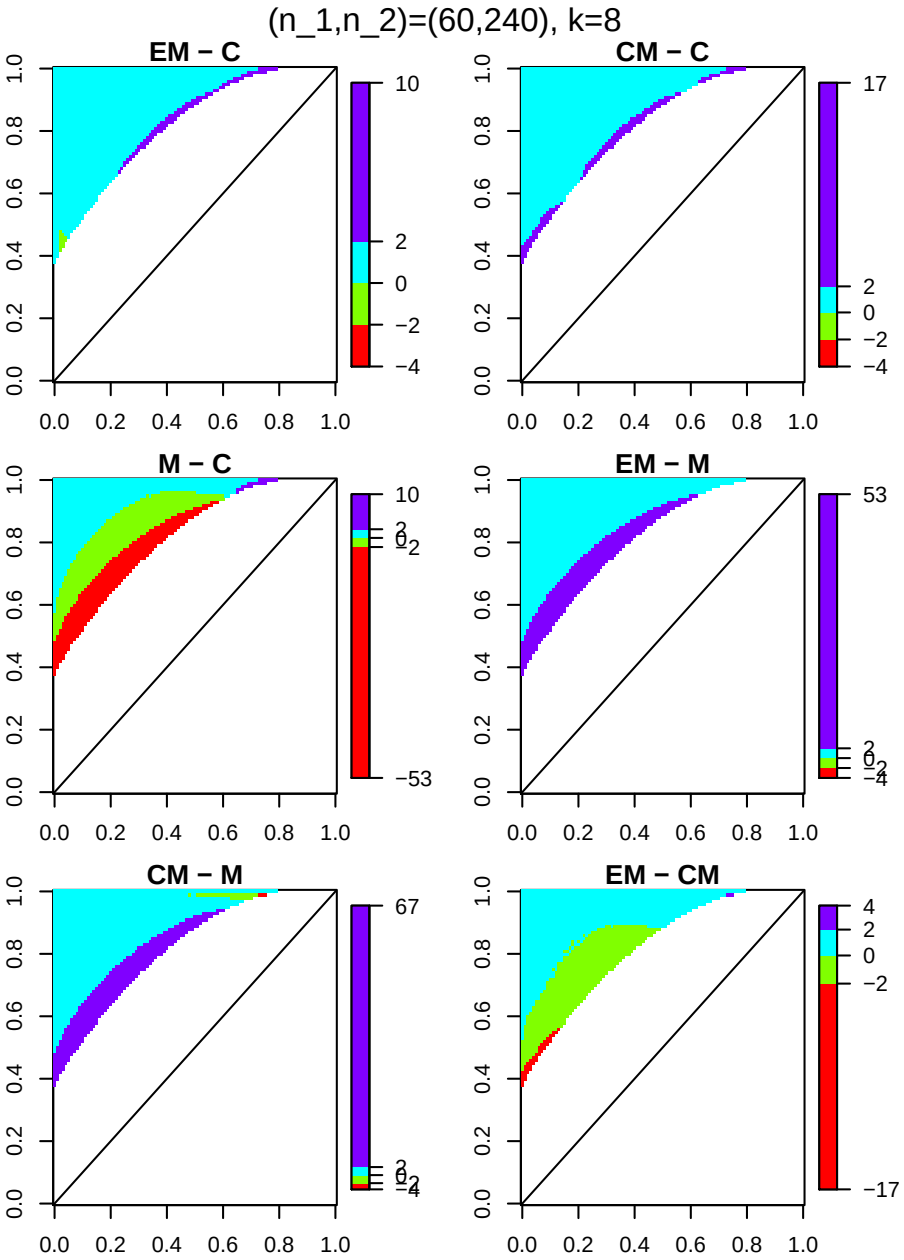


Figure 5.9: Plots of points in the parameter space that gives differences between to power functions in four considered intervals when $n_1 = 60, n_2 = 240$ and $k = 8$. In each panel we consider the differences between two power functions. For instance "EM vs C" means we consider the differences $\gamma_{CM} - \gamma_C$ and consider which points give differences in the four intervals considered. We always consider four intervals, even if there are no points in some of them. This is to ensure that the same color is used for the equivalent intervals for different pairs of power functions. The θ_1-axis is along the abscissa and the θ_2-axis is along the ordinate. We only consider grid points in which the power of at least one of all power functions is 80 % or greater and where the grid is specified at the beginning of this chapter. To the right of each panel there is a legend specifying which colors have been used for points that give differences in the four intervals of differences considered when creating the figure.

One reasonable question is why we use the empirical cumulative function and not for instance histograms. For starters it is easy to read of the median, percentiles and quartiles of a plot of this function. This is much harder to do for a plot of the histogram. Another reason is that the histogram is sensitive with respect to the bin size. Depending on the chosen number of bins the resulting histograms may differ significantly (plots not shown). It is also easier to compare several different cumulative difference functions when plotting them in the same figure than comparing histograms on top of each other.

In this section we consider some of the combined plots of the cumulative difference functions. In each of the plots we have plotted the functions for a specific choice of n_1, n_2, k. Along the abscissa in these plots are the difference in power in percent points and along the ordinate are the values of the cumulative difference functions. This function is only evaluated in the unique differences of the power functions considered, which means the spacing along the abscissa between the points of the cumulative difference functions need not be constant. In each figure the dot thickness is modulated so that all aspects of the curves are visualized.

$(n_1, n_2) = (10, 10), k = 2$ We see in Figure 5.10 that $\gamma_{EM} = \gamma_{CM}, \gamma_{EM} = \gamma_M, \gamma_{CM} = \gamma_M$ since the cumulative difference function is 1 in the point 0, so that $\gamma_{EM} = \gamma_{CM} = \gamma_M$. We therefore observe that the differences between either of γ_{CM}, γ_{EM}, γ_M with γ_C are exactly the same. When considering either one of the these differences we can see from the plot that the minimum difference is 0, that the maximum difference is about 22.5 and that the median difference is about 8. The median difference is quite high, which means performing an M-step on the C p-value will most likely give a test with significantly higher power.

To get a better understanding of the plot of the cumulative difference function it may help to picture a continuous function drawn between the points of the cumulative distribution function. When the slope of this function is large a higher proportion of points have the difference between the power function at this point than if the slope is smaller. When considering this imaginary curve for the cumulative difference function for the difference between any of the power functions with γ_C, we observe that the imaginary curve has greatest slope around 0 and smallest slope from about 17.5 to 22, which means the most frequently occurring difference is about 0 and that not many points give differences in the interval 17.5 to 22.

$(n_1, n_2) = (10, 40), k = 3$ In Figure 5.11 we see that none of the power functions are exactly equal in all the points, which is in agreement with Table 5.2. We

also observe that the differences between the power functions γ_{EM} and γ_C appear to be larger than for the other power functions since the slope of the imaginary line for the blue points is less in the interval $(0, 6)$ than for the other imaginary lines and most of the functions takes values on $[0, \infty)$. The difference function illustrated by the ochre points also takes negative values, but the proportion of these points is very small and smaller than the proportion of points in the interval $(0, 6)$ for the differences of power functions given by the blue points.

$\boldsymbol{(n_1, n_2) = (97, 103), k = 2}$ We see in Figure 5.12 that for all plots of the cumulative difference functions the differences are concentrated around 0, since the curves either start close to the difference of 0 or the difference functions takes values close to zero when evaluated in points smaller than zero and then the steeply rise to 1 for differences a little larger than 0 in both cases. We also observe that the cumulative difference function for the differences $\gamma_{EM} - \gamma_C$ and $\gamma_{EM} - \gamma_{CM}$ takes values for differences as large as about 30. However, we observe that the proportion of points that give these differences is very small in both cases.

$\boldsymbol{(n_1, n_2) = (60, 240), k = 8}$ We observe in Figure 5.13 that the differences between γ_{EM} and γ_C, γ_{EM} and γ_{CM}, γ_M and γ_C and γ_{CM} and γ_C are very small, since the plots of the cumulative difference functions of these differences change from about 0 to about 1 in a small region around the difference 0. This means $\gamma_{EM} \approx \gamma_{CM} \approx \gamma_C$. We also observe that γ_M is smaller than γ_C in about 25 % of the points and for about 12 % of the points in total the difference is smaller than -12%. The proportion of points where γ_M is greater than γ_C is negligible. We can also observe that the two distributions of the differences between γ_{CM} and γ_M and γ_{EM} and γ_M are very similar. We see that about 25% of the points give positive differences between the mentioned power functions and that the differences are larger than 12% for about 12.5% of the points in total. It is important to understand that even if the distributions of the differences between γ_{CM} and γ_M and γ_{CM} and γ_M are very similar, we do not know if the same points in the parameter space give the same differences within each of the two pairs of mentioned power functions by only considering the plots of the two cumulative difference functions alone. However, since we have also plotted the cumulative difference functions for the differences between γ_{EM} and γ_{CM} and this function is found to change from about 0 to about 1 in a small region around the difference 0, in most of the points the difference between γ_{CM} and γ_M is very close to the difference between γ_{EM} and γ_M.

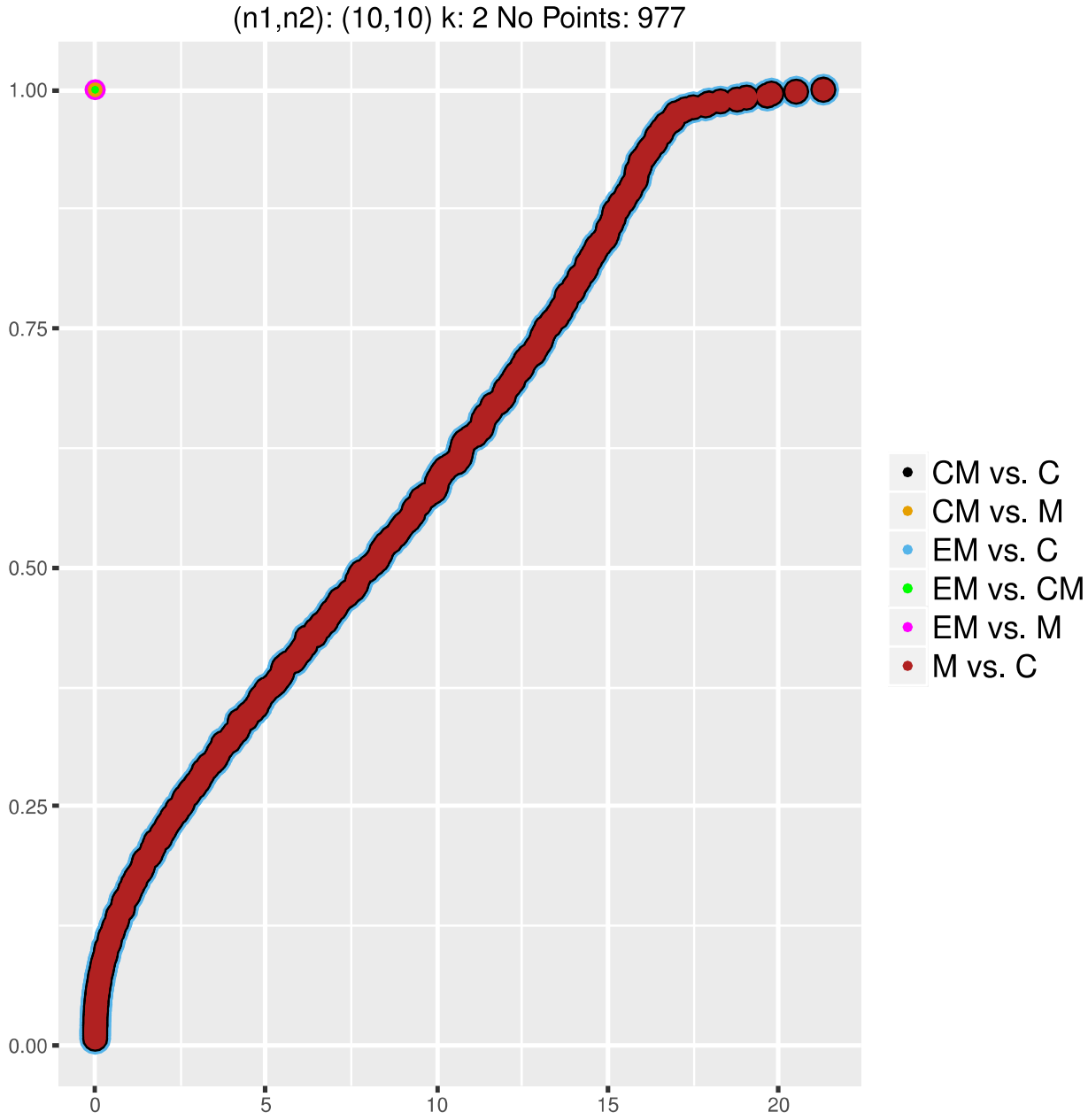


Figure 5.10: Combined plots of the cumulative difference functions for the differences between the power functions in each possible pair of power functions when $n_1 = 10, n_2 = 10, k = 2$. The differences are long the abcissa and the values of the cumulative difference functions are along the ordinate. To the right of the plot there is a legend specifying which color has been used in plotting each cumulative difference function. For instance the black dot in front of "CM vs. C" means the cumulative difference function of $\gamma_{CM} - \gamma_C$ has been plotted using black color. In the title of the plot the number of points where we have taken the differences between the power functions in each pair of power functions are shown. This number corresponds to the number of points where at least one of the power functions has power 80 % or greater. We observe that the cumulative difference functions of respectively $\gamma_{CM} - \gamma_C$, $\gamma_{EM} - \gamma_C$ and $\gamma_M - \gamma_C$ are on top of each other. The same holds for the cumulative difference functions of respectively $\gamma_{EM} - \gamma_{CM}$, $\gamma_{CM} - \gamma_M$ and $\gamma_{EM} - \gamma_M$.

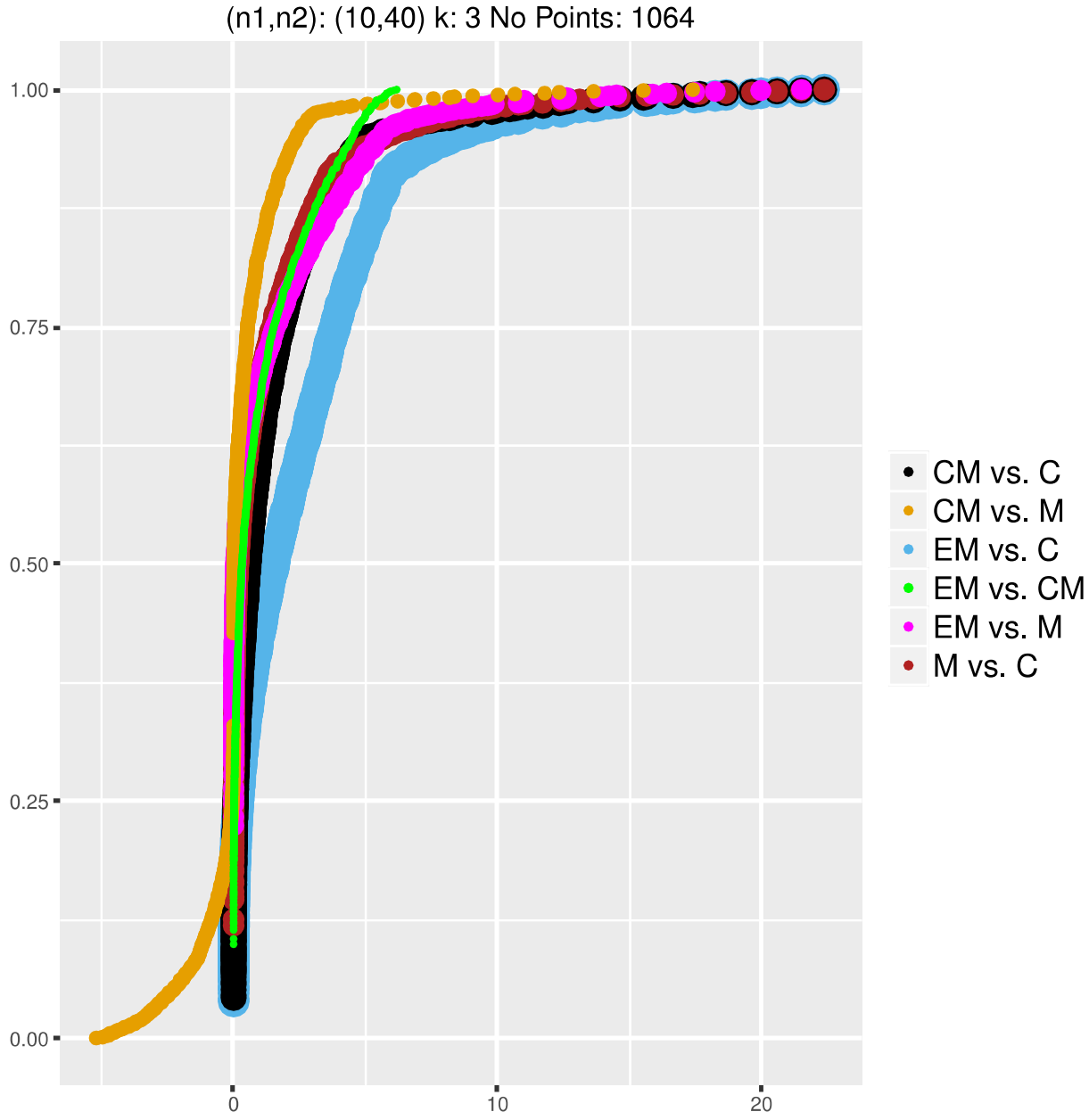


Figure 5.11: Combined plots of the cumulative difference functions for the differences between the power functions in each possible pair of power functions when $n_1 = 10, n_2 = 40, k = 4$. The differences are long the abcissa and the values of the cumulative difference functions are along the ordinate. To the right of the plot there is a legend specifying which color has been used in plotting each cumulative difference function. For instance the green dot in front of "EM vs. CM" means the cumulative difference function of $\gamma_{EM} - \gamma_{CM}$ has been plotted using green color. In the title of the plot the number of points where we have taken the differences between the power functions in each pair of power functions are shown. This number corresponds to the number of points where at least one of the power functions has power 80 % or greater.

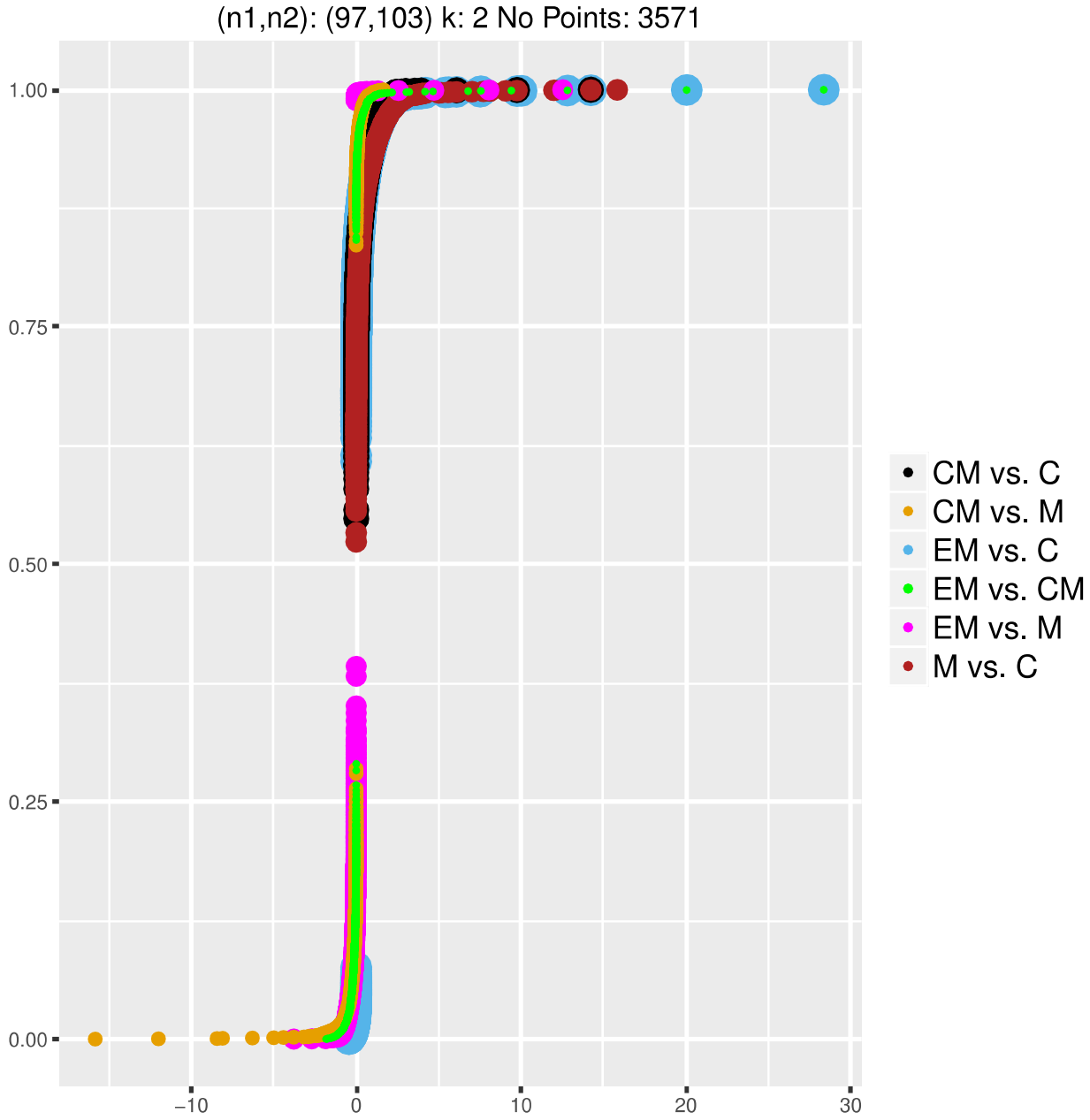


Figure 5.12: Combined plots of the cumulative difference functions for the differences between the power functions in each possible pair of power functions when $n_1 = 97, n_2 = 103, k = 2$. The differences are long the abcissa and the values of the cumulative difference functions are along the ordinate. To the right of the plot there is a legend specifying which color has been used in plotting each cumulative difference function. For instance the blue dot in front of "EM vs. C" means the cumulative difference function of $\gamma_{EM} - \gamma_C$ has been plotted using blue color. In the title of the plot the number of points where we have taken the differences between the power functions in each pair of power functions are shown. This number corresponds to the number of points where at least one of the power functions has power 80 % or greater.

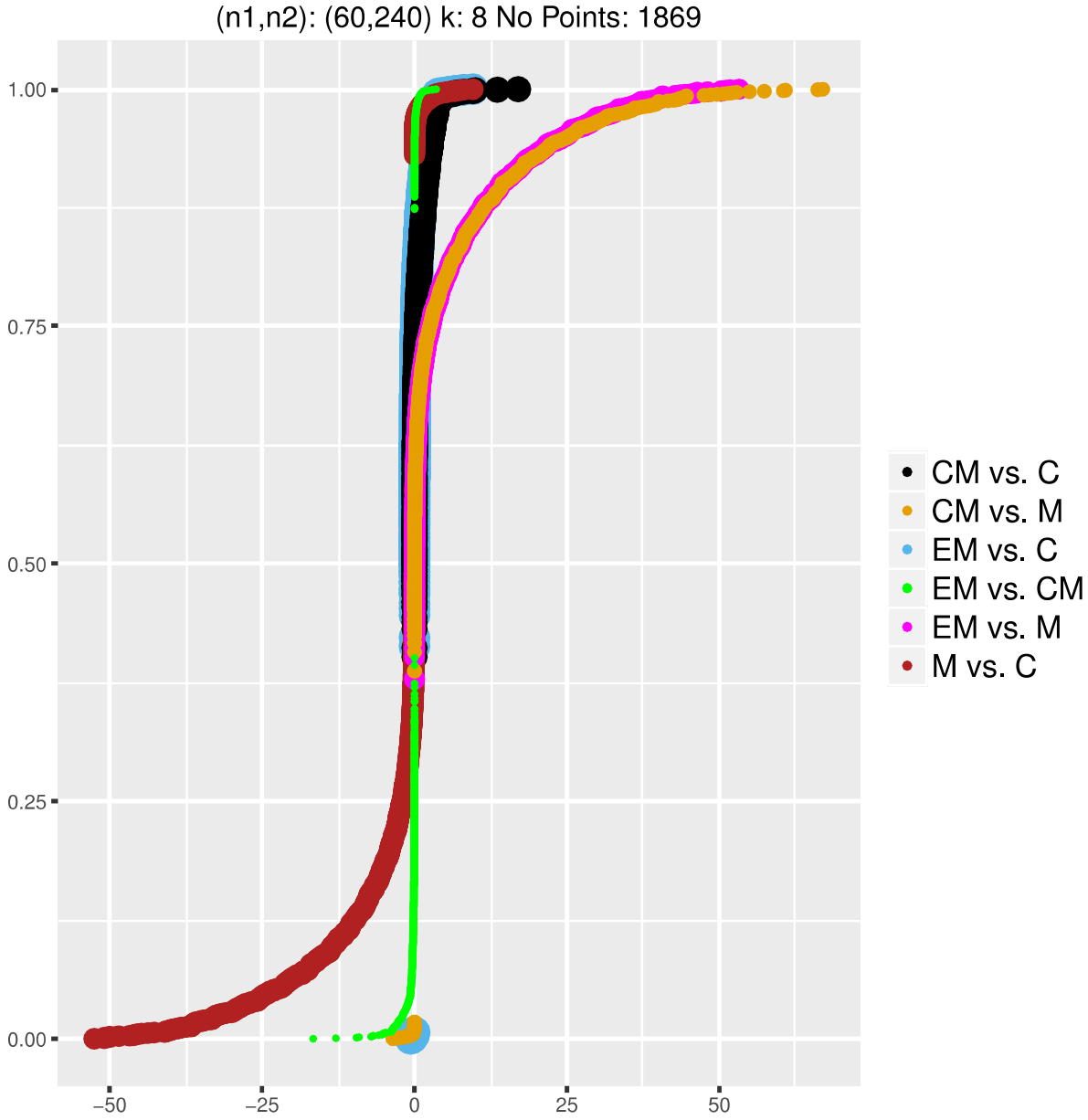


Figure 5.13: Combined plots of the cumulative difference functions for the differences between the power functions in each possible pair of power functions when $n_1 = 60, n_2 = 240, k = 8$. The differences are long the abcissa and the values of the cumulative difference functions are along the ordinate. To the right of the plot there is a legend specifying which color has been used in plotting each cumulative difference function. For instance the black dot in front of "CM vs. C" means the cumulative difference function of $\gamma_{CM} - \gamma_C$ has been plotted using black color. In the title of the plot the number of points where we have taken the differences between the power functions in each pair of power functions are shown. This number corresponds to the number of points where at least one of the power functions has power 80 % or greater.

5.2.4 Median power differences

In Table 5.4 the median differences between the power functions in each pair of power functions are shown. We have only taken the median of power differences in points where the power of at least one of all of the power functions is equal to or greater than 80%. We observe that median power differences only are substantially greater than 0 when one of the power functions considered is γ_C. We also observe that the median power increase is greatest for the two smallest designs considered, i.e $(10, 10)$ and $(4, 16)$. When considering the median power differences in the balanced design the number of points studied is below 50 when $k = 5, 6, 7, 8$ and in the unbalanced design when $k = 4, \ldots, 8$, so that these cases are most likely of no practical importance (since the points in parameter space corresponding to these points are most likely of no practical importance). However, for the other values of k the power increase compared to γ_C is quite large and may be of practical importance.

One sensible question to ask is why we have chosen the median to summarize the differences in power between power functions. When the differences are unique and there is an odd number of points, the median can be shown to minimize the following total *loss function* (proof not given)

$$L_{\text{tot}}(s) = \sum_{x_i} L_1(x_i, s), \tag{5.1}$$

where s is the number summarizing the numbers x_i, $i = 1, \ldots, n$ and

$$L_1(x_i, s) = |s - x_i| \tag{5.2}$$

is called the *absolute error loss* and gives the cost of summarizing the number x_i with s, i.e the cost incurred by the potential discrepancy between s and x_i. The loss function given in Equation (5.2) is not the only possible loss function. Another possible loss function is the *squared error loss function*

$$L_2(x_i, s) = (s - x_i)^2$$

(Casella & Berger 2002, p. 348–349). The minimizer of Equation (5.1) when replacing L_1 with L_2 is the mean of $x_i, i = 1, \ldots, n$, which is also a commonly used summary statistic (proof not shown).

We now compare the two loss functions L_1 and L_2 to better understand why we use the median and not the mean as a summary statistic. All loss functions are non-negative and $L(a, a) = 0$, meaning no cost occurs when summarizing the number a with itself. For the two considered loss functions $L(a, x_i) = L(x_i, a)$, which means summarizing the number x_i with s_1 is equally bad as summarizing the number x_i

with the number s_2 when $s_1 > x_i$, $s_2 < x_2$ and $|x_i - s_1| = |x_i - s_2|$. In fact, L_1 and L_2 are metrics. However, loss functions need not be metrics in general since they need not be symmetric, i.e $L(x_i, s) = L(s, x_i)$ does not need to hold. This occurs for instance when it considered to be more imprecise to summarize the number with s_2 than with s_1 in the example given. Since $|\epsilon| > \epsilon^2$ when $\epsilon \in (-1, 1)$ we have that $L_1 > L_2$ when $|x_i - s| < 1$. And since $|\epsilon| < \epsilon^2$ when $\epsilon \in (-\infty, -1) \cup (1, \infty)$, we have that $L_1 < L_2$ when $|x_i - s| \in (1, \infty)$. This means L_1 places more weight on small differences compared to L_2 which places relatively more weight on larger differences. Alternatively, if we increase the distance between x_i and s by one unit the increase in cost by using L_1 is 1 and is independent of the initial distance between s and x_i. However, when using L_2 the loss is greater if the original distance between x_i and s is large compared to when it is small. We do not regard values of (θ_1, θ_2) that give large differences in power between two power functions as more likely values of (θ_1, θ_2) under H_1. We therefore consider L_1 as a better cost function than L_2 and therefore use the median and not the mean as the summary statistic. We could of course weigh the different $L_1(x_i, s)$ differently than done in Equation (5.1). One possibility would be to weigh them according to some prior belief of the true values of (θ_1, θ_2) under H_1, where we place more weight on more likely values and less weight on the remaining values of (θ_1, θ_2). By using the expression for the total loss in Equation (5.1) we regard each value of (θ_1, θ_2) in which we consider the differences in power as an equally likely value of (θ_1, θ_2) under H_1.

5.3 Example (d) in Section 4.1.1 revisited

In this section we create the realisations of the different p-values in example (d) in Section 4.1.1. Since the A and E p-values are not valid, we should not use these p-values to create a level α-test. We choose $\alpha = 0.05$. In Table 5.6 the different realisations of the valid p-values considered in this thesis are shown. From Section 5.2.3 we know that $\gamma_{CM} \approx \gamma_C \approx \gamma_{EM} \approx \gamma_M$ when $n_1 = 97, n_2 = 103, k = 2$ in points where the power of at least one of the power functions is above or equal to 80%. Which one of the p-values should then be used? Is it in general wise to calculate the different realisations of the different p-values, as we have done in this example, and then choose the p-value that has the lowest realisation since the power functions are almost equal in (hopefully) the interesting part of the parameter space? Could this procedure be used when performing multiple hypothesis tests? Since each p-value is valid is seems at first glance reasonable that the new p-value should be valid.

If we always choose the smallest realisation of the p-values for all outcomes we are

Table 5.4: Median power differences of the power functions in all of the points where the power of at least one power function is equal to or larger than 80 %. Column 1 gives the sample sizes in the experiment, column 2 gives the significance level specified as $5 \cdot 10^{-k}$ and k ranges from 2 to 8, column 3 gives the number of points in which we evaluate the differences of the power functions (which means we consider the same points when taking the differences of all of the possible power functions) columns 4 to 9 each give the median power differences where we consider the difference between the leftmost power function in the column and the rightmost power function in the same column. For instance "EM vs. C" means we consider the median of $\gamma_{EM}(\theta_1, \theta_2; \alpha) - \gamma_C(\theta_1, \theta_2; \alpha)$ over the grid of (θ_1, θ_2) under H_1 where the power of at least one test for the specified values of n_1, n_2 and k is above or equal to 80 %.

(n_1, n_2)	k	No. points	EM vs. C	CM vs C	M vs. C	EM vs. M	CM vs. M	EM vs. CM
(10,10)	2	977	8.048	8.048	8.048	0.000	0.000	0.000
	3	384	11.459	11.459	11.459	0.000	0.000	0.000
	4	136	20.787	20.787	20.787	0.000	0.000	0.000
	5	45	33.720	33.720	33.720	0.000	0.000	0.000
	6	6	86.114	86.114	86.114	0.000	0.000	0.000
	7	0						
	8	0						
(25, 25)	2	2160	0.367	0.365	0.367	0.000	0.000	0.000
	3	1531	0.991	0.898	0.898	0.001	0.000	0.001
	4	1081	1.976	1.885	1.885	0.000	0.000	0.000
	5	745	1.024	1.024	0.738	0.003	0.003	0.000
	6	512	2.177	2.177	0.884	0.213	0.213	0.000
	7	362	1.935	1.935	1.935	0.000	0.000	0.000
	8	242	5.541	5.541	5.541	0.000	0.000	0.000
(50, 50)	2	2966	0.023	0.020	0.021	0.000	-0.000	0.000
	3	2394	0.036	0.033	0.033	0.000	0.000	0.000
	4	1987	0.104	0.104	0.104	0.000	0.000	0.000
	5	1650	0.112	0.112	0.063	0.000	0.000	0.000
	6	1400	0.255	0.250	0.165	0.000	0.000	0.000
	7	1178	0.250	0.137	0.013	0.051	0.019	0.025
	8	988	0.353	0.312	0.091	0.039	0.000	0.020
(97,103)	2	3571	0.000	0.000	0.000	0.000	0.000	0.000
	3	3135	0.000	0.000	0.000	0.000	0.000	0.000
	4	2800	0.000	0.000	0.000	0.000	0.000	0.000
	5	2520	0.000	0.000	0.000	0.000	0.000	0.000
	6	2289	0.000	0.000	0.000	0.000	0.000	0.000
	7	2091	0.001	0.001	0.000	0.000	0.000	0.000
	8	1907	0.001	0.001	0.000	0.000	0.000	0.000
(150, 150)	2	3836	0.000	0.000	0.000	0.000	0.000	0.000
	3	3470	0.000	0.000	0.000	0.000	0.000	0.000
	4	3192	0.000	0.000	0.000	0.000	0.000	0.000
	5	2952	0.000	0.000	0.000	0.000	0.000	0.000
	6	2755	0.000	0.000	0.000	0.000	0.000	0.000
	7	2568	0.000	0.000	0.000	0.000	0.000	0.000
	8	2415	0.000	0.000	0.000	0.000	0.000	0.000
(4,16)	2	642	6.669	6.669	6.669	0.000	0.000	0.000
	3	194	9.895	9.895	9.895	0.000	0.000	0.000
	4	25	20.074	20.074	20.074	0.000	0.000	0.000
	5	8	86.838	86.838	86.838	0.000	0.000	0.000
	6	0						
	7	0						
	8	0						
(10,40)	2	1731	0.861	0.426	0.758	0.003	0.000	0.008
	3	1064	1.539	0.562	0.178	0.082	0.013	0.323
	4	653	3.462	3.462	0.710	0.116	0.116	0.000
	5	414	4.514	4.514	1.770	0.044	0.044	0.000
	6	241	6.438	5.947	3.589	0.008	0.008	0.000
	7	129	8.635	9.698	5.678	0.035	0.040	-0.005
	8	60	7.067	5.112	7.067	0.000	-0.556	0.556
(20,80)	2	2535	0.102	0.030	0.064	0.004	-0.000	0.012
	3	1916	0.253	0.086	0.003	0.061	0.004	0.050
	4	1484	0.380	0.191	-0.484	0.990	0.611	0.027
	5	1167	0.583	0.312	-0.082	0.817	0.369	0.071
	6	910	0.731	0.575	-0.021	0.512	0.364	0.091
	7	718	0.884	1.019	-0.005	0.449	0.449	0.000
	8	555	0.897	1.210	-0.002	0.125	0.146	-0.042
(60,240)	2	3538	0.000	0.000	0.000	0.000	0.000	0.000
	3	3110	0.000	0.000	0.000	0.000	0.000	0.000
	4	2767	0.000	0.000	-0.000	0.001	0.001	0.000
	5	2483	0.000	0.000	-0.000	0.002	0.001	0.000
	6	2246	0.001	0.000	-0.005	0.009	0.005	0.000
	7	2046	0.001	0.001	-0.001	0.006	0.004	0.000
	8	1869	0.001	0.002	-0.003	0.010	0.007	0.000

Table 5.6: The different realisation of either of the C, M, C $\circ$ M or E $\circ$ M p-values. In the first row the different P-values considered are shown and on the second row the realisations of the p-values are given.

p-value	p_C	p_{CM}	p_M	p_{EM}
$p(\boldsymbol{x})$	0.101	0.0897	0.0914	0.0864

Table 5.7: Type I error probabilities of the level 0.05 test based on $p_{\min}(\boldsymbol{X})$ when $n_1 = 97, n_2 = 103$ at the grid of θ-values used under H_0

θ	0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
$\Pr_\theta(p_{\min}(\boldsymbol{X}) \leq 0.05)$	0	5.01	4.72	4.99	5.08	5.09	5.09	5.13	4.82	4.76	4.94

in fact using the test statistic

$$p_{\min}(\boldsymbol{X}) = \min(p_M(\boldsymbol{X}), p_{EM}(\boldsymbol{X}), p_{CM}(\boldsymbol{X}), p_C(\boldsymbol{X}))$$

When calculating $\Pr_{\boldsymbol{\theta}}(p_{min}(\boldsymbol{X}))$ we sum over least as many outcomes by construction as when calculating each of the corresponding probabilities for the other p-values. So, if $\Pr_\theta(p(\boldsymbol{X}) \leq \alpha) = \alpha$ for one of the p-values and we sum over one additional outcome with positive probability when calculating $\Pr_{\boldsymbol{\theta}}(p_{min}(\boldsymbol{X}) \leq \alpha)$ compared to when calculating $\Pr_\theta(p(\boldsymbol{X}) \leq \alpha)$ we have that $p_{\min}(\boldsymbol{X})$ is not valid. In Table 5.7 we show the type I error probabilities when $n_1 = 97, n_2 = 103, k = 2$ and when using $p_{\min}(\boldsymbol{X})$ to create the level 0.05 test. We see that $p_{\min}(\boldsymbol{X})$ is not valid.

We see that $\gamma_{CM} \approx \gamma_C \approx \gamma_{EM} \approx \gamma_M$ in points where the power of at least one of the tests is above 80 % does not necessarily mean that the realisations of the p-values are the same when evaluated in the same outcomes. What does then $\gamma_{CM} \approx \gamma_C \approx \gamma_{EM} \approx \gamma_M$ mean? Given that the true values of the success probabilities are in the considered region over parameter space, i.e the region where at least one of the power functions has power at least 80 %,$\gamma_{CM} \approx \gamma_C \approx \gamma_{EM} \approx \gamma_M$ means the long frequencies we observe realisations of the p-values below or equal to 0.05 for all of the p-values will almost be the same. So in the long run we will get almost the same estimates of the type II error probabilities using either one of the tests (given that the (θ_1, θ_2) is in the considered region of the parameter space) (also recall that $\Pr_\theta(\text{type I error}) = 1 - \gamma$). To sum up, we must choose one valid p-value in a given set-up. In this example we have chosen $\alpha = 0.05$. We therefore will not reject the null hypothesis using any of the p-values. We note that Zhao et al. (2011) use a lower significance level than we do since they perform multiple tests.

5.4 Further work

In this section we present topics that can be part of further work.

5.4.1 Better comparison of power functions under H_0

In Section 5.1 we did simple comparisons of the power functions under H_0. The main focus was to show that the A and E p-values in general not are valid and that the remaining p-values are. In many situations we observed that the power functions γ_M, γ_{EM} and γ_{CM} take almost the same values and that these values are higher than the values taken by γ_C. It is of interest to quantify the power increase also under H_0 and see if $\gamma_C \approx \gamma_{CM} \approx \gamma_{EM}$ holds in general under H_0 or if patterns similar to the ones observed under H_1 also hold under H_0. One possible approach would be to use the same grid along $\theta_1 = \theta_2$ as the grid increment used under H_1 and thereafter create similar tables and figures for the power functions as the ones created under H_1.

5.4.2 More smaller, balanced and unbalanced designs

We have observed that the power functions $\gamma_{EM}, \gamma_{CM}, \gamma_M$ are significantly greater than γ_C for small unbalanced and balanced designs. These changes were greatest for the two designs $(10, 10)$ and $(4, 16)$. It is of interest to know if similar results hold for smaller designs than $(10, 10)$ and $(4, 16)$ and if so how large the power increases are. We could also study other unbalanced designs than $n_2 = 4n_1$, for instance $n_2 = 3n_1$ and $n_2 = 2n_1$.

5.4.3 The Berger and Boos p-value

When calculating the realisations of the M p-value we maximize the tail probability over the entire parameter space possible for the nuisance parameter θ under H_0. Berger & Boos (1994) suggest instead maximizing over a confidence set for θ under H_0. The p-value is given by

$$p_{BB}(\boldsymbol{x}) = \sup_{\theta \in C_\beta} \Pr_\theta(T(\boldsymbol{X} \geq T(\boldsymbol{x})) + \beta,$$

where C_β is a $1 - \beta$ confidence set for θ under the null hypothesis. One possible confidence set for θ under H_0 in our case is the Pearson-Clopper confidence interval,

see for instance Casella & Berger (2002, p. 454). Berger & Boos (1994) show that this p-value is valid. This method may be preferred over the M method when the set of values possible for θ under H_0 is unbounded, since maximization over the entire parameter space may be impossible. If we use $-p_C(\boldsymbol{X})$ as test statistic in the BB method we get

$$\begin{aligned} p_{CBB}(\boldsymbol{x}) &= \sup_{\theta \in C_\beta} \Pr_\theta(-p_C(\boldsymbol{X}) \geq -p_C(\boldsymbol{x})) + \beta \\ &= \sup_{\theta \in C_\beta} \Pr_\theta(p_C(\boldsymbol{X}) \leq p_C(\boldsymbol{x})) + \beta \\ &\leq p_C(\boldsymbol{x}) + \beta, \end{aligned}$$

since $\Pr_\theta(p_C(\boldsymbol{X}) \leq \alpha) \leq \alpha$ holds for all $\alpha \in [0, 1]$ as $p_C(\boldsymbol{X})$ is valid and we can replace α with $p_C(\boldsymbol{x})$. Therefore applying a BB step on the C p-value does not necessarily give a test with uniformly at least as high power as the test with the same level test based on the C p-value. This is one potential drawback with the method. However, applying the BB method on any test statistic will result in a valid p value since the p-value is valid, meaning it is possible to make the A p-value and E p-value valid by applying a BB step on them (or to be more correct we use the negative of either the A p-value and E p-value as test statistic in the BB method).

5.4.4 Different null and alternative hypothesis

Instead of the null hypothesis in Equation (4.1), one could consider

$$H_0 : |\theta_1 - \theta_2| \leq \epsilon,$$

where ϵ is the magnitude of the biggest difference of no practical significance. The alternative hypothesis then takes the form

$$H_1 : |\theta_1 - \theta_2| > \epsilon,$$

and all $(\theta_1, \theta_2) \in \Theta_0^c$ are of practical importance. When we consider different hypothesis tests testing the above null hypothesis against the alternative hypothesis we want tests with as high power as possible for all $(\theta_1, \theta_2) \in \Theta_0^c$ since all $(\theta_1, \theta_2) \in \Theta_0^c$ now are practically important. With the new null hypothesis there is therefore no need to check that the power is low at parameter values that are not practically significant.

Under the new null hypothesis the joint pmf is

$$\begin{aligned} f(x;\theta_1,\theta_2) = &\binom{n_1}{x_1}\binom{n_2}{x_2}\theta_1^{x_1}\theta_2^{x_2}(1-\theta_1)^{n_1-x_1}(1-\theta)^{n_2-x_2} \\ =&\theta_1^{s_1}\theta_2^{s_2}(1-\theta_1)^{n_1-s_1}(1-\theta_2)^{n_2-s_2}\binom{n_1}{x_1}\binom{n_2}{x_2} \end{aligned}$$

so that $S(X_1, X_2) = (X_1, X_2)$ is sufficient for (θ_1, θ_2) by the factorization theorem. The entire data set is always sufficient. We then have

$$\Pr(X_1 = x_1, X_2 = x_2 \mid S(X) = (s_1, s_2)) = \begin{cases} 1 & \text{if } x_1 = s_1, x_2 = s_2 \\ 0 & \text{else} \end{cases},$$

which means the C p-value realisations are 1 for any test statistic. This means the power of the test based on the C p-value will be 0 for all (θ_1, θ_2) under the alternative hypothesis. We therefore should try to find another sufficient statistic for (θ_1, θ_2) under the null hypothesis that gives a test with better power properties.

Chapter 6

Short discussion and conclusion

In this chapter we give a short overall discussion and conclusion. We have done most of the discussion in the previous chapters so that the discussion and conclusion presented in this chapter are meant to give an overview.

We know the p-value is a random variable and not a constant probability, see Section 3.3. This may be hard to understand since many introductory texts in statistics teach that the p-value is a probability. For instance Devore et al. (2012, p. 456) define the p-value as "The ***P*-value** is the probability, calculated assuming that the null hypothesis is true, of obtaining a value of the test statistic at least as contradictory to H_0 as the value calculated from the available sample." and they emphasize that "The P-value is a probability.". Another example has already been given in Section 3.3, where we said that Walpole et al. (2012, p. 333–334) clearly view the test statistic as a random variable and the p-value as not. One may question if introductory texts is the first place students should be taught that the p-value is a random variable. One of the reasons is that students may not understand the distinction between a random variable and the realisations of the random variable. This if fairly easy to understand, where for instance a discrete random variable X takes the different possible realisations x according to a probability distribution. This means before an experiment the outcome is undecided and that in the experiment the random variable X takes the different realisations with certain probabilities. With this knowledge it should be easy to understand that when obtaining two different outcomes in two runs of an experiment we potentially sum over different outcomes when calculating the realisations of a particular p-value, so that the realisations of the p-value may not be the same value. Perhaps the most straightforward exercise would be to evaluate the p-value in all the possible outcomes when the set of outcomes is finite, as done in Section 3.3.

Casella & Berger (2002, p. 397–399) define the p-value as a random variable and also define validity. This is a commonly used book in theoretical statistics courses[1]. The section in the book is fairly short and the book is written at a high theoretical level. This may make it harder for the students to realise that the p-value is not a probability but a random variable, even if they are at a more advanced level. One issue with first introducing the p-value as a random variable in a theoretical statistics course is that far from all users of statistics take such a course. If one does not understand that the p-value is a random variable it is highly likely that one neither understands the concept of validity and that not all p-values are valid. It is also highly unlikely that one understands that there exist different p-values and that the tests based on these p-values may have different power functions.

We have seen in Section 5.1 that the concept of validity is important. It is the foundation for creating tests that never exceed the chosen significance level under the null hypothesis. We have seen that p_M, p_C, p_{CM}, p_{EM} are valid and that p_A and p_E are not. The most severe examples where either of the test based on p_A or p_E exceed the significance levels are found in unbalanced designs. The maximum relative type I probability found for the A p-value is $1.2054 \cdot 10^{-6}$ and occurs when $n_1 = 60, n_2 = 240, \alpha = 5 \cdot 10^{-8}$ and $\theta = \theta_1 = \theta_2 = 0.05$. The maximum relative type I probability found for the E p-value is $5.60 \cdot 10^{-6}$ and occurs when $n_1 = 97, n_2 = 103, \alpha = 5 \cdot 10^{-6}$ and $\theta = \theta_1 = \theta_2 = 0.15$.

Under the alternative hypothesis, the E $\circ$ M and C $\circ$ M p-values are found in general to give level α tests with highest power (only tests based on valid p-values studied). We have also observed that the power differences occur in the region(s) of the points considered that most likely is(are) of greatest scientific importance. In balanced designs for the largest designs studied, i.e $(97, 103)$ and $(150, 150)$, the level α tests based on all the studied p-values, i.e the E $\circ$ M, C $\circ$ M, C and M p-values have power functions that take almost identical values in the majority of parameter points studied. In unbalanced designs the differences between E $\circ$ M, C $\circ$ M, C are found to be small in the largest studied design, $(60, 240)$. For the two smallest designs studied, $(n_1, n_2) = (4, 16)$ and $(n_1, n_2) = (10, 10)$ the power increase of γ_{CM} compared with γ_C can be quite substantial. So the M step may increase the power substantially for designs with small sizes. We get similar results when compare γ_{EM} or γ_M with γ_C as when comparing γ_{CM} with γ_C.

The M step is theoretically optimal since (1) it makes a invalid p-value valid (if we let the negative of the p-value serve as test statistic in the M method) and (2) if we

[1]This book has for instance been used in STAT210 at the University of Bergen in the spring of 2016, in the course TMA4295 Statistical Inference at NTNU in numerous of past years, see respectively `http://www.uib.no/emne/STAT210` and `http://www.ntnu.edu/studies/courses/TMA4295#tab=omEmnet` for further information.

let the negative of a valid p-value serve as test statistic in the M method the power function of the level α test based on the resulting p-value will be uniformly least as high as the power function of the test with same level based on the original p-value. We have seen that the M step maintains the ordering of the test statistic. This means that it is the ordering induced by the test statistic that determines if the power of the level α test based on the p-value will have good or bad properties. The C p-value also have a good property; it is valid. The C method is in general also faster to compute than the M method since in the M method we must calculate the tail probability in a grid of values of $\theta_1 = \theta_2$ and we must consider more outcomes when calculating the tail probabilities. The E $\circ$ M and C $\circ$ M methods are also more time consuming than the C method since all the realisations of the E or C p-values must be calculated when considering respectively the E $\circ$ M or C $\circ$ M methods. For instance it takes 36 seconds, 7 minutes and 42 seconds, 11 minutes and 39 seconds and 8 minutes and 16 seconds to calculate all the realisations of respectively the C, M, E $\circ$ M, C $\circ$ M p-values when $n_1 = n_2 = 150$. Using the R implementation of the M method takes 5 hours, 21 minutes and 44 seconds. When using the R implementation the grid on $[0, 1/2]$ has increment 0.05, which gives a much sparse grid than the grid used in the power study. We see that the M method is quite computationally intensive and clearly benefits from parallel programming. However, due to better and better computers the M step becomes more and more relevant.

Since the C method is quicker than the M method and $\gamma_C \approx \gamma_{CM} \approx \gamma_{EM}$ for the largest study designs, we recommend using the C step for large sample sizes. For medium to small sample sizes we recommend using either the E $\circ$ M or C $\circ$ M p-value as they in general are found to give tests which are the most powerful and more powerful than the test with the same level based on the C p-value. It might be tempting to calculate the realisations of all studied p-values in this thesis and choose the one with the smallest realisation. In Section 5.3 we have seen that this approach gives a p-value that is not valid. We therefore not recommended using the "minimum" p-value.

When comparing the studied power functions we have used both tables and graphical aids. By plotting where in parameter space the differences in power occur, we can easily see if the points are of scientific interest. For instance if most of the points where there are great differences in power had been close to $(1, 0)$, then the power differences would most likely be of little practical importance. However, we have seen that among the studied points the power differences are in the interesting part of the parameter space. This is much harder to do by only comparing power functions in the same grid point and trying to eyeball which points in parameter space are of interest. The dataset quickly becomes unmanageable if one shrink

the grid increment in the power study. Also, by only doing this comparison of the power functions, one can only make a table with a selection of values of the power functions showing interesting features or differences between the power functions. There is limited amount of information in such tables. Only Table D.1 in this thesis has been constructed in this way. The reason we have included the table, when we know there is limited information in such a table, is that the purpose of it is to show each case where one of the power functions exceeds the significance level. We note that the entries in Table D.1 may not reflect the overall trends of the power functions. Table 5.2 and the difference plots provide much more information about the differences between the power functions than Table D.1 does. Increasing the parameter points under H_1 will not make it any harder to construct the difference plot or Table 5.2, but increasing the dataset under H_0 will make it more time consuming to make Table D.1. Constructing tables similar to Table D.1 with the aim of power comparisons will be infeasible with increasing number of parameter points.

When comparing power functions we have also plotted the cumulative difference functions. The plots of the cumulative difference functions carry much more information about the differences than Table 5.2. For instance if the fraction of points that give differences in the interval 0 to 2 % is 1, we do not know the distribution of the differences within this interval. It could be that most points give differences close to 0 or that most differences are close to 2. The use of difference plots and plots of the cumulative difference functions have proved to be quite helpful in comparing power functions. We therefore recommend using these tools when comparing power functions. We note that it does take time to create the plots of the cumulative difference functions, where one needs to set the dot sizes and plotting order so that each function is visible in the figure. We also note that it takes times to construct the difference plots and the plots of the power functions. The hardest part is to understand how to create a matrix of the differences or power values that the plotting function will accept. Despite the mentioned difficulties, we strongly recommend creating these plots.

Appendix A

Basic definitions and concepts in biology and population genetics

In this appendix we give a short introduction to the definitions and concepts needed to understand example (d) in Section 4.1.1. This introduction is a revised version of Appendix B in the projet's thesis Aanes (2016).

DNA DNA (deoxyribonucleic acid) is a long sequence made up of the four bases adenine (A), guanine (G), cytosine (C) and thymine (T). Two sequences of DNA lines up with one another to form a spiral that is called a double helix. The two sequences lines up so that A or T from one sequence is paired up with respectively C or G from the other sequence. Given one of the sequences we can tell the other one (i.e if we look at the excerpt ATCG from one sequence, we know that it is paired up with TAGC). We therefore only consider one of the sequences (National Human Genome Research Institute 2015*b*).

Chromosome The structure in which a single DNA molecule is stored is called the chromosome. A human has 46 chromosomes. The chromosomes pairs up so that each human has 23 pairs. In each pair, one chromosome is inherited from the father and one is inherited from the mother (National Human Genome Research Institute 2015*a*). One of the pairs consists of the sex chromosomes. The rest of the pairs are called autosomes (Ziegler & König 2010, p. 3).

Gene A gene is a sequence of DNA that codes for a protein (Ziegler & König 2010, p. 7).

Locus A locus (plural loci) is a sequence of DNA that may or may not code for a protein (Halliburton 2004, p. 28). An autosomal locus is a locus on one of

the autosomes.

Allele: An allele is an alternative form (i.e alternative sequence) of a locus (Halliburton 2004, p. 28).

Genotype: The chromosomes are ordered in pairs. When we consider the same locus on each of the two chromosomes in a pair, two alleles are present. The pair of alleles (considered unordered) is said to be the genotype for a particular individual at the locus. For instance, if a locus has two alleles labelled R and T and if an individual has the two alleles R and R, the genotype of the individual will be RR (Thompson 1986, p.2).

SNP In a SNP, single nucleotide polymorphism, a single base in one sequence of DNA is substituted with another base. For instance the base A could be replaced by G. Since for a SNP one base is replaced by another, four alleles are possible. However, most SNPs have only two alleles (we say that they are diallelic) (Ziegler & König 2010, p. 54). We only consider diallelic SNP-loci in this thesis.

Law of segregation The law of segregation states that for a particular locus in an individual, one allele is inherited from the mother and the other allele is inherited from the father (Thompson 1986, p. 1).

Phenotype A phenotype is the expression of a trait in an individual, i.e the trait we can *observe* (Genetics Home Reference 2015). There are two types of traits, qualitative traits and quantitative traits (Halliburton 2004, p. 525). Qualitative traits show discrete phenotypes and are controlled by mostly one gene. Quantitative traits show a continuous distribution of phenotypes and are controlled by many genes and by environmental factors. Eye color is an example of a qualitative trait and blue eyes is then a possible phenotype. Weight is an example of a quantitative trait and the actual weight of an individual is then the phenotype for that individual.

Gametes A gamete is the reproductive cell of an organism (The Editors of Encyclopædia Britannica 2016). Humans have either sperm cells or egg cells. Each gamete carry only one copy of one of the chromosomes in each pair.

Genetic marker A genetic marker is a locus with known location and that is polymorphic (Ziegler & König 2010, p. 47), i.e there exists more than one allele at the locus in the population (Halliburton 2004, p. 28).

Law of independent assortment When an indivudal passes on a gamete to an offspring, the law of independent assortment states that which of the two chromosomes is passed on in one pair is independent of which of the two

chromosomes that is passed on in the other(Ziegler & König 2010, p. 22).

Recombination Recombination is any process that results in the set of alleles an individual passes on to an offspring is not the same as the set of alleles the individual inherited from either the mother or the father(Halliburton 2004, p. 91). For instance, we know that an offspring gets for each of the autosomal chromosomes one of the chromosomes from the father and the other from the mother. If we consider two different pairs of chromosomes the law of independent assortment tells us that which of the two chromosomes is passed on in one pair is independent of which of the two chromosomes that is passed on in the other. This means the individual may pass on one chromosome inherited from the father and another chromosome inherited from the mother, so that new combinations of alleles are passed on. Other types of recombination are also possible. Let us consider two loci on a chromosome which is passed on to an offspring. It is possible that the allele on one of the two loci on this chromosome and the allele at the same loci in the other chromosome (i.e the other chromosome in the pair) have been swapped during meiosis (the process where gametes are made) but the alleles at the other considered loci have not been swapped (Ziegler & König 2010, p. 9). This means the chromosome in the individual and the copy of it which is passed on to the offspring differ by one allele (assuming the alleles that have been swapped are different).

Linkage When we consider two loci on the same chromosome and if the alleles at these two loci are not passed on to an individual according to the law of independent assortment (so that which allele is passed on to the offspring is not independent of which allele is passed on at the other loci), we say the two loci are *linked* (Halliburton 2004, p. 94). Otherwise they are *unlinked*. The closer the two loci are, the less likely it is the two loci will act according to the law of independent assortment. This means the closer the two loci are the more likely it is that the alleles at the two loci are inherited together.

Linkage disequilibrium Two loci are in linkage disequilibrium if the alleles at one of the loci are not randomly associated with the alleles at the other loci, i.e they do not act according to the law of independent assortment (Halliburton 2004, p. 93–94). Note that the two loci can be on two different chromosomes, so that linkage is not a necessary condition to have linkage disequilibrium. This means the law of independent assortment does not always hold, even if the two loci considered are on two different chromosomes. The four forces of evolution can cause gametic disequilibrium.

Association study The goal of an association study is to determine if any of

the considered genetic markers are associated with the disease under study. These studies are based on historical recombination. The recombination uncouples all but the most tightly linked markers from the causal locus. When one finds a genetic marker that is associated with a disease, it is hopefully tightly linked to the causal locus, since if so further studies may localize the causal locus. If not, so that the genetic marker and locus only are in linkage disequlibirum, the localization will be much harder, if not impossible Mackay et al. (2009).

Appendix B

New coordinates of point rotated 180 degrees around (1/2,1/2)

In this section we show that the new coordinates of a point (θ_1, θ_2) that has been rotated 180 degrees around $(1/2, 1/2)$ is given by $(1 - \theta_1, \theta_2)$. From Lay (2012, p. 140) the new coordinates of a point (θ_1, θ_2) that has been rotated x degrees counter-clockwise around the origin is

$$\begin{pmatrix} \tilde{\theta}_1 \\ \tilde{\theta}_2 \end{pmatrix} = \begin{pmatrix} \cos(x) & \sin(x) \\ -\sin(x) & \cos(x) \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \tag{B.1}$$

To find the new coordinates of the rotated point we do three steps: 1) make a shift of coordinates where we move the original point to the origin, 2) use Equation (B.1) on the translated point and 3) transform the coordinates of the rotated point, which are expressed in the new coordinate system, to coordinates in the original system.

The coordinate in the new coordinate system are given by

$$(\psi_1, \psi_2) = (\theta_1 - \frac{1}{2}, \theta_2 - \frac{1}{2}) \tag{B.2}$$

If we want to go back to original coordinates we use the formula

$$(\theta_1, \theta_2) = (\psi_1 + \frac{1}{2}, \psi_2 + \frac{1}{2}) \tag{B.3}$$

When we rotate the point in the new coordinate system 180 degrees around the origin, the new coordinates become from Equation (B.1)

$$\begin{pmatrix} \tilde{\psi}_1 \\ \tilde{\psi}_2 \end{pmatrix} = \begin{bmatrix} \cos(180) & \sin(180) \\ -\sin(180) & \cos(180) \end{bmatrix} \begin{pmatrix} \psi_1 \\ \psi_2 \end{pmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \begin{pmatrix} \theta_1 - \frac{1}{2} \\ \theta_2 - \frac{1}{2} \end{pmatrix} = \begin{pmatrix} -\theta_1 + \frac{1}{2} \\ -\theta_2 + \frac{1}{2} \end{pmatrix} \tag{B.4}$$

In Equation (B.4) the coordinates of the rotated point are in the new coordinate system, but we have written them as a function of the original point (θ_1, θ_2). By using Equation (B.3) we get the coordinates of the rotated point in the original coordinate system

$$\begin{pmatrix} \tilde{\theta}_1 \\ \tilde{\theta}_2 \end{pmatrix} = \begin{pmatrix} \tilde{\psi}_1 \\ \tilde{\psi}_2 \end{pmatrix} + \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix} = \begin{pmatrix} 1 - \theta_1 \\ 1 - \theta_2 \end{pmatrix} \tag{B.5}$$

Appendix C

R-code used in Section 5

In this appendix we show an excerpt of the code used in the the power calculations. We show for instance how the A and C p-values can be calculated. For the E and M p-values we recommend using parallel programming.

```
cases=c(10,10)
cases=rbind(cases,c(25,25))
cases=rbind(cases,c(50,50))
cases=rbind(cases,c(97,103))
cases=rbind(cases,c(150,150))
cases=rbind(cases,c(4,16))
cases=rbind(cases,c(10,40))
cases=rbind(cases,c(20,80))
cases=rbind(cases,c(60,240))


for(kn in 1:dim(cases)[1])
{
  n1=cases[kn,1]
  n2=cases[kn,2]

# Create all possible tables
y<-expand.grid(seq(from=0,to=n1),seq(from=0,to=n2))


# write tables to file, which later will be used by the E, N
    and power programs
write.table(y,"tab.txt",row=F,col=F)

#Create table of the values of the test statistic
```

```
t2<-0
t2=(y[,1]/n1-y[,2]/n2)^2/((y[,1]+y[,2])/(n1+n2)*(1-(y[,1]+y
   [,2])/(n1+n2))*(1/n1+1/n2))
t2[is.na(t2)]=-999

write.table(t2,"z.txt",row=F,col=F)

# Calculate p_A and p_C and write to files that later will
   be used in power calculations:

# Calculate A:
avals=1-pchisq(t2,df=1)

#Calculate C:

cvals=rep(NA,length=length(t2))
for ( i in 1:length(t2))
{
  sufficient_statistic=y[i,1]+y[i,2]
  T_observed=t2[i]
  cond_outcomes_bool=y[,1]+y[,2]==sufficient_statistic & t2
     >=T_observed-10^(-10)
  cvals[i]=  sum(dhyper(y[cond_outcomes_bool,1],n1,n2,
     sufficient_statistic))
}


write.table(-avals,paste("Z-",n1,"-",n2,"-","A.txt",sep=""),
   row=F,col=F)
write.table(-cvals,paste("Z-",n1,"-",n2,"-","C.txt",sep=""),
   row=F,col=F)
```

Appendix D

Table of type I error probabilities

In Table D.1 we show the type I error probabilities at selected values of $\theta = \theta_1 = \theta_2$ and n_1, n_2, k studied in Chapter 5.

Table D.1: Type I error probabilities at selected $\theta = \theta_1 = \theta_2$ and significance levels $5 \cdot 10^{-k}$, where k can take the values $2, 3, \ldots, 8$. When the significance level is 5×10^{-k} the probabilities of type I error are divided by 10^{-k}. The first column specifies the sample sizes n_1 and n_2, the second column gives $\theta \cdot 100$, the third column gives the value of k in $5 \cdot 10^{-k}$ and column 4 to 9 gives respectively $\gamma_E(\theta, \theta; \alpha), \gamma_C(\theta, \theta; \alpha), \gamma_A(\theta, \theta; \alpha), \gamma_M(\theta, \theta; \alpha), \gamma_{CM}(\theta, \theta; \alpha)$ and $\gamma_{EM}(\theta, \theta; \alpha)$. We show all cases where one of the power functions exceeds the level and also a selection of other points. The different values are calculated using enumeration.

(n_1, n_2)	k	θ	E	A	M	C	CM	EM
(10,10)	2	5	0.00	0.00	0.00	0.00	0.00	0.00
		15	0.04	0.04	0.04	0.00	0.04	0.04
		25	0.58	0.58	0.58	0.04	0.58	0.58
		35	2.18	2.18	2.18	0.19	2.18	2.18
		45	3.77	3.77	3.77	0.37	3.77	3.77
		50	4.02	4.02	4.02	0.40	4.02	4.02
	6	5	0.00	0.00	0.00	0.00	0.00	0.00
		15	0.00	0.00	0.00	0.00	0.00	0.00
		25	0.11	0.00	0.11	0.00	0.11	0.11
		35	0.74	0.00	0.74	0.00	0.74	0.74
		45	1.72	0.00	1.72	0.00	1.72	1.72
		50	1.91	0.00	1.91	0.00	1.91	1.91

Continued on next page

(n_1, n_2)	k	θ	E	A	M	C	CM	EM
(25,25)	2	15	3.81	5.46	3.81	1.89	3.08	3.81
		20	3.92	5.49	3.91	2.22	3.77	3.91
		25	4.18	5.41	4.16	2.18	4.14	4.16
		30	4.63	5.37	4.48	2.29	4.48	4.48
		35	5.06	5.60	4.63	2.60	4.63	4.63
		40	5.08	6.02	4.36	2.95	4.36	4.36
		45	4.72	6.36	3.93	3.20	3.93	3.93
		50	4.49	6.49	3.73	3.28	3.73	3.73
	3	40	4.99	5.71	4.97	2.24	4.97	4.99
		45	4.89	6.36	4.89	2.50	4.89	4.89
		50	4.76	6.61	4.76	2.61	4.76	4.76
	8	5	0.00	0.00	0.00	0.00	0.00	0.00
		15	0.00	0.00	0.00	0.00	0.00	0.00
		25	0.32	0.02	0.32	0.04	0.32	0.32
		35	3.07	0.21	3.07	0.48	3.07	3.07
		45	4.98	0.45	4.98	0.94	4.98	4.98
		50	4.99	0.48	4.99	0.96	4.99	4.99
(50,50)	2	10	4.84	5.06	3.83	1.79	3.22	4.84
		15	4.74	5.50	4.18	2.39	3.61	4.74
		20	4.85	5.22	4.27	2.82	4.04	4.85
		25	4.74	5.07	4.47	3.02	4.08	4.74
		40	5.05	5.24	4.88	3.21	4.87	4.88
		45	4.93	5.57	4.52	3.43	4.52	4.52
		50	4.69	5.69	4.19	3.52	4.19	4.19
	3	25	4.83	5.06	4.18	2.62	4.18	4.83
		30	4.67	5.23	4.39	2.80	4.39	4.66
		35	4.80	5.25	4.68	2.81	4.68	4.72
		40	5.17	5.52	4.81	3.03	4.81	4.81
		45	4.85	5.45	4.32	3.37	4.32	4.32
		50	4.45	5.18	3.97	3.52	3.97	3.97
	5	30	5.00	3.40	4.19	2.18	4.26	4.26
	8	5	0.00	0.00	0.00	0.00	0.00	0.00
		15	0.25	0.01	0.02	0.06	0.25	0.33
		25	2.51	0.64	0.85	0.82	2.32	3.74
		35	4.48	1.80	2.75	1.50	2.98	4.55
		45	4.37	2.38	4.01	1.59	4.02	4.37
		50	4.20	2.19	4.11	1.66	4.11	4.20
(150,150)	2	5	4.84	4.84	4.84	2.43	3.44	4.84
		10	4.79	4.98	4.79	3.30	3.87	4.79

Continued on next page

(n_1, n_2)	k	θ	E	A	M	C	CM	EM
		15	4.76	4.93	4.76	3.41	4.36	4.76
		20	4.90	5.03	4.75	3.55	4.48	4.74
		25	4.90	5.04	4.83	3.73	4.64	4.68
		30	4.97	5.03	4.78	3.77	4.74	4.74
		35	4.89	4.89	4.71	3.69	4.71	4.71
		40	5.16	5.17	4.94	3.91	4.94	4.94
		45	5.16	5.53	4.48	4.21	4.48	4.48
		50	4.64	5.66	4.33	4.31	4.33	4.33
	3	30	4.88	5.06	4.81	3.48	4.81	4.18
		35	5.00	5.02	4.82	3.48	4.82	4.32
		40	5.04	5.17	4.88	3.61	4.88	4.24
	4	10	5.06	3.54	3.54	2.33	3.54	4.71
		25	5.02	4.86	4.60	3.07	4.46	4.46
		40	5.08	5.05	4.91	3.30	4.91	4.83
		45	5.05	4.85	4.65	3.54	4.65	4.46
	5	45	5.12	4.91	4.69	3.52	4.69	4.69
	6	35	5.01	4.55	4.79	2.94	4.79	4.79
		40	5.10	4.63	4.94	2.95	4.94	4.94
	7	15	5.35	1.89	2.19	2.24	3.75	4.89
		20	5.07	2.70	3.29	2.48	4.47	4.15
	8	5	0.19	0.00	0.00	0.05	0.05	0.19
		15	3.80	1.12	1.70	1.83	3.48	3.80
		25	4.79	2.97	3.21	2.31	4.08	4.75
		30	5.03	3.27	3.96	2.56	4.28	4.69
		35	4.85	3.65	4.19	2.57	4.70	4.71
		45	5.06	4.03	4.99	2.83	4.99	4.99
		50	4.60	3.46	4.60	3.19	4.60	4.60
(4,16)	2	5	1.15	8.69	1.15	0.64	1.15	1.15
		10	2.85	8.25	2.85	1.19	2.85	2.85
		15	3.73	6.47	3.73	1.42	3.73	3.73
		20	3.86	5.07	3.86	1.48	3.86	3.86
		45	3.34	5.36	3.34	1.41	3.34	3.34
		50	3.39	5.57	3.39	1.42	3.39	3.39
	3	5	0.39	6.35	0.39	0.22	0.39	0.39
		10	1.95	10.95	1.95	0.75	1.95	1.95
		15	3.66	10.90	3.66	1.14	3.66	3.66
		20	4.56	8.88	4.56	1.28	4.56	4.56
		25	4.56	6.67	4.56	1.24	4.56	4.56
	4	10	0.51	7.19	0.51	0.19	0.51	0.51

Continued on next page

(n_1, n_2)	k	θ	E	A	M	C	CM	EM
		15	1.44	9.96	1.44	0.38	1.44	1.44
		20	2.25	9.46	2.25	0.45	2.25	2.25
		25	2.48	7.18	2.48	0.39	2.48	2.48
	5	5	0.28	0.28	0.28	0.00	0.28	0.28
		15	3.76	3.76	3.76	0.00	3.76	3.76
		25	3.92	3.92	3.92	0.00	3.92	3.92
		35	1.52	1.52	1.52	0.00	1.52	1.52
		45	0.31	0.31	0.31	0.00	0.31	0.31
		50	0.19	0.19	0.19	0.00	0.19	0.19
(10,40)	2	5	3.79	7.84	3.79	1.77	1.77	3.79
		25	4.70	4.70	4.09	3.35	4.29	4.70
		40	4.75	5.11	4.25	2.67	4.39	4.75
		45	4.68	5.40	4.16	2.66	3.78	4.68
		50	4.61	5.51	4.11	2.68	3.49	4.61
	3	5	2.05	14.49	2.05	0.44	1.79	2.05
		10	4.10	11.04	4.10	1.65	2.51	4.10
		15	4.13	7.05	4.12	2.25	2.67	4.13
		20	3.71	5.27	3.59	2.31	3.16	3.71
	4	5	0.90	15.02	0.90	0.040	0.90	0.90
		10	3.50	21.73	3.50	0.59	3.50	3.50
		15	3.88	15.06	3.87	1.13	3.88	3.88
		20	3.82	10.56	3.70	1.43	3.82	3.82
		25	3.60	7.78	3.08	1.55	3.60	3.60
		30	3.49	5.48	2.28	1.54	3.49	3.49
	5	5	0.90	15.02	0.90	0.04	0.90	0.90
		10	3.71	32.12	3.71	0.42	3.71	3.71
		15	4.82	26.22	4.81	1.04	4.82	4.82
		20	4.84	17.74	4.77	1.45	4.84	4.84
		25	4.36	12.24	3.99	1.54	4.36	4.36
		30	3.78	8.04	2.78	1.49	3.78	3.78
	6	5	0.38	8.95	0.38	0.01	0.38	0.38
		10	2.90	35.27	2.90	0.22	2.90	2.90
		15	4.63	36.39	4.63	0.69	4.63	4.63
		20	4.72	24.64	4.69	1.06	4.72	4.73
		25	4.08	15.30	3.83	1.17	4.08	4.22
		30	3.34	9.05	2.51	1.16	3.34	4.00
	7	10	1.63	28.27	1.63	0.07	1.63	1.63
		15	3.35	38.95	3.35	0.31	3.37	3.36
		20	3.45	28.07	3.45	0.59	3.74	3.58

Continued on next page

(n_1, n_2)	k	θ	E	A	M	C	CM	EM
		25	2.62	15.52	2.62	0.74	3.95	3.13
		30	1.56	7.69	1.56	0.66	4.31	2.45
	8	10	0.74	16.25	0.74	0.08	0.62	0.74
		15	2.91	33.41	2.91	0.47	1.80	2.91
		20	4.52	33.67	4.52	0.89	2.36	4.52
		25	3.74	23.64	3.74	0.86	2.33	3.74
		30	1.86	11.99	1.86	0.51	1.76	1.86
		40	0.13	1.04	0.13	0.05	0.31	0.13
		50	0.00	0.04	0.00	0.00	0.02	0.00
(20,80)	2	15	5.19	4.67	4.19	2.84	3.98	4.75
		20	4.79	5.01	4.58	3.23	4.23	4.69
		25	4.85	5.06	4.59	3.22	4.60	4.83
		35	4.97	4.97	4.35	3.16	4.65	4.72
		45	4.74	5.01	4.18	3.86	3.95	4.69
		50	4.63	5.07	4.08	3.96	3.97	4.62
	3	5	3.36	8.40	3.36	0.95	1.88	3.36
	4	5	3.47	18.76	3.41	0.78	1.21	3.47
		10	4.39	12.67	2.74	1.87	2.64	4.39
		15	4.89	10.05	1.77	2.35	2.72	4.89
		20	4.36	7.43	1.09	2.29	3.44	4.36
		25	4.33	5.31	0.85	2.25	4.22	4.33
		40	4.65	3.93	0.41	2.41	4.69	4.65
		50	4.43	3.95	0.33	2.55	4.01	4.43
	5	5	4.59	29.78	4.59	0.83	0.88	4.59
		10	3.85	16.69	3.27	1.70	2.59	3.85
		15	4.76	12.39	2.54	2.05	2.87	4.76
		20	4.89	9.25	1.92	2.02	3.32	4.89
		25	4.77	6.91	1.25	1.92	3.50	4.77
		30	4.82	5.79	0.75	1.95	3.70	4.82
		50	4.21	2.51	0.15	2.35	4.53	4.21
	6	5	0.83	45.89	0.83	0.61	0.79	0.83
		10	3.86	32.74	3.85	1.44	2.56	3.86
		15	4.57	25.40	4.18	1.52	3.41	4.57
		20	4.32	19.18	2.41	1.54	3.57	4.32
		25	4.26	12.51	1.24	1.61	3.29	4.26
		30	3.90	7.54	0.75	1.67	3.15	3.90
		50	4.47	1.50	0.05	1.59	4.42	4.47
	7	5	0.63	57.07	0.63	0.04	0.63	0.63
		10	3.29	57.87	3.29	0.60	3.30	3.30

Continued on next page

(n_1, n_2)	k	θ	E	A	M	C	CM	EM
		15	3.74	42.80	3.69	1.61	4.04	4.04
		20	3.35	24.07	2.60	2.01	4.39	4.39
		25	3.92	12.36	1.52	1.63	4.60	4.60
		30	4.26	7.51	0.78	1.39	4.40	4.40
		50	3.58	0.54	0.02	1.95	4.78	3.58
	8	5	0.41	57.84	0.41	0.02	0.41	0.41
		10	3.79	92.49	3.79	0.51	3.51	3.51
		15	5.00	71.00	4.97	1.18	3.49	3.45
		20	3.79	42.03	3.36	1.38	3.31	2.84
		25	2.91	19.81	1.42	1.48	3.74	2.58
		30	3.23	10.45	0.55	1.35	3.97	2.50
		35	3.37	5.35	0.22	1.36	4.28	2.57
		50	1.84	0.29	0.00	1.29	2.77	1.84
(60,240)	3	5	5.57	4.49	3.66	3.12	3.45	4.49
		35	4.96	5.01	4.55	4.11	4.88	4.71
		40	4.95	5.01	4.51	3.71	4.69	4.78
		45	5.05	5.18	4.53	3.92	4.03	4.69
		50	4.87	5.29	4.62	4.02	4.02	4.63
	3	5	3.91	5.98	3.65	2.05	2.96	3.91
		10	5.02	5.09	3.61	3.02	3.78	4.75
		15	4.99	5.03	3.45	3.39	4.07	4.93
		35	4.93	4.85	3.40	3.91	4.40	4.93
		50	4.92	4.91	3.38	3.38	4.16	4.92
	4	5	4.28	10.62	2.92	1.94	2.99	4.28
		10	4.58	7.32	1.73	2.43	3.29	4.58
		15	4.79	5.90	1.41	2.90	3.78	4.78
		20	5.10	5.46	1.15	3.06	4.14	4.80
		25	4.75	5.24	0.99	3.28	4.33	4.61
		40	4.86	4.69	0.78	3.43	4.73	4.86
		50	4.65	4.40	0.81	4.25	4.26	4.65
	5	5	4.94	17.76	4.12	1.83	3.14	4.69
		10	4.20	10.28	2.36	2.33	3.27	3.96
		15	5.11	8.25	1.67	2.71	3.63	4.35
		20	4.67	6.74	1.18	2.84	4.08	4.59
		25	4.40	5.83	0.85	3.11	4.37	4.39
		30	4.81	5.02	0.67	3.18	4.30	4.68
		40	4.81	4.41	0.53	3.07	4.87	4.74
		50	4.86	3.88	0.43	3.73	3.79	4.85
	6	5	4.65	30.62	3.59	1.64	1.64	4.65

Continued on next page

(n_1, n_2)	k	θ	E	A	M	C	CM	EM
		10	5.18	18.20	1.52	2.72	2.73	4.86
		15	4.61	13.38	1.02	2.38	2.83	4.48
		20	4.76	9.48	0.54	2.54	3.21	4.66
		25	5.24	7.63	0.36	2.60	3.70	4.93
		30	5.00	6.38	0.24	2.71	3.52	4.75
		40	4.77	3.92	0.13	3.28	4.05	4.72
		50	4.77	3.39	0.09	2.66	4.99	4.57
	7	5	5.00	54.01	4.90	1.50	2.57	3.64
		10	4.08	35.27	2.71	1.93	3.73	4.06
		15	4.61	20.33	1.64	2.27	3.94	4.61
		20	4.63	13.57	0.82	2.43	3.84	4.62
		25	4.79	9.12	0.49	2.56	3.86	4.51
		30	4.97	7.21	0.31	2.58	4.18	4.77
		40	4.72	3.72	0.11	2.93	4.95	4.72
		50	4.64	2.93	0.07	2.83	4.65	4.64
	8	5	4.61	120.54	4.61	1.10	2.79	4.61
		10	4.64	60.03	3.10	2.17	3.81	4.54
		15	4.88	31.20	1.36	2.44	4.13	4.49
		20	4.57	20.01	0.71	2.24	4.57	4.56
		25	4.92	13.16	0.42	2.23	4.92	4.59
		30	4.78	7.62	0.24	2.32	4.80	4.46
		40	4.63	3.55	0.07	2.74	4.94	4.10
		50	4.34	2.16	0.03	3.54	4.73	4.31
(97,103)	2	5	5.01	5.08	4.89	2.85	3.52	4.51
		15	4.98	4.98	4.80	3.90	4.46	4.59
		25	4.95	5.06	4.92	4.09	4.89	4.70
		30	4.98	5.15	4.79	4.16	4.87	4.58
		45	4.84	4.94	4.75	4.64	4.66	4.62
		50	4.96	5.26	4.94	4.76	4.76	4.59
	3	30	5.06	5.06	4.74	3.88	4.75	4.75
		35	5.15	5.23	4.84	3.95	4.92	4.97
		50	4.84	5.26	4.83	4.59	4.60	4.84
	4	40	4.95	5.06	4.75	3.79	4.93	4.78
		45	4.96	5.06	4.81	3.66	4.64	4.81
		50	5.02	5.05	4.96	3.74	4.08	4.96
	5	10	5.46	1.90	1.51	2.61	3.43	4.54
		15	5.09	2.91	2.61	3.18	3.54	4.68
		20	5.12	3.58	3.44	3.16	4.08	4.44
		35	4.82	4.40	4.12	3.61	4.59	4.25

Continued on next page

(n_1, n_2)	k	θ	E	A	M	C	CM	EM
		45	4.98	4.98	4.85	3.54	4.66	4.26
		50	5.03	5.03	5.00	3.58	4.10	4.46
	6	15	5.60	2.03	2.28	2.90	3.86	4.82
	7	15	5.06	1.05	1.26	2.58	3.61	4.51
		25	5.02	2.77	3.66	3.25	4.29	4.84
		35	5.03	3.79	4.75	3.33	4.29	4.97
		45	4.85	4.06	4.85	3.27	4.74	4.75
		50	4.72	4.06	4.72	2.97	4.71	4.71

Bibliography

Aanes, F. L. (2016), *The kinship coefficient: a statistical approach*, Department of Mathematical Sciences, Norwegian University of Science and Technology.

Adams, R. A. & Essex, C. (2010), *Calculus. A Complete Course*, 7 edn, Pearson.

Altman, D. (1991), *Practical Statistics for Medical Research*, 1 edn, Chapman & Hall/CRC.

Ambrosius, W. T., ed. (2007), *Topis in Biostatistics*, Methods in Molecular Biology, 1 edn, Humana Press.

Bakke, Ø. & Langaas, M. (n.d.), Increased test power with unconditional maximization enumeration as a post-processing step in discrete distributions.

Balding, D., Bishop, M. & Cannings, C., eds (2003), *Handbook of Statistical Genetics Volume 2*, 2 edn, Wiley.

Bayarri, M. & Berger, J. (2000), 'P Values for Composite Null Models', *Journal of the American Statistical Association* **95**(452), 1127–1142.

Berger, R. L. & Boos, D. D. (1994), ' P Values Maximized Over a Confidence Set for the Nuisance Parameter', *Journal of the American Statistical Association* **89**(427), 1012–1016.

Bland, J. M. & Altman, D. G. (2004), 'The logrank test.', *BMJ (Clinical research ed.)* **328**(7447), 1073.

Borgan, Ø. (2007), 'Hypotestesting. Notat til STK1110', Available from `http://www.uio.no/studier/emner/matnat/math/STK1110/h08/Annet/test08.pdf`.

Boschloo, R. (1970), 'Raised conditional level of significance for the 2×2-table when testing the equality of two probabilities', *Statistica Neerlandica* **24**(1).

Casella, G. & Berger, R. L. (2002), *Statistical Inference, International Student Edition*, Duxbury Advanced Series, 2 edn, Brooks/Cole Cengage Learning.

Christensen, R. (1998), *Analysis of Variance, Design, and Regression: Applied Statistical Methods*, 1 edn, CRC Press LLC.

Devore, D. L., Berk, K. N. & Olkin (2012), *Modern Mathematical Statistics with Applications*, Springer Texts in Statistics, 2 edn, Springer.

Dudbridge, F. & Gusnanto, A. (2008), 'Estimation of significance threshold for genomwide association scans ', *Genetic Epidemiology* **32**, 227–234.

Genetics Home Reference (2015), 'Phenotype'. [Online; accessed 10-November-2015].
URL: *http://http://ghr.nlm.nih.gov/glossary=phenotype*

Georgi, H.-O. (2014), *Using R for Introductory Statistics*, The R Series, 2 edn, CRC Press.

Günther, C. C., Bakke, Ø., Rue, H. & Langaas, M. (2009), *Statistical hypothesis testing for categorical data using enumeration in the presence of nuisance parameters*, Preprint statistics no. 4/2009, Department of Mathematical Sciences, Norwegian University of Science and Technology.

Halliburton, R. (2004), *Introduction to Population Genetics*, 1 edn, Pearson Prentice Hall.

Hogg, R. V., McKean, J. & Craig, A. T. (2014), *Introduction to Mathematical Statistics, Pearson New International Edition*, 7 edn, Pearson Education Limited.

Høyland, A. (1986), *Del 2, Statistisk metodelære*, 4 edn, Tapir.

Kateri, M. (2014), *Contingency Table Analysis, Methods and Implementation Using R* , Statistics for Industry and Technology, 1 edn, Birkhäuser.

Kiess, H. (2007), 'Difficult concepts: Research hypotheses vs. statistical hypotheses'. [Online, accessed 18.03.2016].
URL: *https://statisticalsage.wordpress.com/2011/09/21/difficult-concepts-research-hypotheses-vs-statistical-hypotheses/*

Krishnamoorthy, K. (2015), *Handbook of Statistical Distributions with Applications*, 2 edn, Chapman and Hall/CRC.

Krishnamoorthy, K. & Thomson, J. (2004), 'A more powerful test for comparing two Poisson means ', *Journal of Statistical Planning and Inference* **119**(1), 23–35.

Lay, D. C. (2012), *Linear Algebra and Its Applications*, 4 edn, Addison-Wesly.

Lin, M., Lucas Jr, H. C. & Shmueli, G. (2013), 'Research Commentary–Too Big to Fail: Large Samples and the p-Value Problem', *Information Systems Research* **24**(4), 906–917.

Lloyd, C. J. (2008), 'Exact p-values for discrete models obtained by estimation and maximization', *Australian & New Zealand Journal of Statistics* **50**(4), 329–345.

Lydersen, S., Langaas, M. & Bakke, y. (2012), '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* **82**(9), 1311–1316.

Mackay, T. F., Stone, E. A. & Ayroles, J. F. (2009), 'The genetics of quantitative traits: challenges and prospects', *Nat Rev Genet* **10**(8), 565–577.

Mehrotra, D. V., Chan, I. S. & Berger, R. L. (2004), 'A cautionary note on exact uncondtional inference for a difference between two independent binomial proportions', *Biometrics* **59**(2), 441–450.

Meyskens, F. L., McLaren, C. E., Pelot, D., Fujikawa-Brooks, S., Carpenter, P. M., Hawk, E., Kelloff, G., Lawson, M. J., Kidao, J., McCracken, J., Albers, C. G., Ahnen, D. J., Turgeon, D. K., Goldschmid, S., Lance, P., Hagedorn, C. H., Gillen, D. L. & Gerner, E. W. (2008), 'Difluoromethylornithine plus sulindac for the prevention of sporadic colorectal adenomas: A randomized placebo-controlled, double-blind trial', *Cancer Prevention Research* **1**(1), 32–38.

Mukhopadhyay, N. (2000), *Probability and Statistical Inference*, Statistics: A Series of Textbooks and Monographs, 1 edn, Marcel Dekker.

National Human Genome Research Institute (2015*a*), 'Chromosomes'. [Online; accessed 01-November-2015].
URL: *https://www.genome.gov/26524120\#top*

National Human Genome Research Institute (2015*b*), 'DNA'. [Online; accessed 01-November-2015].
URL: *http://ghr.nlm.nih.gov/handbook/basics/dna*

Oyana, T. & Margai, F. (2016), *Spatial Analysis: Statistics, Visualization, and Computational Methods*, 1 edn, CRC Press, Taylor & Francis Group.

Panagiotou, O. A. & Ioannidis, J. P. A. (2012), 'What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations', *International Journal of Epidemiology* **41**, 273–286.

R Core Team (2014), *R: A Language and Environment for Statistical Computing*,

R Foundation for Statistical Computing, Vienna, Austria.
URL: *http://www.R-project.org/*

Roman, S. (2012), *Introduction to the Mathematics of Finance, Arbitrage and Option Pricing*, Undergraduate Texts in Mathematics, 2 edn, Springer.

Royall, R. M. (1997), *Statistical Evidence: A likelihood paradigm*, Monographs on Statistics and Applied Probability 71, 1 edn, Chapman & Hall.

Sakpal, T. V. (2004), 'Sample Size Estimation in Clinical Trial', *Perspectives in Clinical Research* **1**(2), 67–69.

Silva, I. d. S. (1999), *Cancer Epidemiology: Principles and Method*, 1 edn, International Agency for Research on Cancer.

Stuart, A. & Ord, J. K. (1991), *Kendall's advanced theory of statistics : 2 : Classical inference and relationship*, 5 edn, Edward Arnold, London.

Taboga, M. (2010), 'Hypothesis testing', Available from `http://www.statlect.com/hypothesis_testing.htm`.

The Editors of Encyclopædia Britannica (2016), 'Gamete'.
URL: *http://global.britannica.com/science/gamete*

Thompson, E. A. (1986), *Pedigree Analysis in Human Genetics*, 2 edn, The Johns Hopkins University Press.

Verbeek, A. & Kroonenberg, P. M. (1985), 'A survey of algorithms for exact distributions of test statistics in $r \times c$ contingency tables with fixed margins', *Computational Statistics and Data Analysis* **3**(C), 159–185.

Verzani, J. (2013), *Stochastics. Introduction to Probability and Statistics*, 2 edn, De Gruyter.

Wacholder, S., Silverman, D. T., McLaughlin, J. K. & Mandel, J. S. (1992), 'Selection of Controls in Case-Control Studies', *American Journal of Epidemiology* **135**(9), 1042–1050.

Walpole, R. E., Myers, R. H., Myers, S. L. & Ye, K. (2012), *Probability & Statistics for Engineers and Scientists*, 9 edn, Pearson.

Wasserstein, R. L. & Lazar, N. A. (2016), 'The ASA's statement on p-values: context, process, and purpose', *The American Statistician* **1305**(March), 1–17.
URL: *http://www.tandfonline.com/doi/full/10.1080/00031305.2016.1154108*

Zhao, Y., Zhang, F.J., Z. S., Duan, H., Y., L., Zhou, Z., W.X., M. & Wang, L. (2011), 'The association of a single nucleotide polymorphism in the promoter re-

gion of the LAMA1 gene with susceptibility to Chinese high myopia', *Molecular Vision* **17**, 1003–1010.

Ziegler, A. & König, I. R. (2010), *A Statistical Approach to Genetic Epidemiology*, 2 edn, Wiley-Blackwell.