Olai Gaarn Skogen

Comparison of p-values for two-sided hypothesis test with non-symmetric distribution

An analysis based on power

Specialization Project in Industrial Mathematics
Supervisor: Øyvind Bakke
January 2025

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

NTNU

This template is
updated and expanded
by Carl Fredrik Berg,
based on an earlier
template created
by Nina Salvesen.

Abstract

P-values are a staple of modern research. Often two-sided hypothesis tests are employed. For symmetric distributions this is perhaps most commonly calculated by doubling the minimal one-sided p-value. For non-symmetric distributions this will produce a valid p-value, but with room for improvment. In particular, the p-value based on the minimal tail probability under the null hypothesis will produce strictly more powerful tests while guaranteeing the same significance level α. In this thesis we have investigated the power of four different p-values for hypothesis tests concerning the binomial and Poisson parameter, as well as Fisher's exact test. The p-values investigated were (i) double the one-sided p-value, (ii) point probability as test statistic, (iii) tail probability as test statistic and (iv) distance from mean as test statistic. We found that the point and tail p-value generally produce more powerful hypothesis tests, with the tail probability also being better at producing unbiased tests. At the time of writing, the R programming default is using the point probability as test statistic and this raises the question of if we should take a closer look at our p-values.

Contents

List of Figures

Abbreviations

List of all abbreviations in alphabetic order:

- **CDF** Cumulative Density Function
- **CLT** Central Limit Theorem
- **iid** Independantly, identically distributed
- **LRT** Likelihood Ratio Test
- **NHST** Null Hypothesis Significance Testing
- **PDF** probability Denisty Function
- **PMF** Probability Mass Function
- **UMP** Uniformly Most Powerful

Chapter 1

Introduction

1.1 Motivation

P-values can be understood as the probability of experiencing an event that is as extreme or more than the observed outcome given the null hypothesis. When our null and alternative hypothesis are one-sided, defining extremity becomes simple. Say we observe the test statistic $T(\mathbf{X})$. Higher values of $T(\mathbf{x})$ will always be seen as a less[1] probable outcome than $T(\mathbf{y})$ when $T(\mathbf{y}) < T(\mathbf{x})$. This leads to a natural rejection criteria of rejection when $T(\mathbf{y})$ is less than some critical value c. However, for a two-sided hypothesis test knowing $T(\mathbf{y}) < T(\mathbf{x})$ will not necessarily tell you which is less or more probable. This means that we need to find a way to rank outcomes between the area in which $T(\mathbf{y}) < T(\mathbf{x})$ means $\mathbf{y}$ is less probable against the area in which it means the opposite. How to rank outcomes between the two regions is not at all obvious.

Let us give a motivating problem. Suppose we have a six sided die and we are particularly interested in making sure that rolling a 6 happens with probability 1/6. Two different experimenters, Anna and Bob, are rolling it 10 times each to investigate this. They both assume the null hypothesis that the die has probability $\theta = 1/6$ to get a 6. Anna never rolls a 6, meanwhile Bob has rolled 6 three times. Which experimenter has then observed the more extreme result given the null hypothesis $H_0 : \theta = 1/6$?

While this problem begins to illustrate what we want to study it might seem a bit artifical and far from reality. So let us consider that Anna and Bob might be researchers who both have spent a year rolling dice and are desperate to show that they have produced the more extreme, and therefore more scientifically relevant, result. Anna begins by arguing that the expected

[1] Or more probable depending on the direction of the hypotheses.

number of 6 rolled is 1.67 and therefore her 0 rolls is further away and should be considered more extreme. This annoys Bob until he realizes that the chance of rolling no 6 given a fair dice is 0.1615 while the chance of rolling three 6 is 0.1550. Thus his result is more extreme. Then Anna finds a new way to claim her result is more extreme and the cycle continues.

Obviously, Anna and Bob are engaging in rampant ad hoc fallacies. It is very bad to change the rejection criteria of a hypothesis test after it has been carried out. But their bickering does highlight something important. Anna and Bob should have decided upon a common way of defining extremity before the experiment started and now it is hard to find a good case for why a one test statistic is better than another. Defining extremity is simply not obvious in general.

1.1.1 Purpose

The purpose of this thesis is to gain more insight into what the agreed upon way of defining extremity should be. This will be done by comparing different ways of defining extremity. We will place an emphasis on p-values and hypothesis tests based on them as they provide a natural framework for the concept of extremity. We will focus mainly on the power produced by tests, as well as looking at unbiasedness when power alone fails to give actionable results. Computing time of different methods could have been interesting but it is not focused on in this thesis.

1.2 Outline

We will first introduce theory concerning hypothesis tests in general. We will clearly define what we mean by two-sided hypothesis tests and outline why we are particularly interested in non-symmetric distributions. We will also define the concept of power and introduce the concept of unbiasedness. Finally, in the hypothesis test chapter, we will clearly define the experiments we will investigate in this thesis; (i) binomial experiment, (ii) Fisher's exact test and (iii) Poisson experiment.

We will then introduce theory on p-values and show a general framework for defining different two-side p-values with a new test statistic $W(\mathbf{X})$ called an extremity statistic[2]. We will then construct two-sided p-values for the different experiments. These p-values are (i) double tail p-value, (ii) point probability p-value, (iii) tail probability p-value and (iv) distance p-value.

[2]Extremity statistic is a term that is only used in this thesis because it is helpful to distinguish different test statistics. It is not commonly used in literature.

In our discussion section we will compare the power functions we calculated for the different combinations of experiment setup and p-value. These results revealed that the point and tail p-values performed noticably better than the double tail and distance p-values. The tail and point p-values are then looked at in greater detail. The power functions did not give particular preferance to one over the other, but we saw some instances of the tail p-value producing more unbiased tests. We then give some suggestions to further research.

Chapter 2

Hypothesis Testing

The scientific method is built upon constantly making hypotheses and testing them. As a hypothesis holds up to repeated experiments it gains more and more credibility. Meanwhile, they are discarded in favor of alternative hypotheses if they fail tests significantly often. To determine exactly what this significant threshold is, researchers employ statistics.

2.1 General principles

Lets take a closer look at the mathematical setup of the null hypothesis significance testing (NHST). In general, a null hypothesis significance test will consist of a null hypotheis and a rejection criterion. The null hypothesis needs to specify the distribution of some observable data while the rejection criterion decides when you reject or fail to reject the null hypotheis. It's usual to define the rejection criteria such that the probability of falsely rejecting a true hypotheis is below a certain threshold α. This probability is often called Type I error and its probability will be equal or below the significance level α of a NHST.

In addition to the null hypothesis, an alternative hypotheis is often given. Its necessity in NHST has historically been disputed (Hubbard and Bayarri, 2003), however it is common to state both your null and alternative hypotheses when doing research today. One example of an hypothesis pair could be the null hypothesis $H_0 : \theta \geq \theta_0$ versus the alternate hypothesis $H_1 : \theta < \theta_0$ where θ is some unknown parameter relevant to an observable stochastic variable. Usually, the alternative hypothesis is set to be the complement of the null hypothesis. So given the parameter space Θ for the parameter θ we have $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^C$ as the hypothesis pair.

2.1.1 Two-sided hypotheses

We distinguish between one-sided and two-sided hypothesis test. One-sided hypothesis tests are only interested in whether or not the parameter θ is above or below a certain value θ_0 and have hypotheses on the following form:

$$H_0 : \theta \leq \theta_0 \quad \text{versus} \quad H_1 : \theta > \theta_0 \tag{2.1}$$

$$\text{or} \quad H_0 : \theta \geq \theta_0 \quad \text{versus} \quad H_1 : \theta < \theta_0 \ . \tag{2.2}$$

While two-sided hypothesis test want to test if the parameter θ is equal (or close to) a value θ_0 and have hypotheses on the following form:

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0 \ . \tag{2.3}$$

There are also other two-sided hypothesis pairs, but we will not be concerned with them in this thesis. When we refer to a two-sided hypothesis test we mean the situation in Eq. 2.3.

Note that while a one-sided hypothesis test often leads to a one-sided rejection region for a test statistic and a two-sided hypothesis test often leads to a two-sided rejection region it is not necessairly the case. We will quickly look at an example.

One-sided rejection region from two-sided hypothesis Consider the normally distributed random sample $X_1, X_2, ...X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, where the variance σ^2 is known and we construct a hypothesis test for the unknown μ. We employ a two-sided hypothesis test $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$.

An unbiased estimator for the parameter μ would be the sample mean, $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. We want rejection for both very small and very large values of $\overline{X}$ as cause for rejcting the null hypotheis as both are unlikely under the null hypothesis. Thus, $\overline{X} < a$ and $\overline{X} > b$, where $a, b \in \mathbb{R}$ and $a < b$, is a natural rejection criteria. This is a two-sided rejection region $R = \{\mathbf{X} : \overline{X} \in (-\infty, a) \cup (b, \infty)\}$.

If we have the symmetric case of $a = \mu_0 - b$ we may define our test statistic as $\mid \overline{X} - \mu_0 \mid$ and obtain $\mid \overline{X} - \mu_0 \mid > \epsilon$ as our new rejection criteria. Thus, we have an example of a one-sided rejection region $R = \{\mathbf{X} :\mid \overline{X} - \mu_0 \mid \in (\epsilon, \infty)\}$ from a two-sided hypothesis. This required looking at a new test statistic, but in the case where the origial test statistic was symmetric it was very natural to do. We shall see that symmetric distributions generally lead to natural measures of extremity coinciding. Thus removing the need to define extremity rigorously.

2.1.2 Level and size α tests

Up until now we have used the words *significance level* without any definition. As previously mentioned, the significance level is an upper bound to the Type I Error. It is possible to drop the word *significance* and instead just talk about *level*, but one will often use both to clarify what we are talking about. The definition for *level* is as follows.

Definition 2.1 *From Casella and Berger (2002, p.385). Let the null hypothesis be $H_0 : \theta \in \Theta_0$ and let the rejection criteria of H_0 be $\mathbf{X} \in R$. Then, for $0 \leq \alpha \leq 1$, a test is a **level** α **test** if*

$$\sup_{\theta \in \Theta_0} P_\theta(\mathbf{X} \in R) \leq \alpha \ . \tag{2.4}$$

This guarantees that the probability of making a Type I Error is below α as all rejection probabilities for $\theta \in \Theta_0$ are below the supremum which again is smaller than α. In the case where the null hypothesis consists of a single point, i.e. $\Theta_0 = \{\theta_0\}$, we may remove the supremum as $\sup_{\theta \in \Theta_0} P_\theta(\cdot) = P_{\theta_0}(\cdot)$. In practice, the supremum produces an optimization problem, and for two-sided null hypotheses we do not need to solve it.

We might have the case where the supremum of the Type I Error is equal to α. We call this a size α test.

Definition 2.2 *From Casella and Berger (2002, p.385). Let the null hypothesis be $H_0 : \theta \in \Theta_0$ and let the rejection criteria of H_0 be $\mathbf{X} \in R$. Then, for $0 \leq \alpha \leq 1$, a test is a **size** α **test** if*

$$\sup_{\theta \in \Theta_0} P_\theta(\mathbf{X} \in R) = \alpha \ . \tag{2.5}$$

We often prefer tests that are size tests to level tests as it allows for more precise control of the Type I Error to other concerns, namely the probability of not rejecting the null hypothesis when it is false. This is called Type II Error. Note that a size α test is also a level α test and thus we will talk about a level test when we have not specified it further.

2.1.3 Power function

The power function is very useful in describing the tradeoff between Type I and Type II Error. It is defined as the probability of rejecting the null hypothesis with respect to the parameter.

Definition 2.3 *From (Casella and Berger, 2002, p.383). The* ***power function*** *of a hypothesis test with rejection region R is the function of θ defined by $\beta(\theta) = P_\theta(\mathbf{X} \in R)$.*

This leads to the following relationship between the errors and the power function (Casella and Berger, 2002, p.383)

$$\beta(\theta) = \begin{cases} \text{probability of a Type I Error} & \text{if } \theta \in \Theta_0 \\ \text{one minus the probability of a Type II Error} & \text{if } \theta \in \Theta_0^C, \end{cases}$$

thus it will be of great interest when we are studying the properties of different ways of calulating the p-value.

2.1.4 Evaluation of Hypothesis Tests

Hypothesis tests are usually evaluated by the probabilities of making mistakes (Casella and Berger, 2002, p. 382), i.e. the Type I and II Errors. We usually start by requiring that both are level α tests and then compare the power function $\beta(\theta)$. In other words, we are controlling Type I Error and then comparing Type II Error, with higher $\beta(\theta)$ signfiying more rejection power and less Type II Error.

This property being optimal is formalized with the definition for the most powerful test (UMP) of a class of tests (Casella and Berger, 2002, p.388)

Definition 2.4 *Let $\mathcal{C}$ be a class of tests for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^C$. A test in class $\mathcal{C}$, with power function $\beta(\theta)$, is a uniformly most powerful (UMP) class $\mathcal{C}$ test if $\beta(\theta) \geq \beta'(\theta)$ for every $\theta \in \Theta_0^C$ ad every $\beta'(\theta)$ that is a power function of a test in class $\mathcal{C}$.*

For this thesis we will consider a test to be better than another if it is always more powerful than other tests with the same level α. One could imagine situations where the expected type 1 error is lower and therefore one test might be better even if another is UMP of it, but we will not look into such cases. This heuristic is backed up by Casella & Berger which state that when a UMP level α test exists "one may as well consider it the best in its class" (Casella and Berger, 2002, p.388).

In some cases, UMP tests of its class are relatively simple to find. Especially, for one-sided tests. We will now see that there are theorems that transforms the problem of finding a UMP level α test for one-sided hypothesis tests into one of finding sufficient statistics with a monotone likelihood ratio (MLR). For one sided hypothesis tests we have the Karlin-Rubin Theorem.

Theorem 2.1 *Consider testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$. Suppose that T is a sufficient statistic for θ and the family of pdfs or pmfs $\{g(t \mid \theta) : \theta \in \Theta\}$ of T has a monotone likelihood ratio (MLR). Then for any t_0, the test that rejects H_0 if $T > t_0$ is a UMP level α test, where $\alpha = P_{\theta_0}(T > t_0)$.*

We will not go in more detail about what sufficient statistics and MLR are. What is important is that Theorem 2.1 reduces the problem of finding an optimal rejection criteria to one of finding a sufficient statistic. This is not necessarily trivial, but it is a fundementally different problem from finding good test statistics for two-sided hypotheses. While sufficient statistics are good candidates for test statistics in the two sided case, they do not decide how to balance the two sides of the hypothesis test against each other.

Therefore, we will focus on the two-sided case as it does not have the same ways of reducing the problem of finding optimal test statistics. Usually, this will require a choice of how to balance the left and right tail of a distribution, which we will explain further in the next section where we investigate an example for the Z-test.

2.2 Why non-symmetric distributions ?

In this section we will give motivation for why we are looking at particular experiments. These experiments typically observe test statistics that are non-symmetric and discretely distributed. We will show why the first is important by looking at the Z-test in detail. We will also explain the concept of unbiasedness.

2.2.1 The Z-test

Consider the normally distributed random sample $X_1, X_2, ...X_n \sim N(\mu, \sigma^2)$ with known variance σ^2. We want a hypothesis test for the unknown parameter μ where we test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$. It can then be shown that

$$Z = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu_0}{\sigma} \tag{2.6}$$

is standard normally distributed under the null hypothesis. Due to the null and alternative hypotheses it makes sense to have a rejection region that consists of a left and right tail. Put mathematically, our rejection region for a two-sided hypothesis test is $R = \{\mathbf{X} : Z \in (-\infty, a) \cup (b, \infty)\}$, where $a, b \in \mathbb{R}$ and $a < b$. Whether the semi-infinite intervals are closed or open

is not that important in this case because Z is continous and we thus have $P(a < Z < b) = P(a \leq Z \leq b)$.

If we define the quantile z_α as the value such that $P(Z > z_\alpha) = \alpha$[1], we can set $a = z_{1-(1-k)\alpha}$ and $b = z_{k\alpha}$, obtaing a hypothesis test with rejection criteria

$$R = \{\mathbf{X} : Z(\mathbf{X}) < z_{1-(1-k)\alpha} \cup Z(\mathbf{X}) > z_{k\alpha}\} \ . \tag{2.7}$$

This will be a size α test because

$$\begin{aligned} P(\mathbf{X} \in R) &= P(Z < a \cup Z > b) = P(Z < z_{1-(1-k)\alpha}) + P(Z > z_{k\alpha}) \\ &= 1 - P(Z > z_{1-(1-k)\alpha}) + P(Z > z_{k\alpha}) \\ &= 1 - (1 - (1-k)\alpha) + k\alpha = \alpha. \end{aligned}$$

Note that while $k = 1/2$ might be the most natural choice, as there is no reason to weigh deviations of small Z higher than those of high Z and vice versa, it is not the only possible choice. Even $k = 0$ and $k = 1$ are allowed, though they clearly yield one-sided hypothesis tests and are thus a bad match with the null and alternative hypotheses being proposed. We have decided to plot the power function $\beta(\mu)$ for differing values k in Figure 2.2.1. Notice that $Z \sim \mathrm{N}(\mu, 1)$ is always the case no matter the parameters n and σ^2 as it is a pivot. We see from Figure 2.2.1 that higher power for negative values of μ is connected to lower power for positive values of μ. This difference is very noticable between the red line ($k = 1/2$) where power is balanced between the two tails and the purple line ($k = 1$) which has one-sided rejection region. This trend holds for increasing values of k.

We thus have no UMP test, but the property of being able to reject at both sides is clearly desirable and we wish to formalize this property.

2.2.2 Unbiasedness

Unbiasedness is a restriction that is often imposed on a class of tests when it is impossible to find a UMP for that class. We restrict ourselves to looking at tests where the rejection is more likely to happen when the alternative hypothesis is true than when the null hypothesis is true. Remembering that $\beta(\theta)$ denotes the chance of rejection given θ we have the following definition for an *unbiased test*:

[1]Note that this probability α is describing the complement of the cumulative, not the cumulative probability $F_Z(x) = P(Z \leq x)$. This is because there are other distributions frequently used for hypothesis testing where larger values are often associated with deviation from the null hypothesize, like the χ^2 distribution.

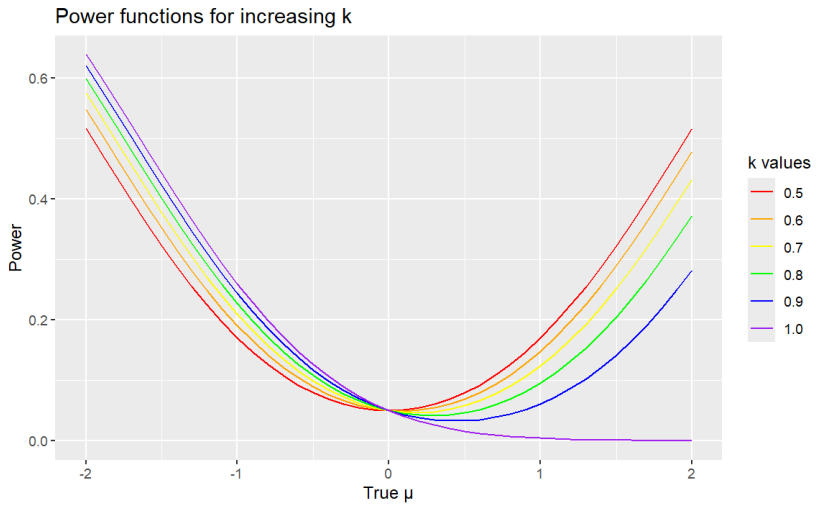


Figure 2.2.1: Power function for varying k in the Z-test where rejection happens for $z_{k\alpha} > Z$ or $Z < z_{1-(1-k)\alpha}$. $Z \sim \mathrm{N}(\mu, 1)$, $H_0 : \mu = \mu_0$, $\alpha = 0.05$ and $k = 0.5, 0.6, 0.7, 0.8, 0.9, 1$.

Definition 2.5 *A test with power function $\beta(\theta)$ and hypotheses $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_0^C$ is **unbiased** if $\beta(\theta') \geq \beta(\theta'')$ for every $\theta' \in \Theta_0^C$ and $\theta'' \in \Theta_0$ (Casella and Berger, 2002, p.387).*

The Z-test is a good example of unbiasedness being a useful concept. When we look at Figure 2.2.1 we see that there is no UMP test. We do however observe that there is only one unbiased test, namely the one shown with the red line where $k = 1/2$. In this case $\beta(\mu_0) = \alpha$, recall that $H_0 : \mu = 0$ making this the supremum of θ in the null hypothesis, is smaller than all $\beta(\mu)$ for $\mu \neq 0$. So the rejection rate for the alternative hypothesis is higher than for the null hypothesis. Thus, if we require unbiasedness, $k = 1/2$ is the only choice among tests with a rejection region as described in Eq. (2.7).

2.2.3 Symmetry

In a Z-test the test statistic has a symmetric distribution when the null hypothesis is true. We will see that this is a property that usually makes natural test-statistics for a two-sided hypothesis test coincide in how they rank the likelihood of the different outcomes of this test-statistic. This will be elaborated on in Chapter 3, where we go in depth on p-values and extremity. For now it is important to mention the null hypotheses that produce symmetric distributions under the null hypothesis when we present the hypothesis test families that we are interested in in this project.

2.2.4 Discreteness

All experiments we study in this thesis have discretely distributed test statistics $T(\mathbf{X})$. This makes it generally difficult to construct size α tests unless we use a randomized test. By a randomized test we mean the case where a realized value of the random variable X might lead to a random rejection. For example, we might reject H_0 for $X = 4$ and not reject for $X = 6$, while $X = 5$ leads to a rejection 20% of the time. As in the example, the random rejection usually happpens at a boundary point between rejection and acceptance regions. Examples of randomized tests that fullfill UMP requirments for the class of unbiased level α tests are discussed in Dunne et al. (1996). We will not consider these randomized tests satisfying as the rejection depend on more than the observed test statistic $T(\mathbf{X})$.

2.3 Experiments

In this section we will throughly define the different experiment setups we will investigate in this thesis. We will omit the rejection critera when defining the experiments in this section, as the rejection criteria will depend on which p-value is used. We will only present a test statistic $T(\mathbf{X})$ for each experiment that the p-values again will be based on.

2.3.1 Binomial experiment

Binomial experiment

We observe the random variable X which is binomially distributed with parameter θ and n number of trials[2]. That is

$$P(X = x) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \quad x = 0, 1, ..., n \ . \tag{2.8}$$

We consider the parameter n to be known and want to test θ with a two-sided test, i.e. $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. We will use the test statistic $T(X) = X$ which has PMF as in Eq. (2.8).

Note that this test is equivalent to testing the parameter of success θ for n identically, independantly distributed Bernoulli trials.

[2]Often one will use p to denote the chance of success. We have used θ to prevent confusion with the p-value, also denoted by p.

If we let $\theta_0 = 1/2$ we see that the PMF under the null hypothesis becomes

$$P(X = x \mid \theta_0) = \binom{n}{x}\theta_0^x(1-\theta_0)^{n-x} = \binom{n}{x}\left(\frac{1}{2}\right)^n , \tag{2.9}$$

and is thus symmetric about $x = n/2$. We will be most interested in the cases where this is not the case. In a similar vein, we should notice that, as a consequence of being equivalent to a sum of n iid Bernoulli trials and the Central Limit Theorem, X converges in distribution to a normal distribution for large n. Thus, we also require that n is relatively small to be interesting. In this thesis we have not looked at $n > 50$.

2.3.2 Fisher's exact test

Consider two independant binomially distributed variables X_a and X_b with respectively n_a and n_b number of trials and chance of success θ_a and θ_b. The number of trials n_a and n_b is known and we want to test if the parameters θ_a and θ_b are equal, i.e. $H_0 : \theta_a = \theta_b$ and $H_1 : \theta_a \neq \theta_b$. This is also a two-sided hypothesis test even though it looks slightly different than Eq. 2.3.

Fisher found that if we define the new random variable $Y = X_a + X_b$ and consider the PMF of X_a given Y it is hypergeometrically distributed (Agresti, 1992, p.134). The derivation is quite simple and we will reproduce one done by Dåsvand (2022). It will use the fact that under the null hypothesis Y has binomial distribution with $n_a + n_b$ trials and chance of success $\theta_a = \theta_b$. Then we get that

$$\begin{aligned} P(X_a = x_a \mid Y = y) &= \frac{P(X_a = x_a \cap Y = y)}{P(Y = y)} \\ &= \frac{\binom{n_a}{x_a}\theta_a^{x_a}(1-\theta_a)^{n_a-x_a}\binom{n_b}{y-x_a}\theta_a^{y-x_a}(1-\theta_a)^{n_b-y+x_a}}{\binom{n_a+n_b}{x_a}\theta_a^{x_a}(1-\theta_a)^{n_a+n_b-x_a}} = \frac{\binom{n_a}{x_a}\binom{n_b}{y-x_a}}{\binom{n_a+n_b}{y}} . \end{aligned}$$

We see that this distribution becomes symmetric when $n_a = n_b$ and so we will be most interested in cases where this is not the case.

Employing that $X_a \mid Y$ is hypergeometrically distributed we may obtain a one-sided test of $H_0 : \theta_a = \theta_b$ against $H_1 : \theta_a > \theta_b$. This can be done by rejecting when $x_a > x_{y,\alpha}$ where $x_{y,\alpha}$ is the smallest value that produces a level α test, i.e. such that $P(X_a > x_{y,\alpha} \mid Y = y) \leq \alpha$. When producing a two-sided test for $H_0 : \theta_a = \theta_b$ against $H_1 : \theta_a \neq \theta_b$ there is more discussion on what constitutes good rejection criteria (Agresti, 1992, section 2.1). We will go through these methods in detail in the chapter 3 about p-values. For now we will present the setup for the experiment as we have done for the binomial experiment, omitting the rejection criteria.

Two-sided Fisher's exact test

We observe the random variables X_a, X_b which are binomially distributed with success parameter θ_a, θ_b and n_a, n_b number of trials respectively. We want to test the two-sided hypothesis $H_0 : \theta_a = \theta_b$ against $H_1 : \theta_a \neq \theta_b$. We will use the test statistic $T(X_a, X_b) = X_a$ which will have hypergeometric distribution given $Y = X_a + X_b = y$.

$$P(X_a = x_a \mid Y = y) = \frac{\binom{n_a}{x_a}\binom{n_b}{y-x_a}}{\binom{n_a+n_b}{c}} \quad x_a = 0, 1, ..., y \ . \tag{2.10}$$

2.3.3 Poisson experiment

Poisson experiment

We observe the random variable X which is Poisson distributed with parameter μ. That is

$$P(X = x) = \frac{\mu^x}{x!}e^{-\mu}, \quad x = 0, 1, ... \ . \tag{2.11}$$

We want to test this parameter μ with a two-sided test $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. We will use the test statistic $T(X) = X$ which will have PMF given by (2.11).

The distribution of X will never be symmetric. It will however have convergence to a normal distribution for $\mu \to \infty$. That means that we are most interested in cases where μ_0 is small. However, we also do not want this parameter too small, as we will see that it may quickly turn into a one-sided hypothesis test when a lot of the probability mass is at or very close to 0.

Chapter 3

P-Values

3.1 General principles

One way of reporting on a finished hypothesis test is to report the outcome (rejection/non-rejection of null hypothesis) and the level α of the test. An alternative way of reporting the findings is the p-value test statistic.

Definition 3.1 *(Casella and Berger, 2002, p.397) A* ***p-value*** $p(\mathbf{X})$ *is a test statistic satisfying* $0 \leq p(\mathbf{x}) \leq 1$ *for every sample point* $\mathbf{x}$*. Small values of* $p(\mathbf{X})$ *give evidence that* H_0 *is false*[1]*. A p-value is* ***valid*** *if, for every* $\theta \in \Theta_0$ *and every* $0 \leq \alpha \leq 1$*,*

$$P_\theta(p(\mathbf{X}) \leq \alpha) \leq \alpha \ . \tag{3.1}$$

A valid p-value p can easily be used to specify a rejection critera for a level α test. This is done by rejecting the null hypothesis whenever $p(\mathbf{x}) \leq \alpha$. We see from the Definition 3.1 that this can at most lead to rejection with probability α whenever H_0 is true.

3.1.1 Test statistics and p-values

As previously stated, a p-value may be thought of as the maximal probability of observing an event that is as extreme or more extreme under the null hypothesis. This ranking of extremity can be done with another test statistic $W(\mathbf{X})$.

[1]Casella and Berger Casella and Berger (2002, p.397) states "H_1 is true" instead of "H_0 is false". To avoid the philosophical debate of whether we can give evidence for hypotheses being true we have stated the theorem in terms of rejection.

Theorem 3.1 *(Casella and Berger, 2002, p.397) Let $W(\mathbf{X})$ be a test statistic such that large values of W give evidence that $H_0 : \theta \in \Theta_0$ is false. For each sample point $\mathbf{x}$, define*

$$p(\mathbf{x}) = \sup_{\theta \in \Theta_0} P_\theta(W(\mathbf{X}) \geq W(\mathbf{x})) \ . \tag{3.2}$$

Then, $p(\mathbf{X})$ is a valid p-value.

I have found it helpful to distinguish between the test statistics $T(\mathbf{X})$ and $W(\mathbf{X})$. $T(\mathbf{X})$ is a test statistic which is a natural start of an observed statistic irregardless of the specific rejection criteria, for example the sample mean when studying the mean of a normal distribution. $W(\mathbf{X})$ on the other hand is based on $T(\mathbf{X})$ again and is a measure of extremity with respect to the null hypothesis. Therefore, I will sometimes reference an "extremity statistic" and it will then refer to $W(\mathbf{X})$. Note that Theorem 3.1 could have been based on a general test statistic $T(\mathbf{X})$.

Note that we might as well have used a statistic that rejects for small values instead of large, even though it does not immediately fulfill the criteria of Theorem 3.1. Consider that we have such a statistic $W'(\mathbf{X})$ where small W' imply rejection. Then simply setting $W(\mathbf{X}) = -W'(\mathbf{X})$ we get a new $W(\mathbf{X})$ which fullfills Theorem 3.1 because now $-W'(\mathbf{X})$ rejects for large values.

We will also employ a slight variation of this theorem.

Theorem 3.2 *(Casella and Berger, 2002, EQ 8.3.10, p.399) Suppose $S(\mathbf{X})$ is a sufficient statistic for $\theta \in \Theta_0$ and let $W(\mathbf{X})$ be a test statistic such that large values of W give evidence that $H_0 : \theta \in \Theta_0$ is false. For each sample point $\mathbf{x}$, define*

$$p(\mathbf{x}) = \sup_{\theta \in \Theta_0} P_\theta(W(\mathbf{X}) \geq W(\mathbf{x}) \mid S = S(\mathbf{x})) \ . \tag{3.3}$$

Then, $p(\mathbf{X})$ is a valid p-value.

Again, we will not go in detail on sufficient statistics. We mention this theorem because we will need it to construct p-values for Fisher's exact test. Here the statistic $S(\mathbf{X})$ is $S(\mathbf{X}) = Y = X_a + X_b$, where Y,X_a and X_b is as in section 2.3.2. It can be shown that $S(\mathbf{X})$ is sufficient under the null hypothesis. Note that it is important that the statistic $S(\mathbf{X})$ is only sufficient under the null as it will have low power outside this region otherwise (Casella and Berger, 2002, p.399).

Notice that in our case we have two-sided hypotheses on the form $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. This means that $\Theta_0 = \{\theta_0\}$ and that $\sup P_\theta(\cdot) =$

$P_{\theta_0}(\cdot)$, i.e. we need only find the probability in a specific case and not the supremum. Fisher's exact test will similarly not need the supremum due to being conditioned on a sufficient statistic.

3.2 Construction of two-sided p-values

We Will now construct the two-sided p-values. The p-values we will construct are (i) double tail, (ii) point p-value, (iii) tail p-value and (iv) distance p-value. We will make some comments on properties of the p-values when it is natural. We will also explicitly show that the p-values coincide for symmetric distributions.

3.2.1 Double tail method

Note that using the Theorem 3.1 or 3.2 is in no way required for a valid p-value. We will start with such a p-value. This p-value uses the double of the smallest one-sided p-value. Because we are dealing with discrete distributions we might have that this one-sided p-value is larger than 1/2, which would make the p-value larger than 1 which is not allowed.

We thus arrive at the following definition for what we will call the **double tail p-value**:

Definition 3.2 *Inspired by Dåsvand (2022, Section 3.2). Let P_l denote the p-value for the left one-sided test $H_0 : \theta = \theta_0$ versus $H_1 : \theta < \theta_0$ (or $H_0 : \theta_a = \theta_b$ versus $H_1 : \theta_a < \theta_b$ for Fisher's exact test) and let P_r denote the p-value for the right one-sided test $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$ (or $H_0 : \theta_a = \theta_b$ versus $H_1 : \theta_a > \theta_b$ for Fisher's exact test). Then the **double tail p-value** is defined as*

$$p_{DT}(\mathbf{x}) = 2 \min\left(P_l, P_r, 1/2\right) \ . \tag{3.4}$$

Proof of validity is given in Appendix B.

We see that we need to calculate the one-sided p-values P_l and P_r. Most often this will be done with a test statistic $T(\mathbf{X})$ for the situation. Let $T(\mathbf{X})$ be a test statistic such that small values indicate rejection of the left one-sided test and large values indicate rejection of the right one-sided test. Then,

$$P_l(\mathbf{x}) = P_{\theta_0}(T(\mathbf{X}) \leq T(\mathbf{x})) \ , \tag{3.5}$$
$$P_r(\mathbf{x}) = P_{\theta_0}(T(\mathbf{X}) \geq T(\mathbf{x})) \ . \tag{3.6}$$

For the binomial and Poisson test we have $T(X) = X$. We can then simply use the CDF of binomial and Poisson distributions respectively to calculate the one-sided p-values.

For Fischer's exact test we have a test statistic $T(X_a, X_b) = X_a$ where the rejection criteria is similar as in Eq. (3.5) and Eq. (3.6), but it is conditioned on the statistic $S = X_a + X_b$ which is sufficient for θ_a under the null hypothesis $H_0 : \theta_a = \theta_b$. Therefore, the one-sided p-values becomes slightly different

$$P_l(\mathbf{x}) = P_{\theta_0}(T(\mathbf{X}) \leq T(\mathbf{x}) \mid S = S(\mathbf{x})) \ , \tag{3.7}$$
$$P_r(\mathbf{x}) = P_{\theta_0}(T(\mathbf{X}) \geq T(\mathbf{x}) \mid S = S(\mathbf{x})) \ . \tag{3.8}$$

Since the double tail p-value does not use either Theorem 3.1 or Theorem 3.2 we will briefly discuss where it comes from. In his influential work *Statistical methods for Research Workers*, Fisher mentions that the p-value for a T-test should be divided by 2 if it is one-sided (Fisher, 1925, p.120). This is equivalent to saying we should multiply the one-sided p-value by 2 to obtain the two-sided p-value. This is of course true for the T-test because it is based on the Student t-distribution which is symmetric, and it will still produce a valid p-value for non-symmetric cases as we show in Appendix B. However, the added machinery of taking the smallest one-sided p-value and also inlcuding the case of both being larger than 1/2 should make us doubt that this is the best way to define it in the non-symmetric case.

3.2.2 Point p-value

Now we will use an extremity test statistic $W_P(\mathbf{X})$ that ranks extremity for defining a p-value. In this case we will use the point probability, so for a discrete sample $\mathbf{X}$ we have that,

$$W_P(\mathbf{x}) = P_{\theta_0}(T(\mathbf{X}) = T(\mathbf{x})) \ . \tag{3.9}$$

The subscript P stands for Point probability. We have choosen to look at the point probability of $T(\mathbf{X}) = T(\mathbf{x})$ instead of the sample point probability $\mathbf{X} = \mathbf{x}$. This is to make it more similar to the tail and distance p-values later.

Then we get the following definition for **point p-value** $p_P(\mathbf{X})$.

Definition 3.3 *Consider a sample* $\mathbf{X}$ *which has a distribution dependant on* θ *and no other unknown parameters. Let the statistic* $W_P(\mathbf{X})$ *denote the point probability under the null hypothesis* $H_0 : \theta = \theta_0$*, i.e.* $P_{\theta_0}(T(\mathbf{X}) = T(\mathbf{x}))$*. Then define the* ***point p-value*** $p_P(\mathbf{X})$ *as*

$$p_P(\mathbf{x}) = P_{\theta_0}(W_P(\mathbf{X}) \leq W_P(\mathbf{x})) \ , \tag{3.10}$$

which we know is valid from Theorem 3.2.

Also, we define the ***conditional point p-value*** *as*

$$p_P(\mathbf{x}) = P_{\theta_0}(W_P(\mathbf{X}) \leq W_P(\mathbf{x}) \mid S = S(\mathbf{x})) \ , \tag{3.11}$$

where $S(\mathbf{X})$ is some statistic which is sufficient for the model when the null hypothesis is true. We know this p-value is valid from Theorem 3.3.

The conditional and non-conditional point p-values are very similar. The author does not promise to always explicitly distinguish between these two as it will often be evident from context.

Using the PDF, instead of point probability, would have been a natural extension to continous distributions, but we have not focused on it in this thesis.

For the binomial test and the Poisson test we can use Eq. (3.10) with $T(X) = X$. For Fisher's exact test we can use Eq. (3.11) with $T(\mathbf{X}) = X_a$ and $S(\mathbf{X}) = X_a + X_b$, where X_a, X_b is defined as in section 2.3.2. This yields the following sums.

Two-sided Point p-values

All probability functions are under the null hypothesis. $I(A)$ denotes the indicator function of the event A.
Binomial test :

$$p(x) = \sum_{i=0}^{n} I[P(X = i) \leq P(X = x)]P(X = i)$$

Fisher exact :

$$\begin{aligned} &Y = X_a + X_b \ , \quad y = x_a + x_b \\ &W_T \mid S(x_a, y) := P(X_a = x_a \mid Y = y) \\ &p(x_a, x_b) = \sum_{i=0}^{y} I[W_T \mid S(i, y) \leq W_T \mid S(x_a, y)]P(X_a = i \mid Y = y) \end{aligned}$$

Poisson test:

$$\begin{aligned} p(x) &= \sum_{i=0}^{\infty} I(P(X = i) \leq P(X = x))P(X = i) \\ &= 1 - \sum_{i=0}^{\infty} I(P(X = i) > P(X = x))P(X = i) \end{aligned}$$

3.2.3 Tail p-value

Now we will use the minimal tail probability as an extremity statistic $W_T(\mathbf{X})$ for ranking the possible outcomes. Calculating the tail probability can be a bit tricky for higher dimensional sample spaces. Say we have $\mathbf{X} = (X_a, X_b)$. Then it becomes hard to say which is biggest between the outcomes $\mathbf{x} = (0, 2)$ and $\mathbf{x} = (1, 1)$.

Luckily, we were careful to define one dimensional test statistics when we defined the tests in Section 2.3. Thus,

$$W_T(\mathbf{x}) = \min(P_{\theta_0}(T(\mathbf{X}) \geq T(\mathbf{x})), P_{\theta_0}(T(\mathbf{X}) \leq T(\mathbf{x}))) \ , \tag{3.12}$$

is a well-defined statistic for our tests. Small values of this W_T give grounds for rejecting the null hypothesis. Thus, we have the following definition for the **tail p-value** :

Definition 3.4 *Consider a sample* $\mathbf{X}$ *which has a distribution dependant on* θ *and no other unknown parameters. Let* $T(\mathbf{X})$ *be a one dimensional test statistic. Then let the statistic* $W_T(\mathbf{x})$ *denote the smallest tail probability* $\min(P(T(\mathbf{X}) \geq T(\mathbf{x})), P(T(\mathbf{X}) \leq T(\mathbf{x})))$. *Then define the* ***tail p-value*** $p_T(\mathbf{X})$ *as*

$$p_T(\mathbf{x}) = P_{\theta_0}(W_T(\mathbf{X}) \leq W_T(\mathbf{x})) \ , \tag{3.13}$$

which we know is valid from Theorem 3.1.

Also, we define the ***conditional tail p-value*** *as*

$$p_T(\mathbf{x}) = P_{\theta_0}(W_T(\mathbf{X}) \leq W_T(\mathbf{x}) \mid S = S(\mathbf{x})) \ , \tag{3.14}$$

where $S(\mathbf{X})$ *is some statistic which is sufficient for the model when the null hypothesis is true. We know this p-value is valid from Theorem 3.2.*

The conditional and non-conditional ways of defining the tail probability is very similar, just as with the point p-value, and the author again does not promise to explicitly distinguish between them as it will often be evident from context.

Similarly to the point p-value, we will use Eq. (3.13) for the binomial and Poisson tests and Eq. (3.14) for Fisher's exact test. For the Binomial and Poisson tests we simply have $T(X) = X$. For Fisher's exact test we will use $T(\mathbf{X}) = X_a$. Now we can write them up as sums.

Two-sided tail p-values

All probability functions are under the null hypothesis. $I(A)$ denotes the indicator function of the event A.
Binomial test :

$$W_T(x) = \min(P(X \leq x), P(X \geq x))$$
$$p(x) = \sum_{i=0}^{n} I(W_T(X) \leq W_T(x))P(X = i)$$

Fisher exact :

$$Y = X_a + X_b, \quad y = x_a + x_b$$
$$W_T \mid S(x_a, y) := \min(P(X_a \leq x_a \mid Y = y), P(X_a \geq x_a \mid Y = y))$$
$$p(x_a, y) = \sum_{i=0}^{y} I(W_T \mid S(i, y) \leq W_T \mid S(x_a, y))P(X_a = i \mid Y = y)$$

Poisson test:

$$W_T(x) = \min(P(X \leq x), P(X \geq x))$$
$$p(x) = \sum_{i=0}^{\infty} I(W_T(X) \leq W(x))P(X = i)$$
$$= 1 - \sum_{i=0}^{\infty} I(W_T(X) < W(x))P(X = i)$$

We observe that the tail p-value is always more powerful than the double tail p-value, i.e. the p-value will always be lower for the tail p-value while still being at the signifcance level α. Let us prove this for a one dimensional sample X.

Let $W_T(X), p_T(X)$ be defined as in Definition 3.4 with one dimensional sample X such that $T(X) = X$. Let $p_{DT}(X)$ be defined as in Definition 3.2 with $P_l(X) = P(X \leq x)$ and $P_r(X) = P(X \geq x)$.

Because $W_T(X)$ has a unimodal distribution we know that the outcomes j such that $W_T(j) \leq W_T(x)$ are those above or below the critical values a and b repectively. Furthermore one of these critical values must be equal to x. This is illustrated in Figure 3.2.1.

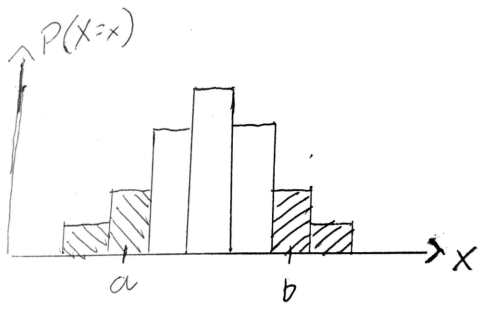


Figure 3.2.1: Critical values a and b for the tail probability statistic.

This means that

$$\begin{aligned} p_T(x) &= P_{\theta_0}((X \leq a) \cup (X \geq b)) \\ &= P_{\theta_0}(X \leq a) + P_{\theta_0}(X \geq b) \ . \end{aligned}$$

Lets take the case where the left tail is smallest, i.e. $W_T(x) = P(X \leq x)$ and $a = x$. We then have that $P_{\theta_0}(X \leq a) = P_{\theta_0}(X \leq x)$. Furthermore, since b is the smallest value that has $W_T(b) = P_{\theta_0}(X \geq b)$ which is smaller than or equal $W_T(x)$, we know that $P_{\theta_0}(X \geq b) \leq P_{\theta_0}(X \leq x)$. This means that we can continue on the derivation;

$$\begin{aligned} p_T(x) &= P_{\theta_0}(X \leq a) + P_{\theta_0}(X \geq b) \\ &= P_{\theta_0}(X \leq x) + P_{\theta_0}(X \geq b) \\ &\leq P_{\theta_0}(X \leq x) + P_{\theta_0}(X \leq x) \\ &= 2P_l(x) = p_{DT}(x) \ . \end{aligned}$$

If we choose that the right tail $P_{\theta_0}(X \geq x)$ was largest we would arrive at $p_T(x) \leq 2P_{\theta_0}(X \geq x) = p_{DT}(x)$ in asimilar fashion.

Notice that for a continous distribution we will have equality between these p-values. To illustrate this consider the case where $W_T(x) = P_{\theta_0}(X \leq x)$ again as before. We then have that $b = \min j : P_{\theta_0}(X \geq j) \leq P_{\theta_0}(X \leq x)$. Assume we have a strict inequality $P_{\theta_0}(X \geq b) < P_{\theta_0}(X \leq x)$. Then, we must also have a value $b - \epsilon$ that is just a tiny bit smaller, but still satisfyies $P_{\theta_0}(X \geq b - \epsilon) < P_{\theta_0}(X \leq x)$ due to continuity. This is of course a contradiction to b being a critical value and thus $P_{\theta_0}(X \geq b) = P_{\theta_0}(X \leq x)$, which again means that $p_T(x) = p_{DT}(x)$.

So in conclusion, the tail p-value is more powerful than the double tail p-value. If the distribution is continous they become equal, but in the case of discrete distributions they will in general be unequal.

3.2.4 Distance p-value

Up until now we have used probabilities under the null hypothesis to gain extremity statistics which then defines p-values. We will now use something a bit different; the distance from the mean.

Just as with the tail test statistic, higher dimensions might cause problems. Therefore, a one dimensional test statistic $T(\mathbf{X})$ will be very important also in the definition for the distance statistic $W_D(\mathbf{X})$,

$$W_D(\mathbf{x}) = |\,\mathrm{E}_\theta[T(\mathbf{X})] - T(\mathbf{X})|\ . \tag{3.15}$$

Note that there are many other test statistics which it would be very natural to call distance statistics. It could for example be cases where the dimensionality reduction is done in a different manner or when measuring from another central tendency other than the mean, like the mode or median. In this thesis we have restricted ourselves to the mean because it was among those proposed by Agresti (1992) for Fisher's exact test.

We define the **distance p-value** $p_D(\mathbf{X})$ (and we want to emphasize that it is very specific to this thesis and is probably not a good name outside this context) as

Definition 3.5 *Let* $\mathbf{X}$ *be a sample with distribution dependant on* θ *and no other unknown parameters. Let* $T(\mathbf{X})$ *be a one dimensional statistic. Then let the statistic* $W_D(\mathbf{X})$ *denote the sum of distances from the means* $W_D(\mathbf{X}) = |\,\mathrm{E}_\theta[T(\mathbf{X})] - T(\mathbf{X})|$*. Then define the* ***distance p-value*** $p_D(\mathbf{X})$ *as*

$$p_D(\mathbf{x}) = P_\theta(W_D(\mathbf{X}) \leq W_D(\mathbf{x}))\ , \tag{3.16}$$

which we know is valid from Theorem 3.1.

Also, we define the ***conditional distance p-value*** *as*

$$p_D(\mathbf{x}) = P_\theta(W_D(\mathbf{X}) \leq W_D(\mathbf{x}) \mid S = S(\mathbf{x}))\ , \tag{3.17}$$

where $S(\mathbf{X})$ *is some statistic which is sufficient for the model when the null hypothesis is true. We know this p-value is valid from Theorem 3.2.*

As with the point and tail p-values, the difference between conditional and non-conditional p-values is not one we will spend time on.

In a similar fashion to before we will use Eq. (3.16) with $T(X) = X$ for the binomial and Poisson experiments. And for Fischer's exact test we will use Eq. (3.17) with $T(X_a, X_b) = X_a$ and $S(X_a, X_b) = X_a + X_b$. We plug in the expected values under the null hypothesis for the different distributions and arrive at the following results.

Two-sided distance p-values

All probability functions are under the null hypothesis. $I(A)$ denotes the indicator function of the event A.
Binomial test :

$$W_D(x) = |\,\mathrm{E}_{\theta_0}[X] - x| = |n\theta_0 - x|$$
$$p(x) = \sum_{i=0}^{n} I(W_D(X) \geq W_D(x))P(X = i)$$

Fisher exact :

$$Y = X_a + X_b, \quad y = x_a + x_b$$
$$W_D \mid S(x_a, y) := |\,\mathrm{E}_{\Theta_0}[X_a \mid Y = y] - x_a|$$
$$= |\frac{n_a y}{n_a + n_b} - x_a|$$
$$p(x_a, y) = \sum_{i=0}^{y} I(W_D \mid S(i, y) \geq W_D \mid S(x_a, y))P(X_a = i \mid Y = y)$$

Poisson test:

$$W_D(x) = |\,\mathrm{E}_{\mu_0}[X] - x| = |\mu_0 - x|$$
$$p(x) = \sum_{i=0}^{\infty} I(W_D(X) \geq W(x))P(X = i)$$

3.2.5 Why are symmetric distributions uninteresting ?

Now it becomes more fruitful to discuss why symmetric distributions are uninteresting as we mentioned in section 2.2.3.

For a discretely distributed random variable X with symmetry about a we have that

$$P(X = a - b) = P(X = a + b)\ . \tag{3.18}$$

Clearly the realizations $a-b$ and $a+b$ have the same point proability statistic W_P. Since the mean of a symmetric distribution will be the center of the symmetry, we have $E[X] = a$ and the distance from the mean is the same for realizations $a - b$ and $a + b$. Thus, $p_P(X) = p_D(X)$.

Furthermore, since $P(X = a - b - i) = P(X = a + b + i)$ for all $i \in \mathbb{N}$ we have that

$$\sum_{i=0}^{\infty} P(X = a - b - i) = \sum_{i=0}^{\infty} P(X = a + b + i) \ . \tag{3.19}$$

We recognize that both sides of the equations are tail probabilities and in particular they are the smallest tail probability for the points $a-b$ and $a+b$. Thus, the tail proability statistic W_T is also equal for these two points and we have $p_P(X) = p_D(X) = p_T(X)$.

As we have seen now, all the statistics become equal for points equally far from the mean. This means that the statistics will rank extremity of outcomes the same and the p-values will coincide. We also see that for the double tail p-value will also coincide because

$$p_T(a + b) = \sum_{i=0}^{\infty} P(X = a + b + i) + \sum_{i=0}^{\infty} P(X = a - b - i) \tag{3.20}$$

$$= 2 \sum_{i=0}^{\infty} P(X = a + b + i) = p_{DT}(a + b) \ . \tag{3.21}$$

Thus, we conclude that $p_{DT}(X) = p_P(X) = p_T(X) = p_D(X)$ for a symmetrically, discretely distributed random variable.

While you can certainly find statistics that rank extremity differently in the symmetric case, it is hard for that same statistic to be natural for the null hypothesis $H_0 : \theta = \theta_0$. We have not looked at any other measures of extremity in this thesis.

3.2.6 Pseudocode

Now we will go through some pseudocode for how the p-values are calculated in this project, as well as how power was calculated.

The double tail p-value $p_{DT}(x)$ is very straightforward. We simply calculate the p-values for the left one-sided test $P_l(x)$ and right one-sided test $P_r(x)$ which can be calculated by CDFs.

3.2.6.1 P-values

Let the random variable X be from a known distribution under the null hypothesis $H_0 : \theta = \theta_0$ and let x be the observed outcome of X. We can then use Algorithm 1 to calculate the double tailed p-value. The input is the parameter θ_0 and the distribution of X. Here we have used that $P_{\theta_0}(X \geq$

Algorithm 1 Double tail p-value

1: `pl` $\leftarrow P_{\theta_0}(X \leq x)$
2: `pr` $\leftarrow 1 -$ `pl` $+ P_{\theta_0}(X = x)$
3: `p_val` $\leftarrow 2\min($`pl`$,$ `pr`$, 1/2)$
4: **return** `p_val`

$x) = 1 - P_{\theta_0}(X \leq x) + P_{\theta_0}(X = x)$ for discretely distributed X. For Fisher's exact test we substitute X for X_a, x for x_a and make every probability conditional on $Y = x_a + x_b$.

Algorithm 2 Extremity statistic p-value

1: `Wx` $\leftarrow W(x)$
2: `p_val` $\leftarrow 0$
3: **for** `i` **in** $0 : n$ **do**
4: `Wi` $\leftarrow W($`i`$)$
5: **if** `Wi` $\leq$ `Wx` **then**
6: `p_val` $\leftarrow$ `p_val` $+ P_{\theta_0}(X =$ `i`$)$
7: **return** `p_val`

Let the random variable X be from a known distribution under the null hypothesis $H_0 : \theta = \theta_0$ and let x be the observed outcome of X. We can then use Algorithm 2 to calculate the p-value when it is based on an extremity statistic $W(X)$. THe input variables are the parameter θ_0, the distribution of X and the extremity statistic $W(X)$. Different p-values will have different extremity statistics. We must also change whether we add the probabilities $P_{\theta_0}(X = i)$ for which i has smaller or larger $W(i)$ than $W(x)$. For the point and tail probability statistics we want to add cases where the $W(i) \leq W(x)$ while we want $W(i) \geq W(x)$ for the distance statistic.

When we have an infinite sample space, as is the case for the Poisson test, we must restructure the algorithm so we do not get an infinite for-loop. We will do this by subtracting the cases where $W(i) > W(x)$ instead. We can do this within a while-loop with a stop condition that activitates whenever the condition for subtracting a point has failed after it has begun succeding. This will work because all the statistics have a unimodal distribution. This is shown in detail in Algorithm 3.

3.2.6.2 Power

The input variables for calculating power is the p-value function `p_func`, the true parameter θ and the signifcance level α. In addition, we must input

Algorithm 3 Extremity statistic p-value, Infinte sample space

```
1: Wx ← W(x)
2: p_val ← 1
3: i ← 0
4: loopExited ← FALSE
5: loopEntered ← FALSE
6: while NOT loopExited do
7:     Wi ← W(i)
8:     if Wi > Wx then
9:         p_val ← p_val − P_θ0(X = i)
10:        loopEntered ← TRUE
11:    else
12:        if loopEntered then
13:            loopExited ← FALSE
14:    i ← i + 1
15: return p_val
```

the null hypothesis parameter θ_0 and the distribution of X to calculate the p-value. We also need to know what experiment we are using as this will decide which algorithm we are using.

We start by looping through the sample space. For example, we would let the outcome i iterate from 0 to n for the binomial experiment. For each outcome i we would then calculate $\texttt{p\_func}(i, \theta_0)$, where the p-value of course is a function of the null hypothesis. We then check if we reject the null hypothesis or not, that is if $\texttt{p\_func}(i, \theta_0) < \alpha$ or not. If the outcome i gives rejection we add the probability of i, this time given the true parameter θ, and add it to the power. When we have iterated through the entire sample space we have our power given θ, θ_0, the p-value and α. Due to the difference in sample space we need slightly different variations of calculating power. The algorithm is described in pseudocode in Algorithm 4 for the binomial experiment. The pseudocode for Fisher's exact test and Poisson experiments can be found in Algorithm 5 and Algorithm 6 respectively which are in Appendix C.

Calculating for a single fixed θ is not that interesting so we must discretize our parameter space and loop over different values of theta, still keeping θ_0, the p-value and α fixed. We have also calculated power for variable θ and θ_0, obtaining a two-dimensional grid of power instead of a one-dimensional slice. We have kept $\alpha = 0.05$ throughout this thesis. The most useful visualization we have found is plotting the difference between two p-values in a heatmap

with the color representing the difference in power. For interesting cases we found it useful to plot a slice of the power functions with fixed θ_0.

Algorithm 4 Calculate Power, Binomial test

```
pwr ← 0
for i in 0 : n do
    pi ← p_func(i, θ_0)
    if pi ≤ α then
        pwr ← pwr + P_θ(X = i)
return pwr
```

Chapter 4

Discussion

In this chapter we will investigate and compare the properties of the different p-values with each other in the case of a binomial experiment, Fisher's exact test and a Poisson experiment. We will do this by comparing plots of realized p-values $p(x)$ against x and plots of the power function $\beta(\theta)$ against θ. This already gives us 4 p-values times 3 tests, making it 12 cases to investigate. In each of these cases it is possible to make many different choices of θ_0 in the null hypothesis $H_0 : \theta = \theta_0$. When we want a power function $\beta(\theta)$ we must also decide the signifcance level α (although this has been set to $\alpha = 0.05$ as that is the norm, throughout this thesis). In addition to these choices before making plots we also have the number of trials, n for the binomial test and both n_a and n_b for Fisher's exact test.

In conclusion, we cannot present all possible outcomes and we simply must make a selection when we are investigating the properties of the p-values. We have therefore choosen examples that most effectively illustrates trends in how the different p-values behave. The rest of the plots can be found in Appendix D.

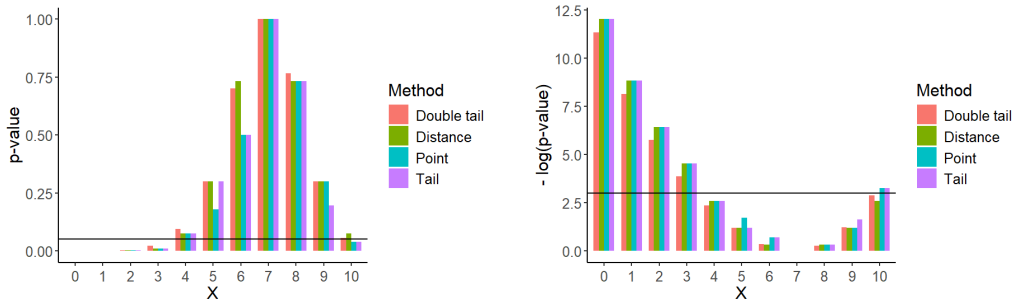
We will first compare on the distance and the double tail p-values to the others, as those were frequently outperformed. Then we will more closely investigate the point and tail p-value as they are most promising.

4.1 Double tail and distance p-values

We will see that in general the double tail and the distance p-value becomes larger than the others. We will look at an realizaed p-values for each experiment and then finally look at power for Fisher's exact test.

4.1.1 Realizations of p-values

First we have used the binomial experiment setup in section 2.3.1 with $n = 10$ and $\theta_0 = 0.7$. The resulting p-values $p(x)$, all methods, are shown in Figure 4.1.1. In additon, we also plot $-\log(p(x))$ to more easily see differences in the p-value where it is close to 0.



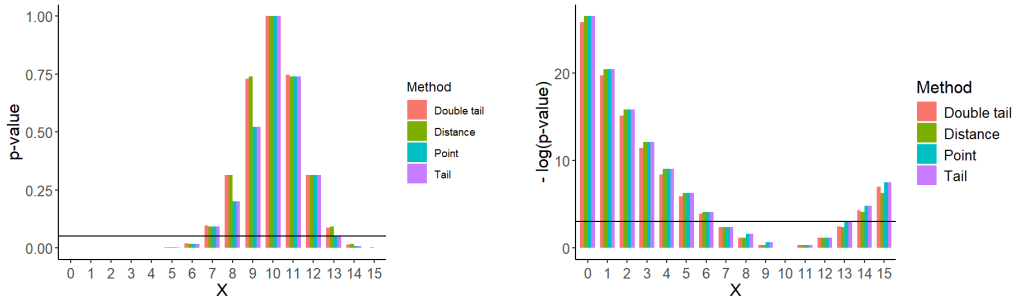
(a) P-value **(b)** - log(P-value)

Figure 4.1.1: P-values as functions of realized values **X** for a binomial distribution with $n = 10, \theta = 0.7$. All methods. The solid line represents the signifcance level $\alpha = 0.05$.

For $X > 4$ we observe that the double tail and distance p-value are always larger than the others, with one exception at $X = 9$ where the point p-value is slightly larger than the double tail p-value, although this is only really noticable in the logarithmic plot. This is an exception to the rule. usually, the double tail and distance p-value is far higher than the point and tail p-value. Which of the double tail and distance p-value is higher is quite fluctuating in this region.

For $X \leq 4$ we see that the double tail p-value is always larger than the others. This makes sense as all the other p-values are based on probability of a statistic being more extreme than the one observed, which at some point will just simplify to $P(X \leq x)$. In this case, when x is a sufficently small outcome of X such that the point, tail and distance statistics all become more extreme for points that are even smaller. Meanwhile the double tail has to be double this probability as $p_{DT}(x) = 2\min(P_l, P_r, 1/2) = 2P_l(x)$. We will see this pattern in future examples and even more can be found in Appendix D.

We now look at the experimental setup of Fisher's exact test and define the variables as in section 2.3.2. You can find the resulting p-values of $n_a = 30$, $n_b = 15$ and $y = x_a + x_b = 15$ in Figure 4.1.2. Again, we have also plotted $-\log(p(x))$ so we can compare small values.



(a) P-value **(b)** - log P-value

Figure 4.1.2: P-values as functions of realized values **X** for Fishers exact test with $n = 30/15$, $c = 15$. All methods. The solid line represents the signifcance level $\alpha = 0.05$.

We see that the double tail and distance p-values are always larger than the point and tail p-values in this realization. This difference is quite large in some cases. More examples of this can be found in Appendix D.

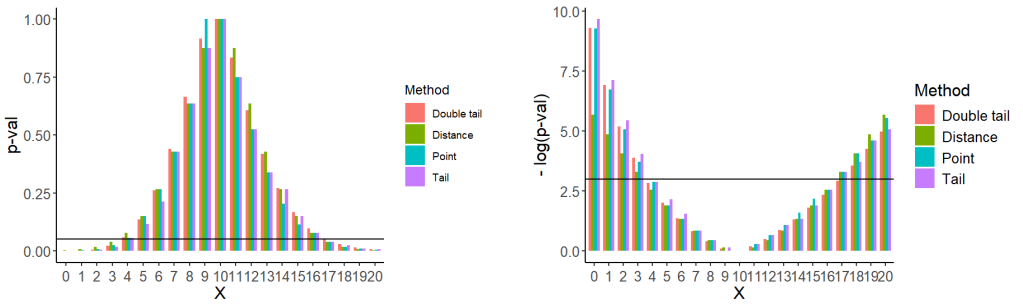
For $X > 6$ we see that in general the distance p-value seems to be the greatest. There are some cases, $X = 7$ and $X = 11$, where the double tail is larger. For $X \leq 6$ we see again the same behaviour as for the binomial test shown in Figure 4.1.1. That is the double tail p-value becomes twice as large as the other p-values while the rest fall on the same value.

We will now investigate realizations for the Poisson experiment with variables defined as in section 2.3.3. Resulting p-values upto $x = 20$ when setting $\mu_0 = 10$ can be found in Figure 4.1.3, along with a logarithmic plot of the same p-values.

The point p-value at $X = 9$ looks strange as it should only be one instance where $p(x) = 1$ when you are using a test statistic. The reason is a tie in the extremity statistic which causes to outcomes to produce the same p-value. This topic will be treated in more detail later.

For $X > 10$ we see that the double tail p-value is always larger than the point and tail p-values. The distance p-value is often larger than the others but not always, for example for $X \geq 17$. However, this is likely connected with distance being a lot larger at $X \leq 4$, the left tail.

For $X < 10$ we see that the double tail is always larger than the tail method, as expected. For the point p-value we see that there are multiple instances of the double tail p-value being smaller than it at $X = 6$, $X = 5$, $X = 3$ and $X = 2$. However, we have several instances of the point p-value being smaller at the right region $X > 10$ so this is in no way indicative of the double tail p-value outperforming the point p-value. It is nonetheless making



(a) P-value **(b)** - log P-value

Figure 4.1.3: P-values as functions of realized values $\mathbf{X}$ for a Poisson distribution with $\mu_0 = 10$. All p-values. The solid line is the signifcance level α.

it diffcult to say that the double tail p-value was outperformed by the point p-value based solely on realizations. The distance p-value is a similar story, as the small p-values it produces at $X \geq 17$ is offset by the large p-values it produces at $X \leq 4$.

4.1.2 Power comparison

We will now switch our attention from realizations of p-values to power. We will only look at Fisher's exact test to save time. As before all examples can be seen in Appendix D.

We start with explaing the plots. Each plot is a comparison of two p-values. The color shows the difference in power between the two p-values. Along the x-axis we have the parameter value θ_a, along the y-axis we have the parameter value θ_b. This means that the null hypothesis is along the diagonal where $\theta_a = \theta_b$.

In Figure 4.1.4 we have a heatmap of the power difference between the double tail and point p-values. The experimental setup is Fishers exact test with $n_a = 40$ and $n_b = 50$. θ_a and θ_b are both discretized into 21 equividistant points from 0 to 1. No colors are indicating that the two methods have the same power. Red areas indicate that the first method (in this case the double tail p-value) has higher power, while green areas indicate that the second method (in this case the point p-value) has higher power. The grey blocks indicate that the power difference is out of the bounds $[-0.1, 0.1]$.

To compare all methods we put it in a matrix like in Figure 4.1.5. Green values indicate that the p-value indicated in the row has more power while red values indicate that the column p-value has more power.

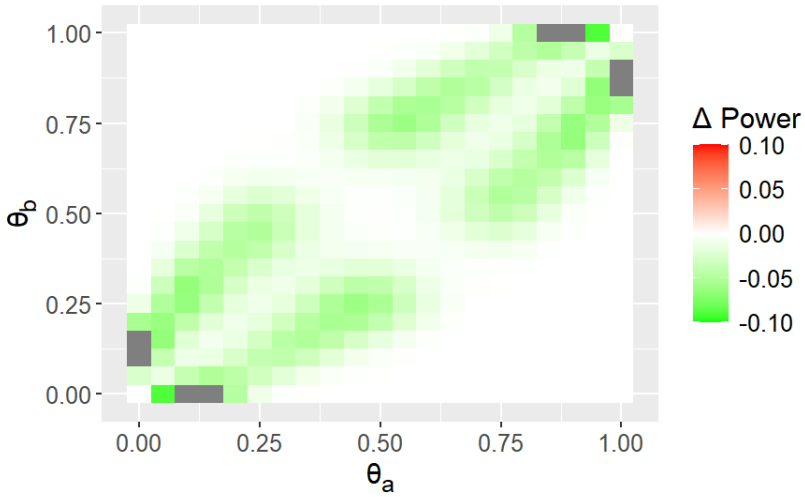


Figure 4.1.4: Power comparison between Double Tail and Point methods for varying θ_a and θ_b in a Fisher's exact test.

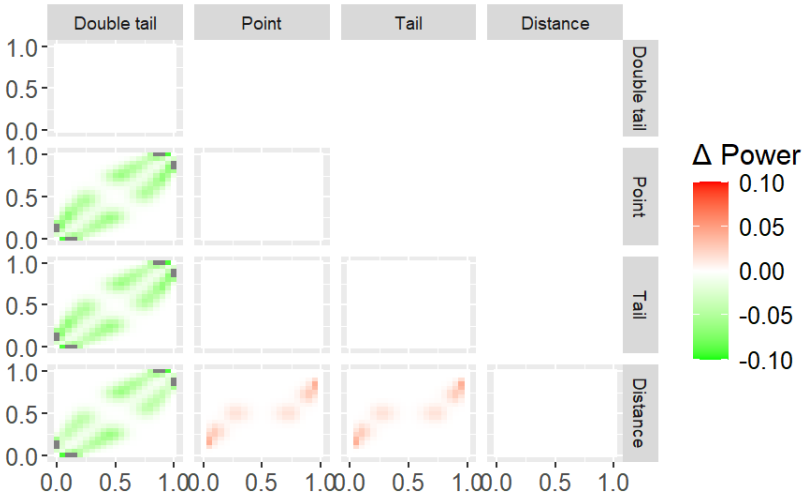


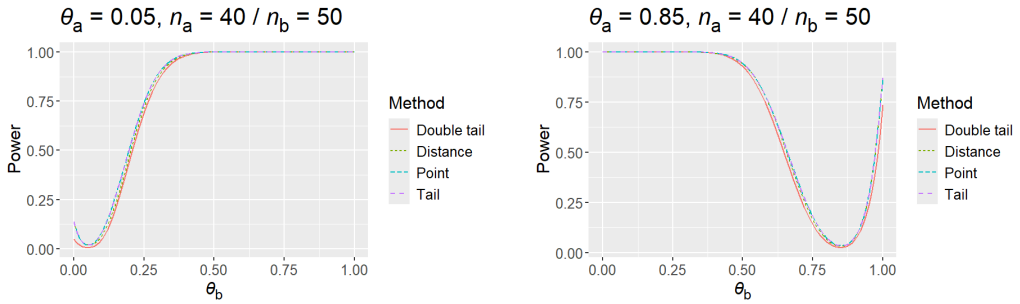
Figure 4.1.5: higher contrast power comparison between all methods for varying θ_a and θ_b in a Fisher's exact test. Black areas indicate that the difference is out of bounds

We see that in general there is a lot of green in the double tail column, indicating it has worse power than the other methods, including the distance p-value. The null hypothesis $H_0 : \theta_a = \theta_b$ lays along the diagonal. The worsening power of the double tail seemingly forms a sort of butterfly pattern along this diagonal. This means that for smaller perturbations in the null

hypothesis the power is a lot weaker than the other methods. For large deviations, the power difference between the p-values is not that noticable.

For the distance row we see a lot of red (except for the one in the double tail column). This means that the distance p-value has lower power compared to the other p-values. Again the areas where the difference in power is noticable seems to be close to the diagonal and large deviations from the null hypothesis doesn't have much difference in power.

To get a more traditional look at the power function we will also look at some slices of the 2D power matrix. We will look at slices that are interesting for both the double tail and the distance p-value. These will be $\theta_a = 0.05$ and $\theta_a = 0.85$ and they are shown as functions of θ_b in Figure 4.1.6.



(a) $\theta_a = 0.05$ **(b)** $\theta_a = 0.85$

Figure 4.1.6: Power functions for some interesting values of θ_a

The double tail has lower power than all the other p-values for all values of θ_b. For this particular value of θ_a it is simply uniformly least powerful among the plotted tests. The distance p-value does not fare much better. It is more powerful than the double tail, but it is less powerful than the point and tail p-values.

4.1.3 Conclusion

To summarize, we have seen some specific examples of how the double tail and distance p-values produce higher p-values than the point and tail p-values. This held for the binomial experiment and Fisher's exact test while a Poisson experiment was more ambigous. In terms of power the double tail and distance p-values are far weaker than the point and tail p-values for Fisher's exact test. Further examples can be seen in Appendix D. We also want to restate that the double tail p-value will necessairly be less powerful than the tail p-value. In general the point and tail p-values seem more promising than the double tail and distance p-values.

4.2 Point vs tail p-value

In the previous section we spent some time on showing how the double tail and distance p-values are frequently outperformed. But that begs the question ; which is more powerful between the point and tail p-values. As we can see from Figure 4.1.5 the point and tail p-values may be equally powerful for Fisher's exact test.

We will try to distinguish these two p-values from eachother. We will make frequent use of the concept of unbiased hypothesis tests that we introduced in section 2.2.2 and see that the tail p-value frequently performs better in this regard. However, we will see cases where the point p-value has strictly more power than the tail p-value too.

4.2.1 Realizations of p-values

We have already seen a lot of $p(x)$ against x graphs, so we will not go through them in as much detail again. We have already seen that they outperform double tail and distance p-values. What we will look at more closely is that these two p-values usually dominate each other on each side about the center when one of them is lower than the other.

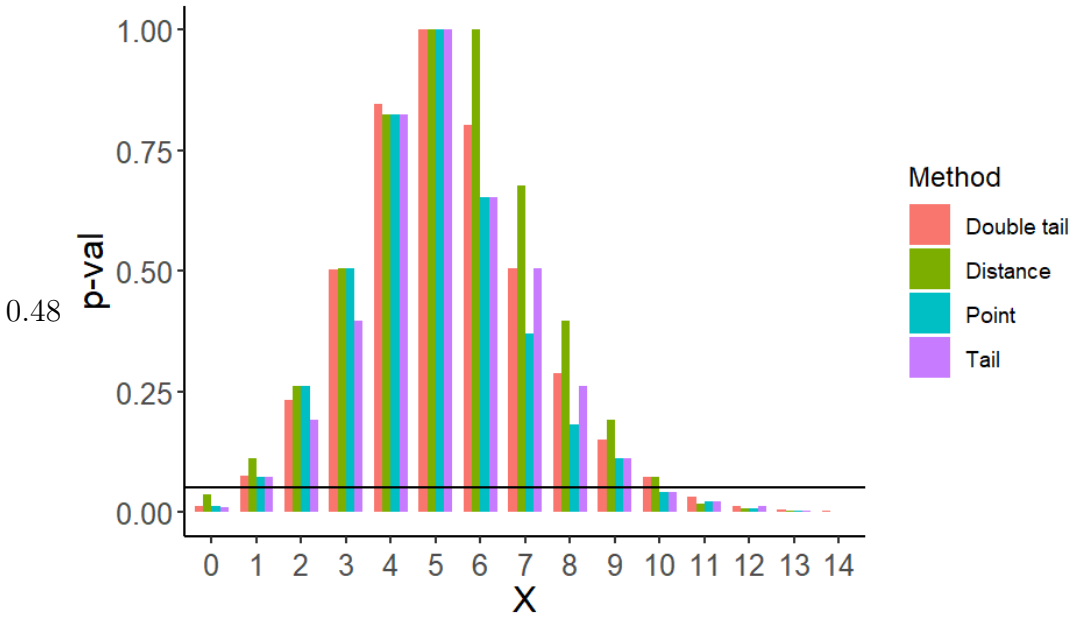
We will plot a Poisson test like the one in Figure 4.1.3, but with $\mu_0 = 5.1$ instead of $\mu_0 = 10$. We are intentionally not using a whole number for μ_0 to avoid a situation with ties for the point probability statistics[1]. We will go in more detail on this later.

We see that the point and tail p-values often produce the same result. At the points $X = 7$ and $X = 8$ we see that the point p-value becomes smaller, while at $X = 3$ and $X = 2$ the tail p-value is smaller. This is an example of the p-value being smaller at one side of the center and larger at the other.

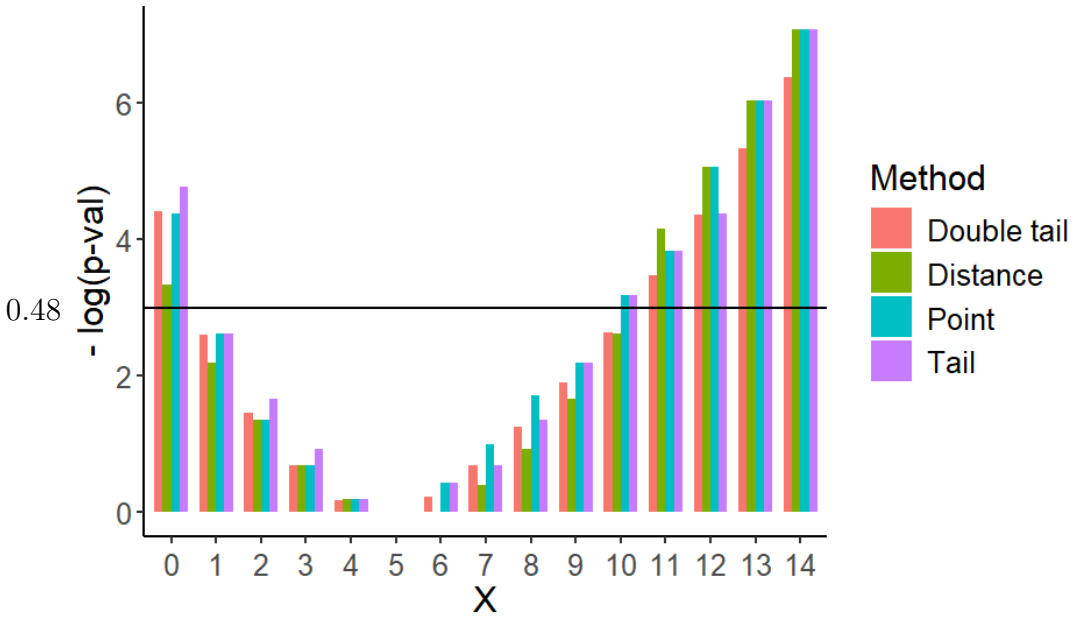
While it is a bit hard to judge from the plot it seems like $p_P(3) = p_T(7)$ where p_P is the point p-value and p_T is the tail p-value. This would make sense as both are p-values calculated by the chance of observing a lower test statistic under the null hypothesis.

If we rank the outcomes x in terms of descending tail p-value $p_T(x)$ we get $5, 4, 6, 7, 3, 8, 2, 9, ...$ If we do the same procedure for the point p-value $p_P(x)$ we get $5, 4, 6, 3, 7, 2, 8, 9, ...$ If we use the definitions of p_P and p_T in

[1]However from the plot we can see that something went wrong for the distance p-value anyway as there is two instances of $p_D(x) = 1$. There is likely a mistake in the calculation of the distance p-value for the Poisson test.



(a) P-value



(b) - log P-value

Figure 4.2.1: P-values as functions of realized values **X** for a Poisson distribution with $\mu_0 = 5.1$. All methods

section 3.2 we get

$$p_P(3) = 1 - P_{\mu_0}(X = 5) - P_{\mu_0}(X = 4) - P_{\mu_0}(X = 6) - P_{\mu_0}(X = 3) \quad (4.1)$$
$$p_T(7) = 1 - P_{\mu_0}(X = 5) - P_{\mu_0}(X = 4) - P_{\mu_0}(X = 6) - P_{\mu_0}(X = 7)\ . \quad (4.2)$$

We see that it is not the same, but if $P_{\mu_0}(X = 3) \approx P_{\mu_0}(X = 7)$ it will be approximately the same. This will often be the case for the cases where they disagree because if there is a large difference then it is unlikely that it will not affect the tail probaility so much that it agrees with the point probability statistic.

We see this pattern multiple times both in the previous Figure 4.1.1 and Figure 4.1.3. More examples can be seen in Appendix D.

4.2.2 Power comparison

It is hard to get a clear preference between point and tail p-values from looking at plots of $p(x)$ vs x. We will now look at power functions $\beta(\theta)$ instead. Since we have already seen Fisher's exact test, and seen no difference in power, we will instead focus on the binomial and Poisson experiments.

We have the binomial experiment as defined in section 2.3.1. We may then calculate the power function $\beta(\theta)$ for discretized values of θ_0 and θ. This discretization is done by splitting the interval $[0, 1]$ into $N = 200$ subintervals with a total of 201 nodes, and we calcualte the power function for each of these nodes.

We are interested in the point and tail p-values and thus we only present the comparison between these. This is represented in Figure 4.2.2. Red areas indicate that the point p-value has more power, while green areas indicate that the tail p-value has more power. Along the y-axis we have the true parameter value θ, along the x-axis we have the null hypothesis parameter θ_0.

We see in Figure 4.2.2 that for the most part the power functions of point and tail p-value are the same for the binomial test (of $n = 10$). We see that the regions where any difference is detected is along lines with constant θ_0 and that in these lines there is a switch between which value becomes more powerful depending on the true parameter θ.

We plot the power functions along one such slice $\theta_0 = 0.71$ in Figure 4.2.3. In this figure we see that the tail p-value is the only one that manages a two-sided rejection region. This makes it a much better candidate for a two-sided hypothesis test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. We see that there are regions where the point p-value has more power than the tail p-value, but the large difference in power that arises for $\theta > 0.71 = \theta_0$ is

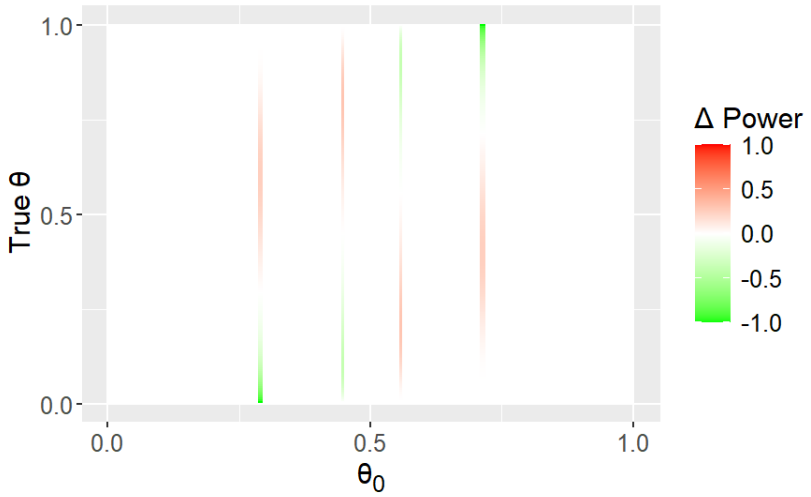


Figure 4.2.2: Power comparison between Point and Tail methods for varying θ_0 and θ in a test for the probability parameter in a binomial distribution. Red areas indicate the point p-value is more powerful while green areas indicate that the tail p-value is more powerful.

clearly more important. Another way of putting this preference for the tail p-value in this case is that the tail p-value produces unbiased (or at least closer to unbiased) results, i.e. $\beta_T(\theta) \geq \beta_T(\theta_0)$ for $\theta \neq \theta_0$. Meanwhile, we have $\beta_P(1) = 0 < \beta_P(\theta_0)$ which is clearly something we wish to avoid.

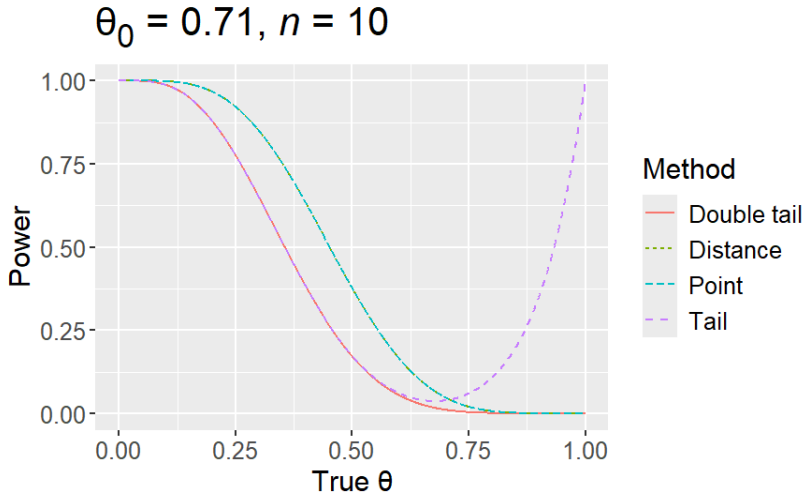


Figure 4.2.3: Power function for all methods for $\theta_0 = 0.71$:

We will now switch our attention to the Poisson experiment, defined as in

section 2.3.3. Again we are only interested in the point and tail p-values and the difference in power is shown in Figure 4.2.4. Green areas have the point p-value being more powerful while red areas indicate that the tail p-value is more powerful. The discretization of μ and μ_0 was done with 50 equividistant points from 0.2 to 20.

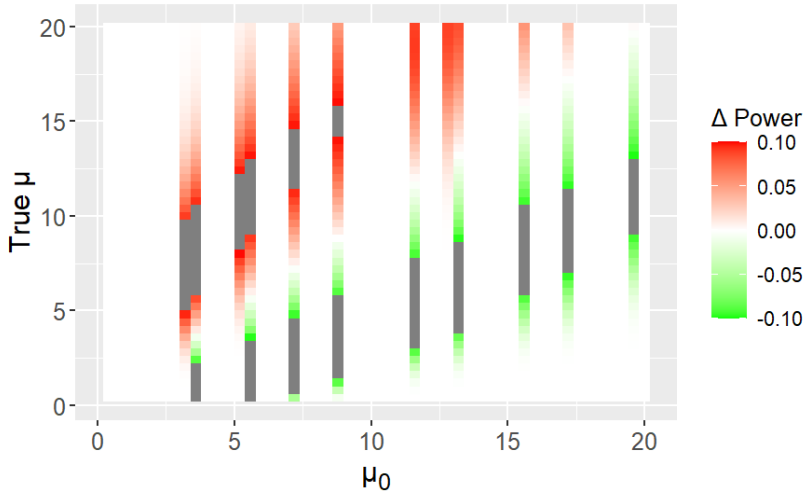


Figure 4.2.4: Example of power heatmap. Power of Point method and Tail method compared with eachother. Red areas indicate that the point p-value being more powerful while green areas indicate tail p-value being more powerful.

We observe in Figure 4.2.4 that when we have any difference they are along constant lines of μ_0. We have both slices where which p-value is more powerful depends on μ and slices where one always is more powerful. In the cases where they are different the tail p-value is better for $\theta < \theta_0$ while the point p-value is better for $\theta > \theta_0$. We take a closer look at some slices, graphing the power functions.

Let us look at the case of $\mu = 19.8$ in Figure 4.2.5 first. Notice that the x-axis has been extended to $\mu = 40$. Here we see that the point p-value starts being less powerful until it takes over right around $\mu > \mu_0$ from which it becomes more powerful than the tail p-value. Interestingly, we clearly see that the point p-value has a minima for the power function $\beta_P(\mu)$ before μ_0 while the tail p-value seems to have its minima at μ_0 (or at least very close to it). Thus, the tail p-value is unbiased (or at least closer to it) while the point p-value is not.

For $\mu_0 = 5.2$ we see maybe the most interesting case of this entire thesis in Figure 4.2.6. Here we see that the point p-value is always more powerful

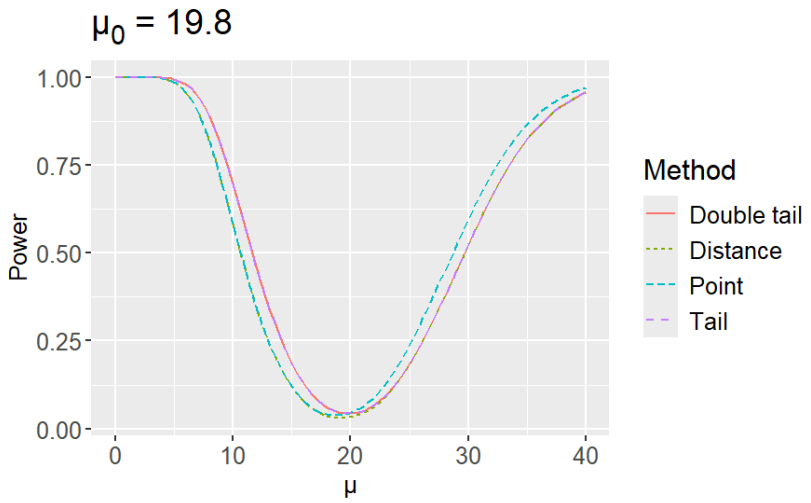


Figure 4.2.5: Power function for all methods for $\mu_0 = 19.8$.

than the tail p-value. However, we also see that the tail p-value is closer to having a minima at $\mu_0 = 5.2$ while the point p-value seems to have it noticably earlier. Thus, the tail p-value is once again better at producing an unbiased hypothesis test. Therefore, it is really a question of which property one deems more desirable, unbiasedness or more power.

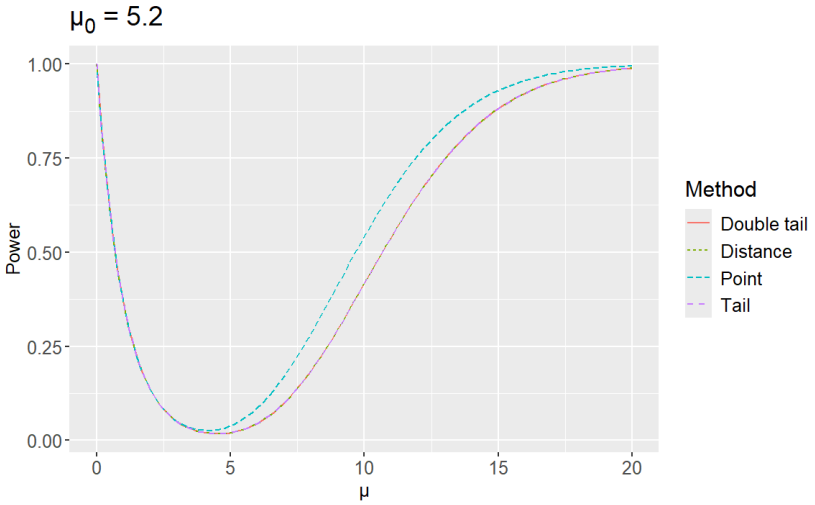


Figure 4.2.6: Power function for all methods for $\mu_0 = 5.2$.

4.2.3 Conlusion

We have only seen three instances of the tail p-value being better than the point p-value at being unbiased so it is dangerous to draw generalized conclusions. It seems like there might be more cases, particularly for the Poisson distribution, but we have not investigated more to the degree that it is presentable. This pattern of the tail p-value being better at producing unbiased hypothesis test would be very interesting if true, seeing as it was not among those recommended by Agresti (1992). This is a very interesting topic for further research. It would also be interesting to see if it was possible to get any differences between point and tail p-values for other parameters in Fisher's exact test, namely trying to make the distribution more asymmetric.

4.2.4 Current practice

The R Documentation for **fisher.test** states that "Two-sided tests are based on the probabilities of the tables, and take as 'more extreme' all tables with probabilities less than or equal to that of the observed table, the p-value being the sum of such probabilities" (R Core Team, 2024c). This means that R uses the point p-value for Fisher's exact test. The documentation for the exact binomial and poisson tests do not make special mention of the p-value used (R Core Team, 2024a) ,(R Core Team, 2024b). However, by running some examples we see that they coincide with our calculations for the point p-value, making it highly likely that thats their method.

P-value method	**Value**
$p_P(x = 5 \mid n = 10, \theta_0 = 0.7)$	0.1785159
$p_T(x = 5 \mid n = 10, \theta_0 = 0.7)$	0.2995767
binom.test	0.1785
$p_P(x = 5 \mid \mu_0 = 10)$	0.1505444
$p_T(x = 5 \mid \mu_0 = 10)$	0.1158264
poisson.test	0.1505

Table 4.2.1: Comparison of realized point and tail p-values with the same parameters done by R Core Team

This makes the point p-value a more convinient default within research for now. But if it should stay like this is another question. Perhaps it would be better to have the tail method be the default.

4.3 Limitations

So far we have concluded that the point and tail p-values produce more powerful tests than the double tail and distance p-values. We have also seen indication that the tail p-value is better at producing unbiased tests. We will go though some limitations that might undermine these results.

One thing we will not discuss is the computational efficency of different p-values. This could be an important consideration in discussing which two-sided p-values should be more established. However, we have not spent enough time investigating computational efficency to discuss it.

We have used numerical calculation of the power functions, not an algebraic expression. Thus, we have only looked at some specififc instances. This can generally be split into two parts, those that are linked to a specific choice of the number of trials and those linked to discretization of the parameter space Θ. We will first look at limitations conncected to having used a specific n. Then we will loo at discretization. Finally, we will also look at problems connected to ties of the extremity statistic.

4.3.1 Specification of known parameter values

For tests of the binomial parameter like in section 2.3.1 we have used the following number of trials $n = 10, 30, 50$. For Fisher's exact test like in section 2.3.2 we have used the following tuples $(n_a, n_b) = (25, 20), (45, 40), (30, 15)$. For the Poisson we have no such additional parameters. For power calculation and comparison we used $n = 10$ for simple binomial and $(n_a, n_b) = (40, 50)$ for Fisher's exact test.

Notably, this might not have created a signifcantly skew distribution for Fisher's exact test. This is a major limitation that should be rectified in further research on the topic as we have seen no difference between the point and the tail p-values for this test and there very well might be for more skew cases. Also, it is very interesting if this does not happen, seeing as Fisher's exact test is the only one of our tests which is constructed with conditioning on a statistic which is only sufficent under the null hypothesis (as is done in Theorem 3.2). If we fail to produce difference between these two it might be connected to this construction and the fact that both are based on probabilites under the null in some capacity.

4.3.2 Discretization of parameter space

Let us split the issue of discretization into cases dependant on which experimental setup that was used.

(i) For the binomial parameter we used $N = 200$ giving us 201 equividistant points as the discretization of $[0, 1]$ with both edges included. This means that the values for θ_0 and true θ that we looped through was $0, \frac{1}{200}, \frac{2}{200}, ..., \frac{199}{200}, 1$. This is a quite fine discretization, but it is entirely possible that there are differences in power that are not picked up upon as many of the lines indicating difference in the power plot, being very thin. In particular, the heatmap of the difference between the point and tail p-value has some lines indicating difference that is just a single line wide. This makes it entirely possible that similar bands exist where there is a difference between tail and point p-values for a small interval of the parameter space but that this interval is small enough that it fits between points.

(ii) For Fisher's exact test we discretized the interval $[0, 1]$ into 21 equividistant points with both the edges included. That is the points $0, \frac{1}{20}, \frac{2}{20}, ..., \frac{19}{20}, 1$. This means that the distance between the discretixed points are quite large and it is very possible that there are values of θ_a, θ_b that will show a difference between point and tail values. We had problems with turning up the $N = 20$ for Fishers exact test. It should however be possible to optimize the code and get a finer resolution.

(iii) For Poisson we used $N = 50$. This is could probably have been done finer and it is entirely possible that we have missed results again. However, when it comes to comparison between the point and tail p-value we had such frequent and consistent results where the point p-values was better for $\mu > \mu_0$ and vice versa. We might find cases where the point p-value is better at being unbiased as this was more infrequent. Finding a systematic way of testing which p-value is closest to producing an unbiased test seems like a good idea anyway.

4.3.3 Ties of extremity statistic

One problem that we discovered was what happened if the test statistic W is equal for two or more points, i.e. $W(x) = W(x')$. This will result in unnecessairly large p-values because both $p(x)$ and $p(x')$ will include the probability for both x and x'. To see the possible improvment consider the test statistic W' which is excatly equal to W except for that it will have $W(x) < W(x')$ in the case where $x = x'$ and $x < x'$. This means that the new p-value p' will have $p'(x) < p(x)$ when x was the lower value in a tie, while being equal in all other cases. Thus, $p'(x) \leq p(x)$ while still having the same level α because W' satisfies 3.1.

Take for example the distance p-value used for testing a Poisson parameter μ. If μ_0 is a whole number we get that $W_D(\mu_0 - k) = |\mu_0 - (\mu_0 - k)| = |k| = |\mu_0 - (\mu_0 + k)| = W_D(\mu_0 + k)$ where k is an integer. This means that

there will be multiple ties. This phenomena will happen in general whenever the means $\mathrm{E}_{\theta_0}[T(\mathbf{X})]$ is a whole or half number.

We saw this very clearly when we calculated the Type I Error for the distance p-value for the binomial experiment (section 2.3.1).

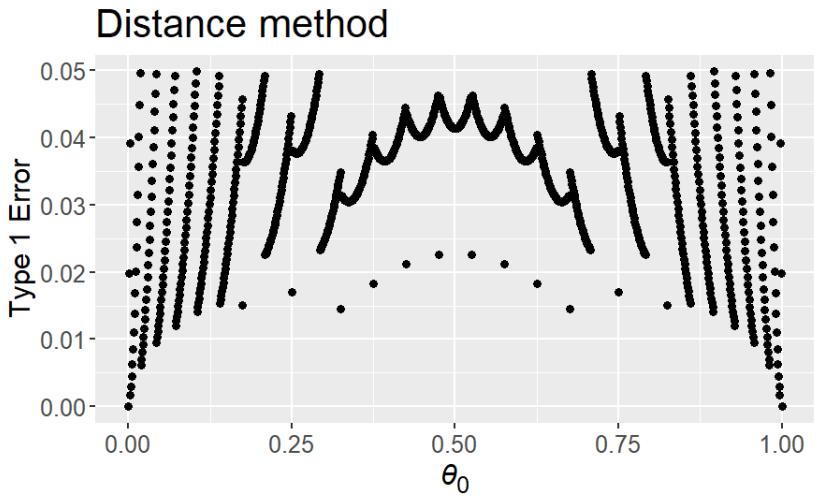


Figure 4.3.1: Type I Error for different $H_0 : \theta = \theta_0$ where θ is the binomial parameter. Distance method

We see in Figure 4.3.1 that there are some sudden drops for the Type I Error. This is at points where we get a sudden tie. It then immediately jumps up again when the tie is resolved by a small increase or decrease in θ_0.

Drops in Type I Error is usually connected to a decrease in power and so it might have influenced some of the results of the distance p-value. We will need to look into the effect ties could have had on power calculations for each different experimental setup.

(i) If it was a problem for the binomial experiment we would expect regularly streaks of very different power, but this doesn't seem to have happened. We therefore conclude that the p-values drops below the level α at the same time and it is not an issue.

(ii) For Fisher's exact test the expected value is less likely to be a whole number since the mean is $\frac{n_a y}{n_a+n_b}$, where the variables are as defined in section 2.3.2. Thus, it doesnt happen as often, but it will have some effect when it does. When power was calculated we iterated through all possible values of y so the effect should have been dispersed out over the whole space as it will get a whole number each time $n_a + n_b$ is a factor in $n_a y$.

(iii) For the Poisson experiment we would also expect streaks where the power is much worse, but from the power heatmap matrix in Appendix D

Figure D.3.7 it seems like the places where it is worse off forms more consistent bands. Thus, ties cannot take the whole blame for distance being bad at acheiving power for the Poisson test. For $N = 50$ p-values in the interval $\mu_0 \in [0, 20]$ we get the discrete points $\mu_0 \in \{0.4, 0.8, 1.2, 1.6, 2, ...\}$ so it is not always an issue. And also the power plots doesn't seem to have regular streaks in them, indicating that this was not a problem. It does not seem like ties had a critical effect on our conclusions.

We see a tie situation for the point p-value and the Poisson test. As we saw in Figure 4.1.3 where we had realizations of $p(x)$ with $\mu_0 = 10$. We see two places where $p_P(x) = 1$ which is due to the fact that

$$P_\mu(X = \mu) = \frac{\mu^\mu}{\mu!}e^{-\mu} = \frac{\mu^{\mu-1}}{(\mu-1)!}e^{-\mu} = P_\mu(X = \mu - 1) \ .$$

However, this is not really an issue for power calculation as both of these point would in practice always have $p_T >> \alpha$ anyway.

Of course, there could be ties in situations we have not even thought about. It does however seem unlikely that it had a critical effect on our results.

Chapter 5

Conclusions

We have investigated the power for four p-values, the double tail p-value, point p-value, tail p-value and distance p-value, applied to some specific experiments. For the binomial and Poisson experiments we found evidence suggesting that the point and tail p-value produce tests of higher power, with the tail p-value in addition seeming to be better at producing unbiased tests. For Fisher's exact test we found evidence of the point and tail p-values producing more powerful tests, but no indication that the tail p-value was better at creating unbiased tests.

A limitation is that during the calculation of power the parameter space was discretized, both for the null parameter θ_0 and for the true θ, or in the case of Fisher's exact test, θ_a and θ_b. This means that there might be cases we are not aware of as power seems to be non-continous when viewed as a parameter of θ_0.

Other parameters that we have not varied sufficently are the trial number parameters n, particularly for Fisher's exact test this might have been a problem where the gap between n_a and n_b was not large enough, leading to relatively symmmetric distributions.

In conclusion, we have some evidence of the point and tail p-value being able to produce more powerful two-sided tests and weak evidence of the tail p-value being better at producing more unbiased two-sided tests making it a better choice overall. This is interesting as the current default in R is to use the point p-value in all the cases we studied.

5.1 Future research

We might need to study the point and tail p-value for other parameter values of Fisher's exact test as the one we currently have used might not have given

a signifcant enough skewness. It would be interesting to see if the equalitt between the point and tail p-values continues. If it does, it might be worth looking into if it is provable. And even more ambitously, if it holds for every test where we condition on a sufficient statistic.

In this thesis we have investigated discrete, non-symmetric distributions, but while the non-symmetric property is important for two-sided p-value discussion to be interesting this is not the case for the discrete property. In fact, there might be interesting p-value properties that arise as a consequence of continuity, which we do not see for discrete distributions. This should be studied further.

A final point which might be interesting to give some proper attention to is ways of constructing unbiased tests. This is a natural requirment and intuitively it seems particularly useful for two-sided hypotheses. It might then be fruitful to bring it more actively into our exploration of two-sided p-values.

References

Agresti, Alan (1992). "A Survey of Exact Inference for Contingency Tables". In: *Statistical Science* 7.1, pp. 131–177.

Casella, George and Roger L. Berger (2002). *Statistical Inference*. Second Edition. Cengage.

Dåsvand, Mathias (2022). "Two-sided p-values for a non-symmetric distribution of test statistics". unpublished Bachelor Thesis, available on NTNU Open.

Dunne, Adrian, Yudi Pawitan, and Liam Doody (1996). "Two-Sided P-Values from Discrete Asymmetric Distributions Based on Uniformly Most Powerful Unbiased Tests". In: *Journal of the Royal Statistical Society. Series D (The Statistician)* 45.4, pp. 397–405.

Fisher, Ronald A. (1925). *Statistical Methods for Research Workers*. 12th Ed.

Hubbard, Raymond and Maria Jesus Bayarri (2003). "Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing". In: *The American Statistician* 57.3, pp. 171–178. DOI: 10.1198/0003130031856.

R Core Team (2024a). *Exact Binomial Test*. Vienna, Austria: R Foundation for Statistical Computing. URL: https://search.r-project.org/R/refmans/stats/html/binom.test.html.

— (2024b). *Exact Poisson Test*. Vienna, Austria: R Foundation for Statistical Computing. URL: https://search.r-project.org/R/refmans/stats/html/poisson.test.html.

— (2024c). *Fisher's Exact Test for Count Data*. Vienna, Austria: R Foundation for Statistical Computing. URL: https://search.r-project.org/R/refmans/stats/html/fisher.test.html.

Appendices

Appendix A

Github repository

Specific code used for the project may be found here. The specifics for a certain plot may not be easily found in the repo, but code for implementation of p-value methods as well as how the comparisons were made is here.

All code is in R.

- `https://github.com/olibolli/prosjektoppgave`

Appendix B

Validity of Double Tail p-value

Denote the two tail probabilities $P_r(x) = P(\mathbf{X} \geq x)$ and $P_l(x) = P(\mathbf{X} \leq x$ and let $p(\mathbf{X}) = 2\min(P_r(x), P_l(x), \frac{1}{2})$. Let α be a number $0 < \alpha < 1$.

$$
\begin{aligned}
& P_\theta(p(\mathbf{X}_{ob}) \leq \alpha) \\
& = P(2\min(P_r, P_l, \frac{1}{2}) \leq \alpha) \\
& = P(\min(P_r, P_l, \frac{1}{2}) \leq \frac{\alpha}{2}) && \text{(divided by 2)} \\
& = P(P_r \leq \frac{\alpha}{2} \cup P_l \leq \frac{\alpha}{2} \cup \frac{1}{2} \leq \frac{\alpha}{2}) \\
& = P(P_r \leq \frac{\alpha}{2}) + P(P_l \leq \frac{\alpha}{2}) + P(\frac{1}{2} \leq \frac{\alpha}{2}) && \text{(disjoint events*)} \\
& = P(P(\mathbf{X} \geq \mathbf{X}_{ob}) \leq \frac{\alpha}{2}) + P(P(\mathbf{X} \leq \mathbf{X}_{ob}) \leq \frac{\alpha}{2}) + 0 && \text{(definition of } P_l \text{ and } P_r) \\
& \leq \frac{\alpha}{2} + \frac{\alpha}{2} && \text{(discrete random variable)} \\
& = \alpha
\end{aligned}
$$

The tail events being disjoint is perhaps more obvious if we recall that $P_r = P(\mathbf{X} \geq x) = 1 - P(\mathbf{X} \leq x) + P(\mathbf{X} = x) \geq 1 - P_l$. This means that if $P_l \leq \frac{\alpha}{2} < \frac{1}{2}$ we necessarily have that $1 - P_l > 1 - \frac{1}{2} = \frac{1}{2}$. Comparing with previous equation we have that $P_r \geq 1 - P_l > \frac{1}{2} > \frac{\alpha}{2}$. Similarly, we have that $P_l > \frac{\alpha}{2}$ when $P_r \leq \frac{\alpha}{2}$. This means that the events $P_l \leq \frac{\alpha}{2}$ and $P_r \leq \frac{\alpha}{2}$ are disjoint.

Appendix C

All Algorithms for Power

Algorithm 5 Calculate Power, Fisher's exact test

```
pwr ← 0
for i in 0 : n_a do
    for j in 0 : n_b do
        pj ← p_func(i, i + j)
        if pj ≤ α then
            pwr ← pwr + P_{θ_a}(X_a = i) P_{θ_b}(X_b = j)
return pwr
```

Algorithm 6 Calculate Power, Poisson test

```
pwr ← 1
loopExited ← FALSE
loopEntered ← FALSE
while NOT loopExited do
    pi ← p_func(i, θ_0)
    if pi > α then
        pwr ← pwr − P_μ(X = i)
        loopEntered ← TRUE
    else
        if loopEntered then
            loopExited ← TRUE
return pwr
```

For the point and tail p-value we made some modifications to the algorithm so that it could quickly determine whether it was over or below the

signifcance level α. These modifications were as following:

- Point : $P(x) > \alpha \implies$ reject
- Tail : $\min(P_l(x), P_r(x)) > \alpha \implies$ reject
- Tail : $\min(P_l(x), P_r(x)) < \frac{\alpha}{2} \implies$ accept

The same methods which are used for calculating power is also used for calculating the Type I Error as it is essentially the same thing, only requiring that $\theta = \theta_0$. This means that we will vary this parameter only while keeping them equal.

Appendix D

Results

D.1 Outcome space

D.1.1 Binomial

D.1.2 Fischer's exact test

D.1.3 Poisson

D.2 Type I Error

D.2.1 Binomial

D.2.2 Fischer's exact test

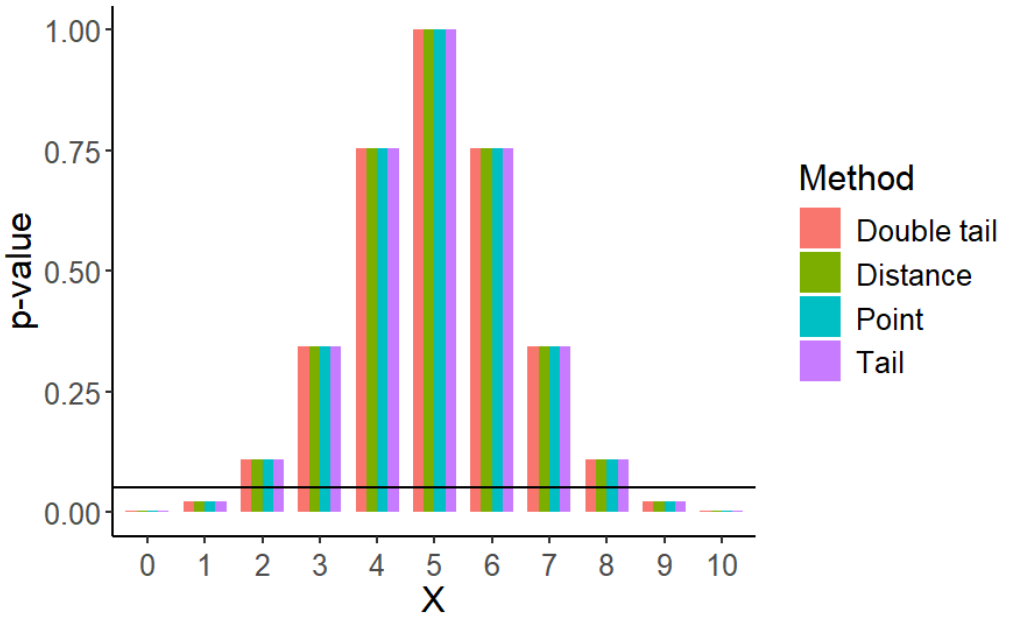
D.2.3 Poisson

D.3 Power

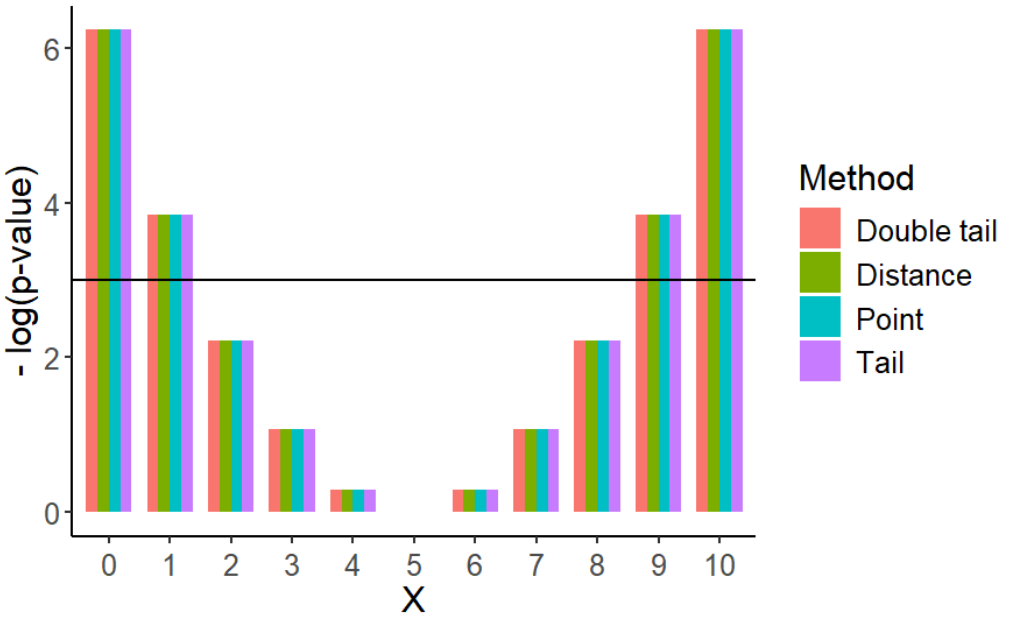
D.3.1 Binomial

D.3.2 Fischer's exact test

D.3.3 Poisson

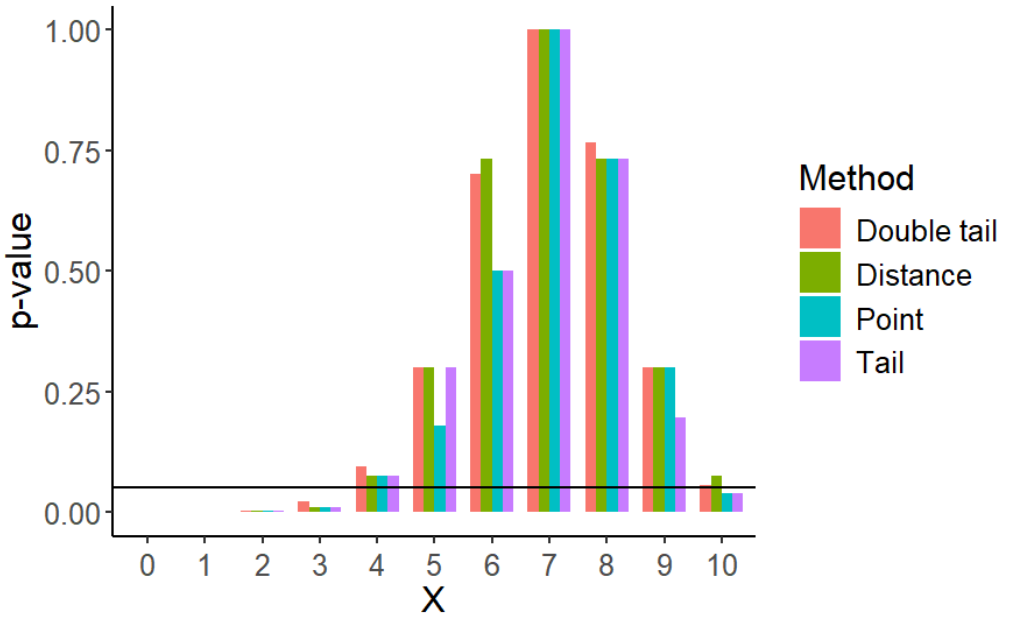


(a) P-value

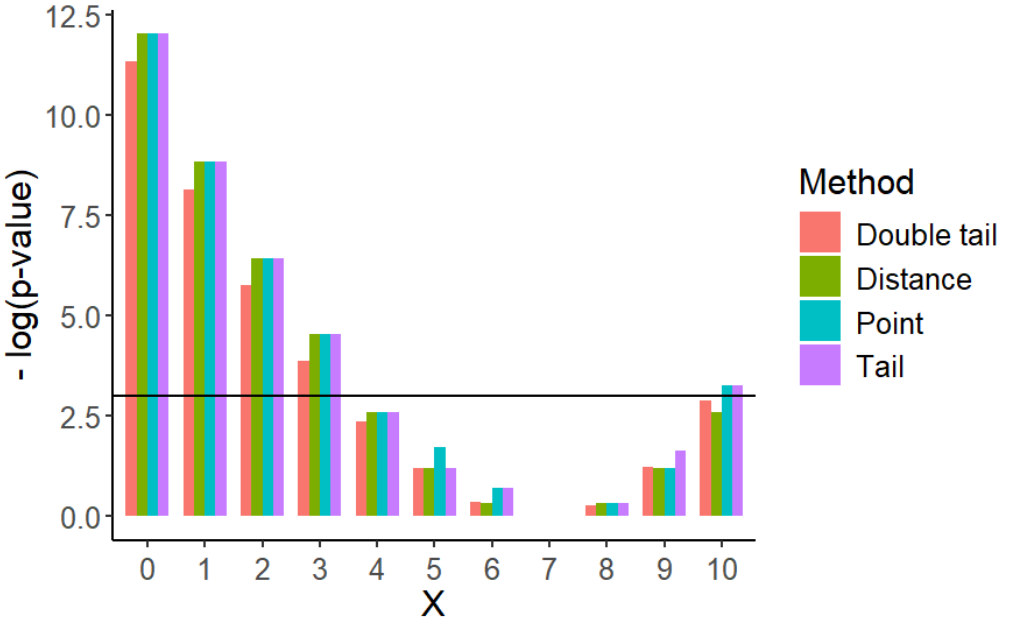


(b) - log P-value

Figure D.1.1: P-values as functions of realized values **X** for a binomial distribution with $n = 10, \theta = 0.5$. All methods

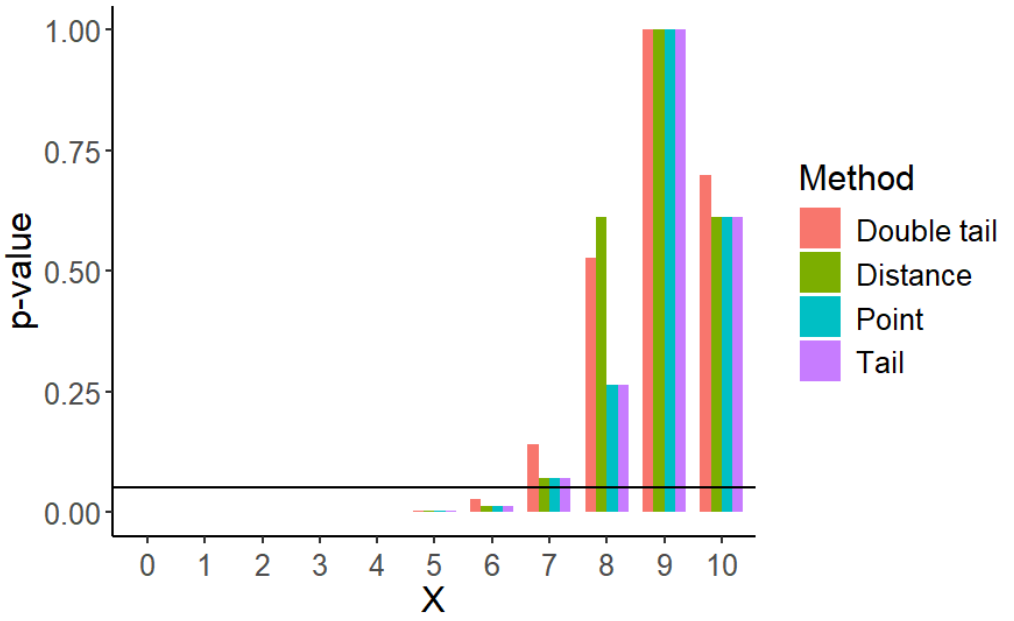


(a) P-value

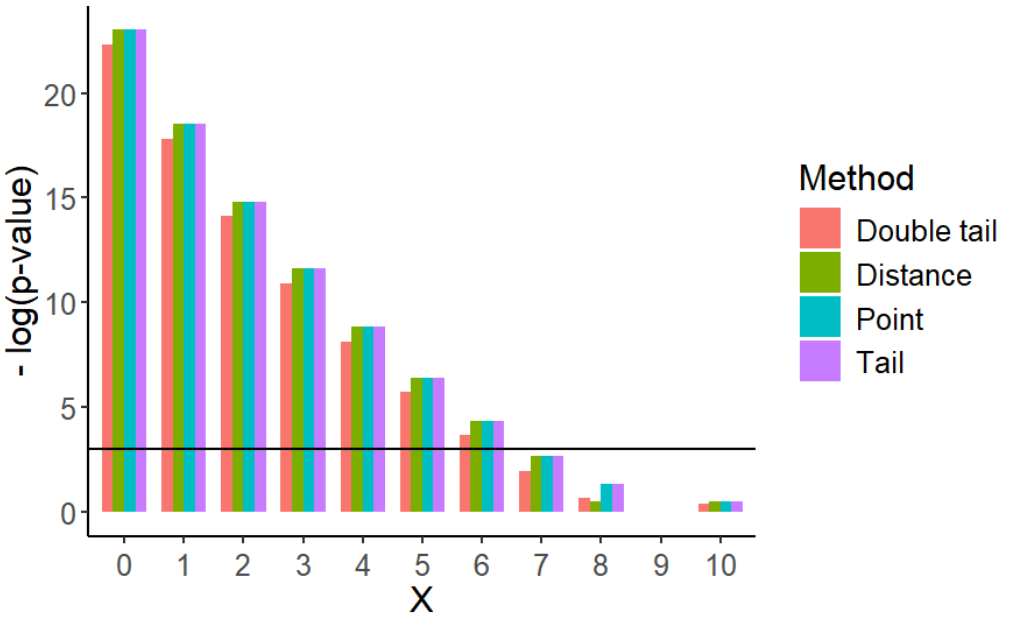


(b) - log P-value

Figure D.1.2: P-values as functions of realized values **X** for a binomial distribution with $n = 10, \theta = 0.7$. All methods

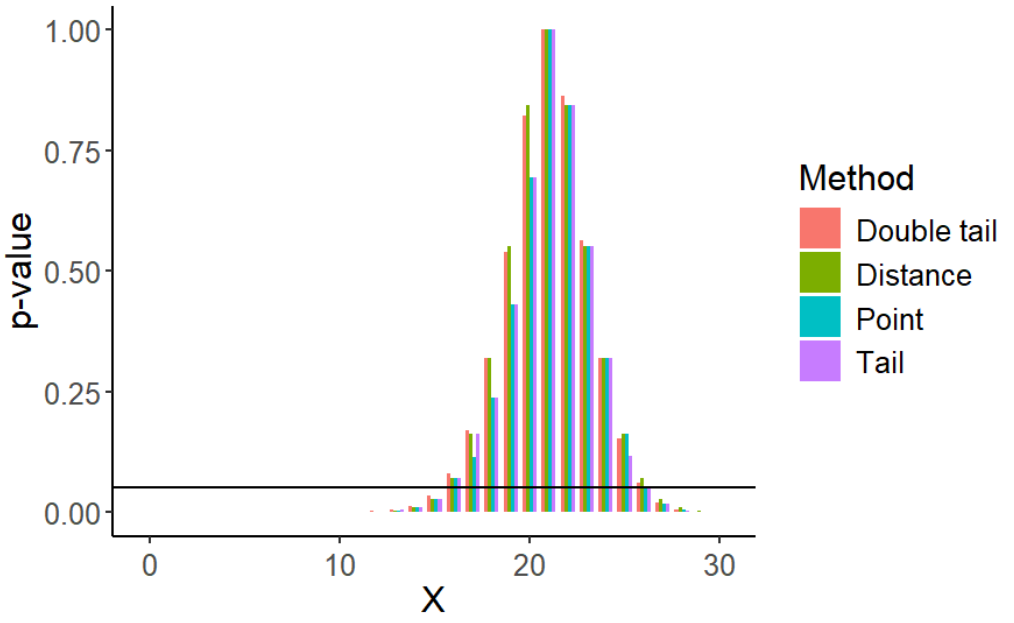


(a) P-value

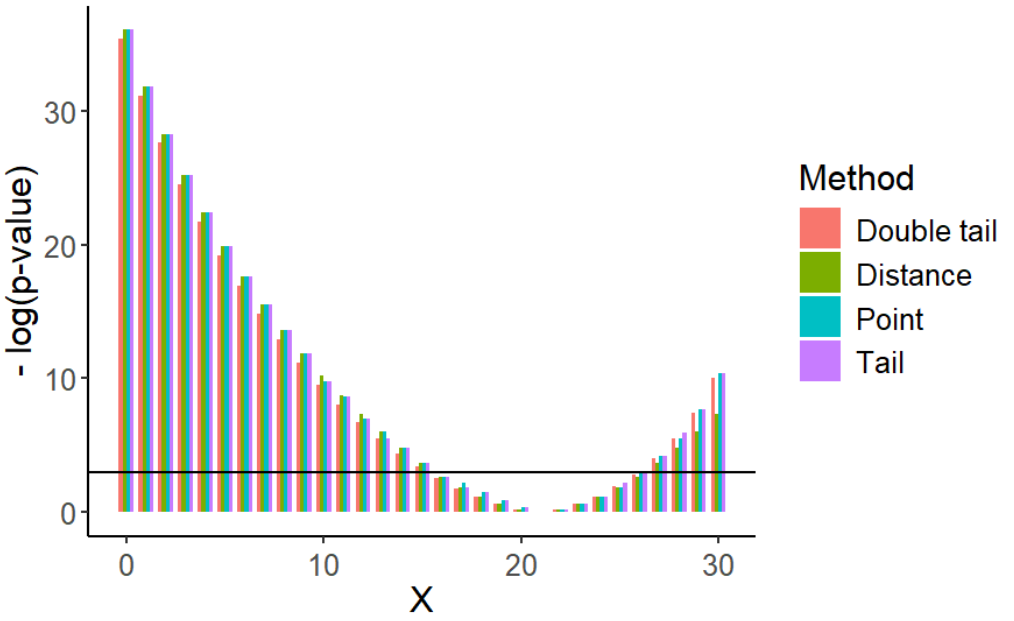


(b) - log P-value

Figure D.1.3: P-values as functions of realized values **X** for a binomial distribution with $n = 10, \theta = 0.9$. All methods

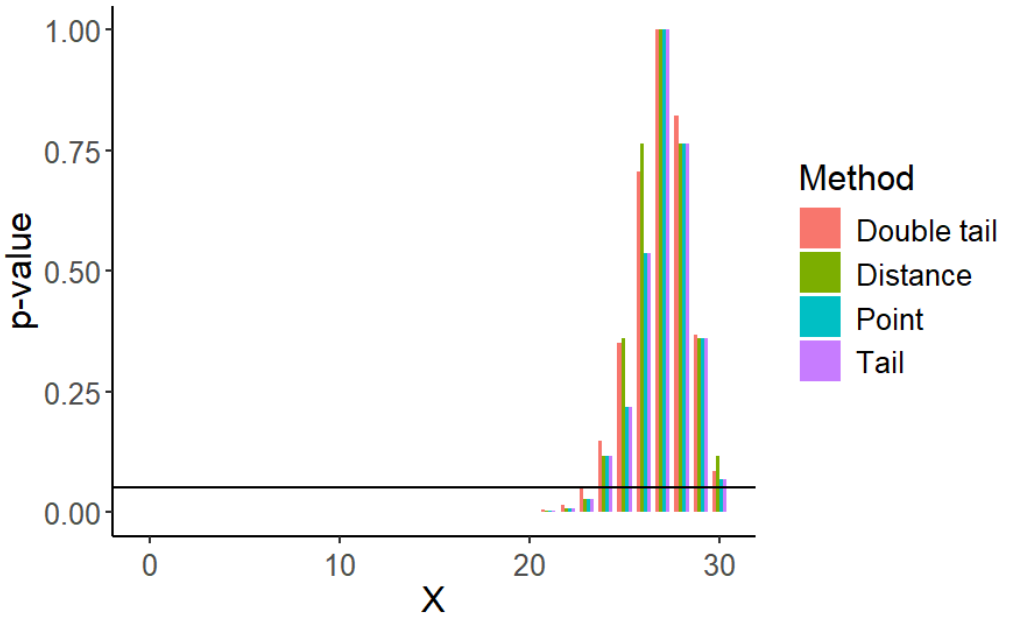


(a) P-value

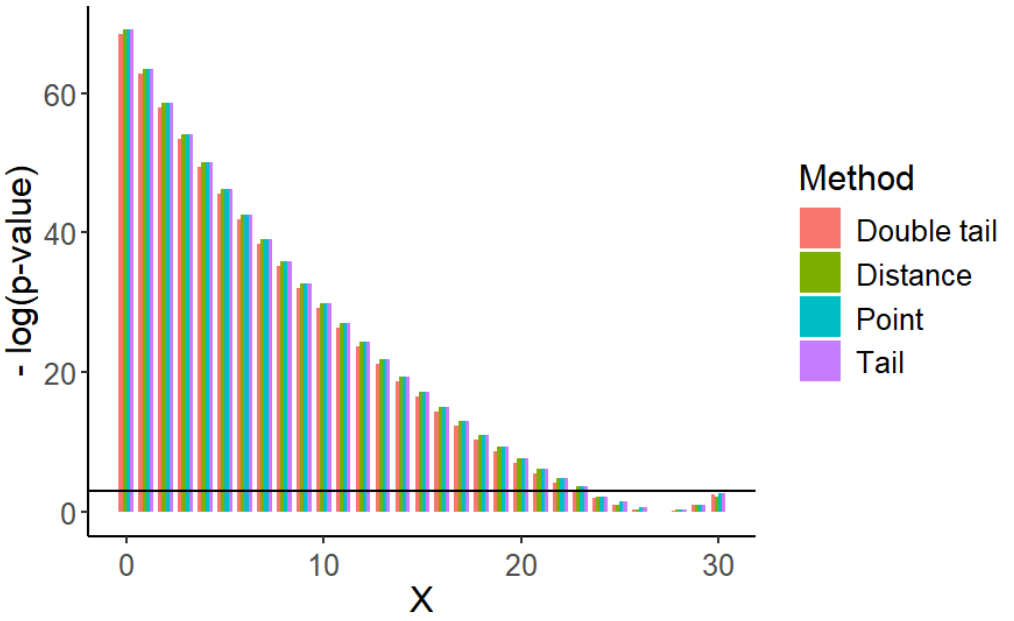


(b) - log P-value

Figure D.1.4: P-values as functions of realized values **X** for a binomial distribution with $n = 30, \theta = 0.7$. All methods

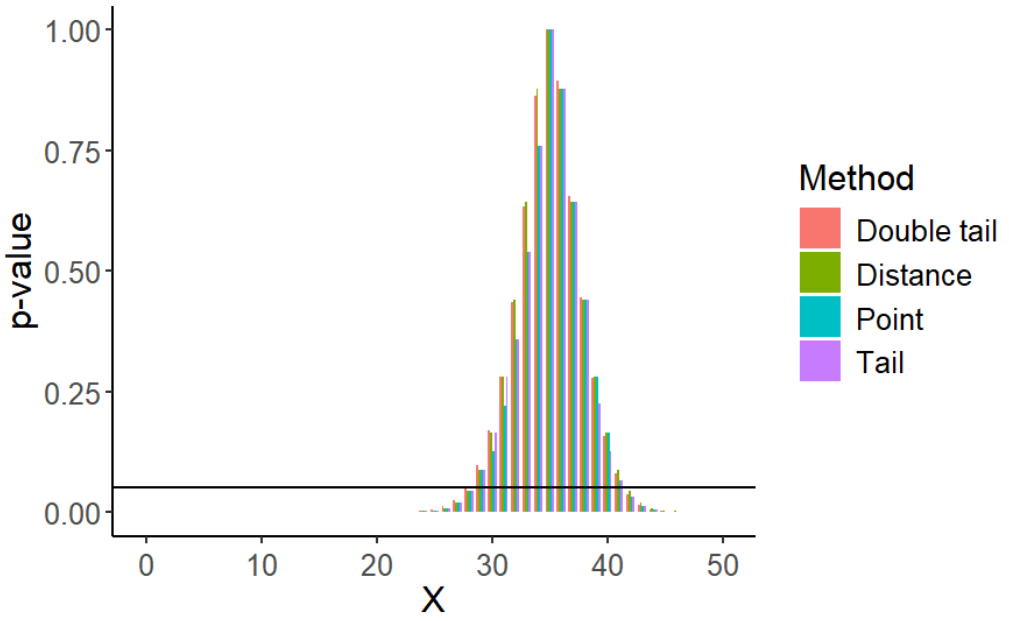


(a) P-value

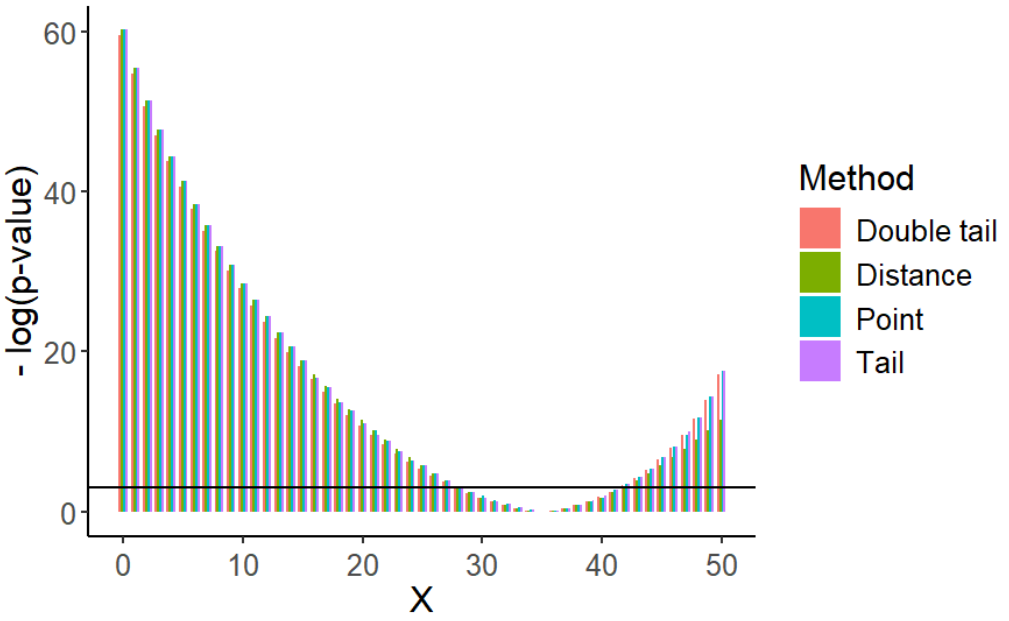


(b) - log P-value

Figure D.1.5: P-values as functions of realized values $\mathbf{X}$ for a binomial distribution with $n = 30, \theta = 0.9$. All methods

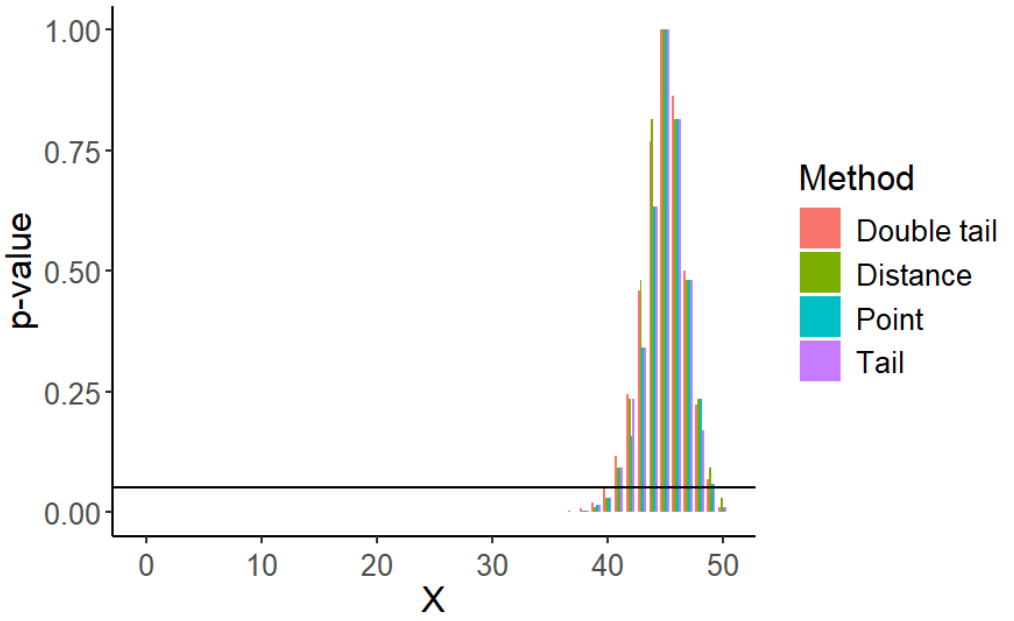


(a) P-value

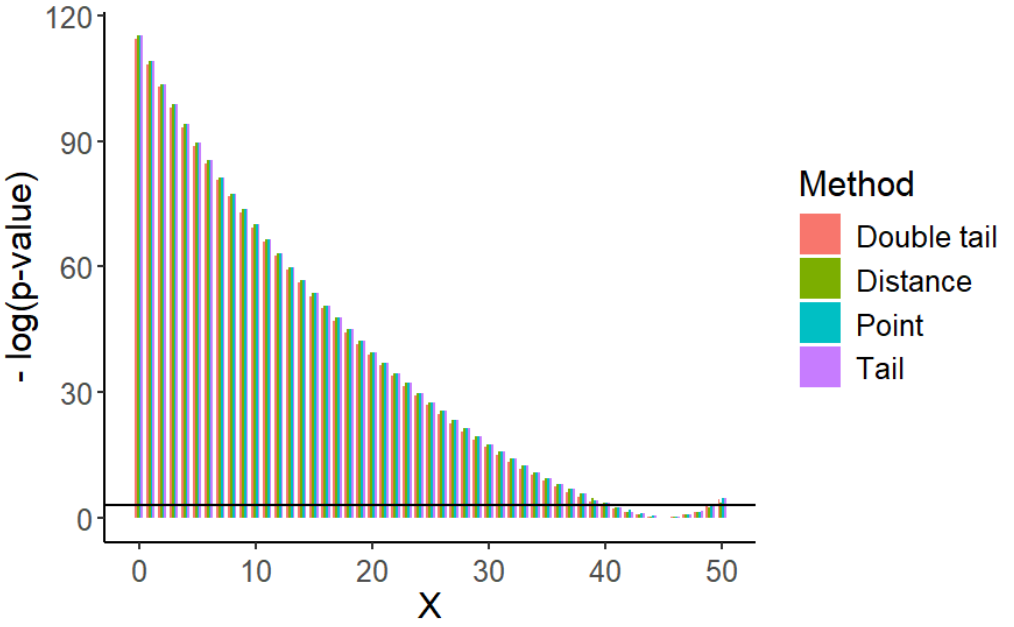


(b) - log P-value

Figure D.1.6: P-values as functions of realized values **X** for a binomial distribution with $n = 50, \theta = 0.7$. All methods

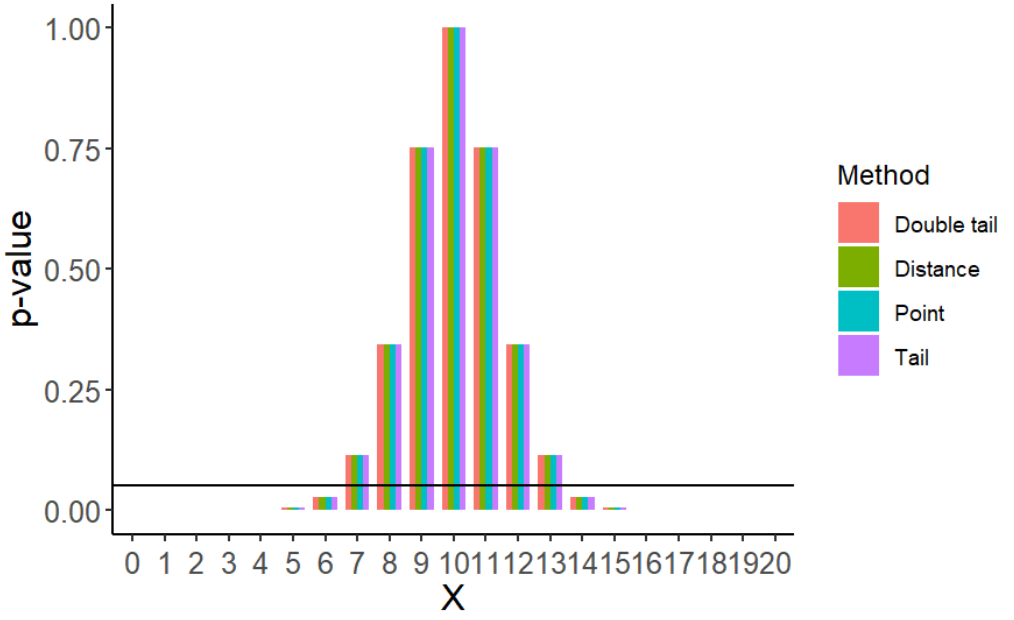


(a) P-value

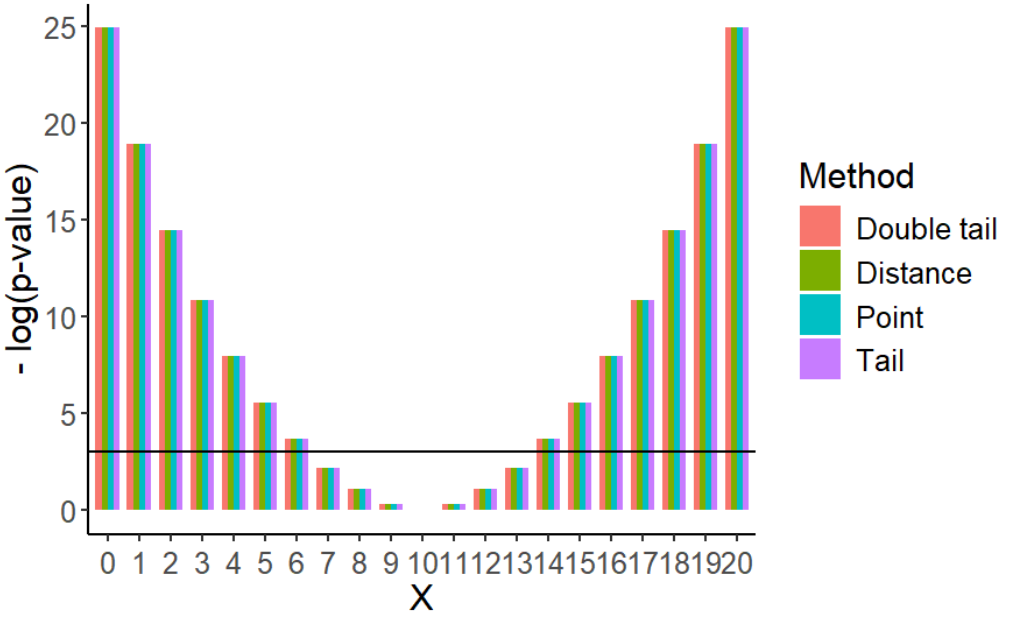


(b) - log P-value

Figure D.1.7: P-values as functions of realized values **X** for a binomial distribution with $n = 50, \theta = 0.9$. All methods

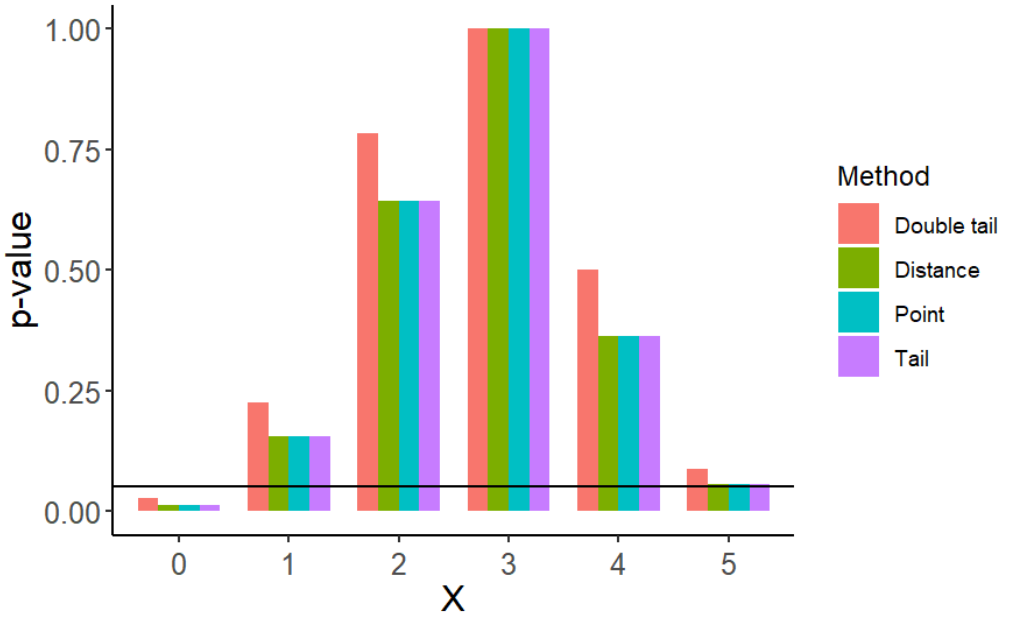


(a) P-value

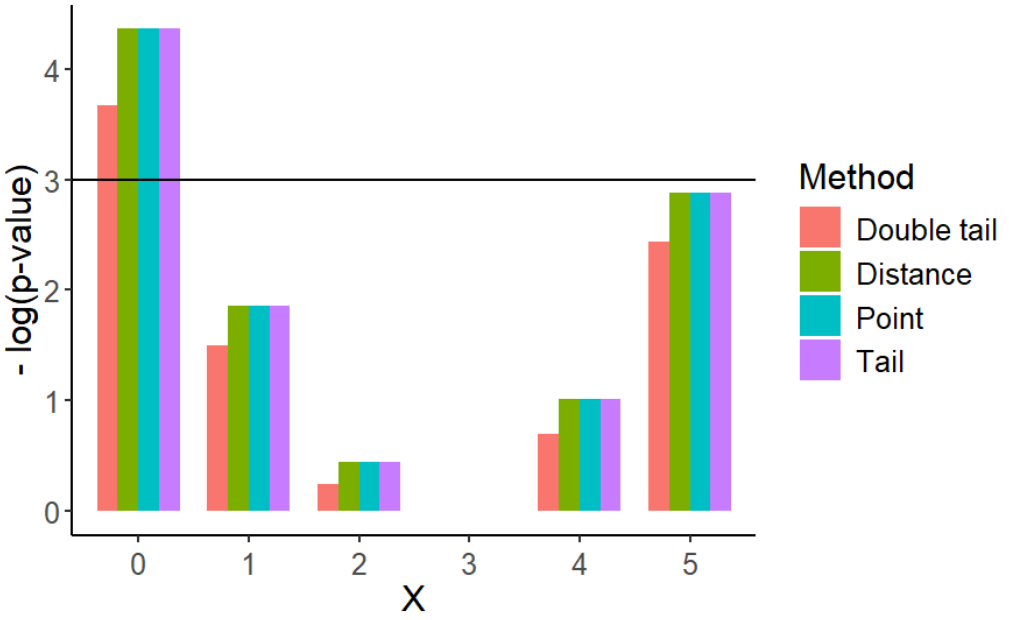


(b) - log P-value

Figure D.1.8: P-values as functions of realized values **X** for the symmetric case of a Fishers exact test with $n = 20/20,\ c = 20$. All methods. $\epsilon = 10^{-10}$ for the tail method.

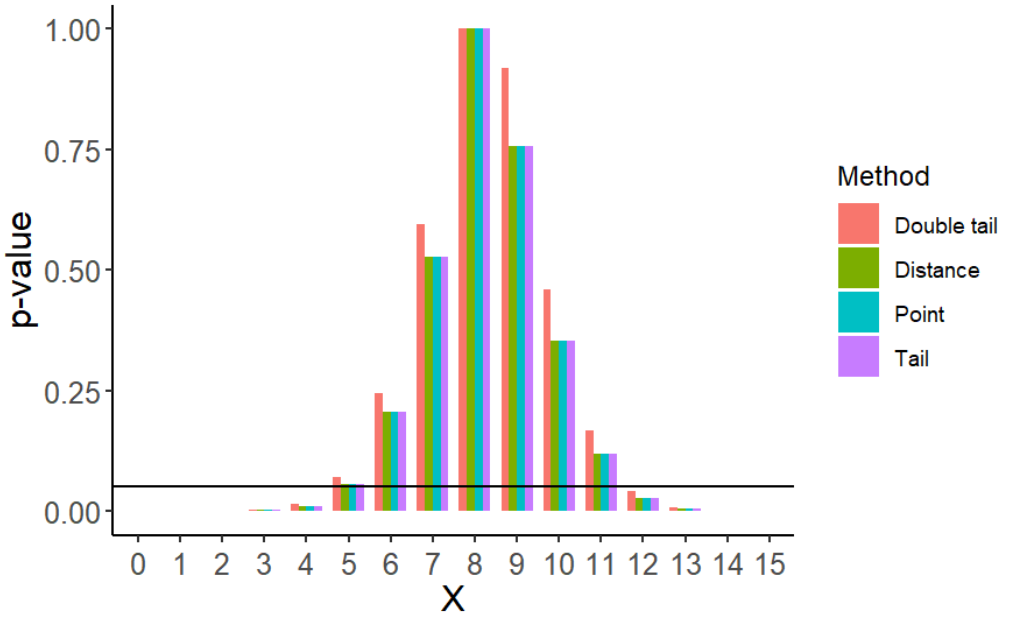


(a) P-value

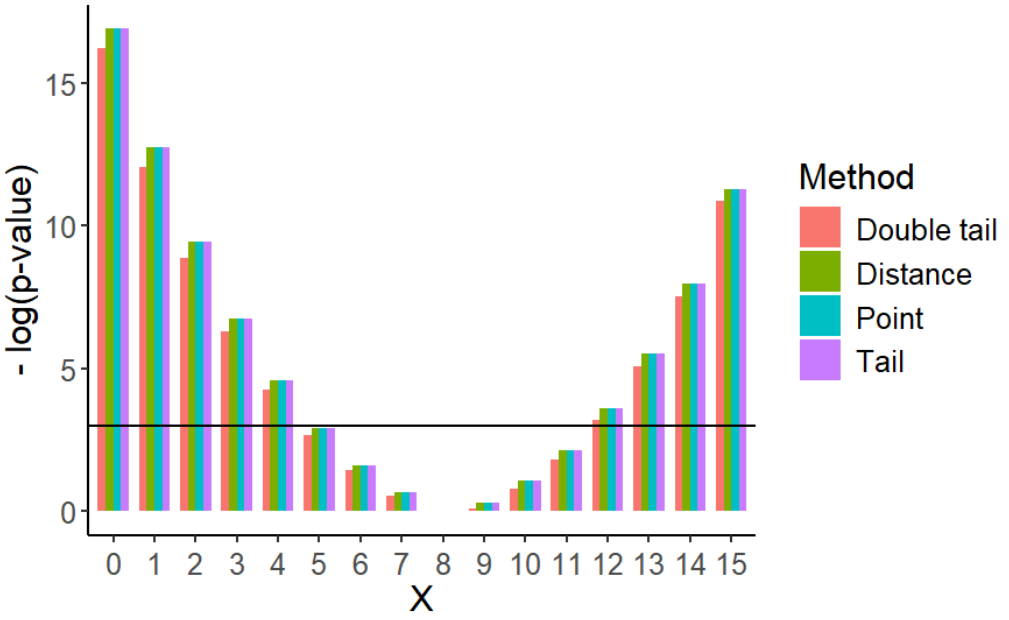


(b) - log P-value

Figure D.1.9: P-values as functions of realized values **X** for Fishers exact test with $n = 25/20,\ c = 5$. All methods

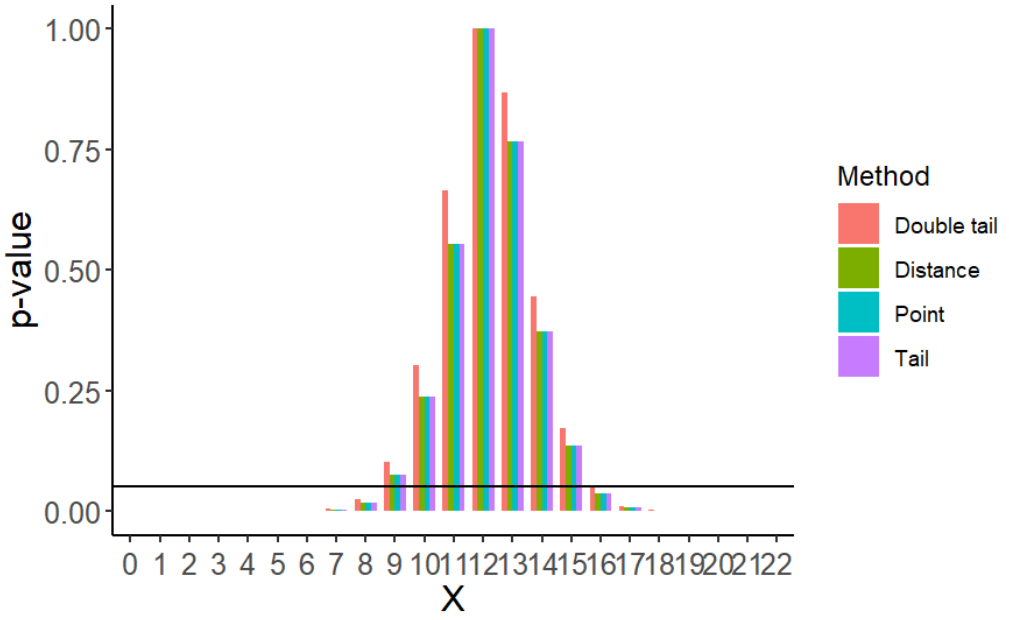


(a) P-value

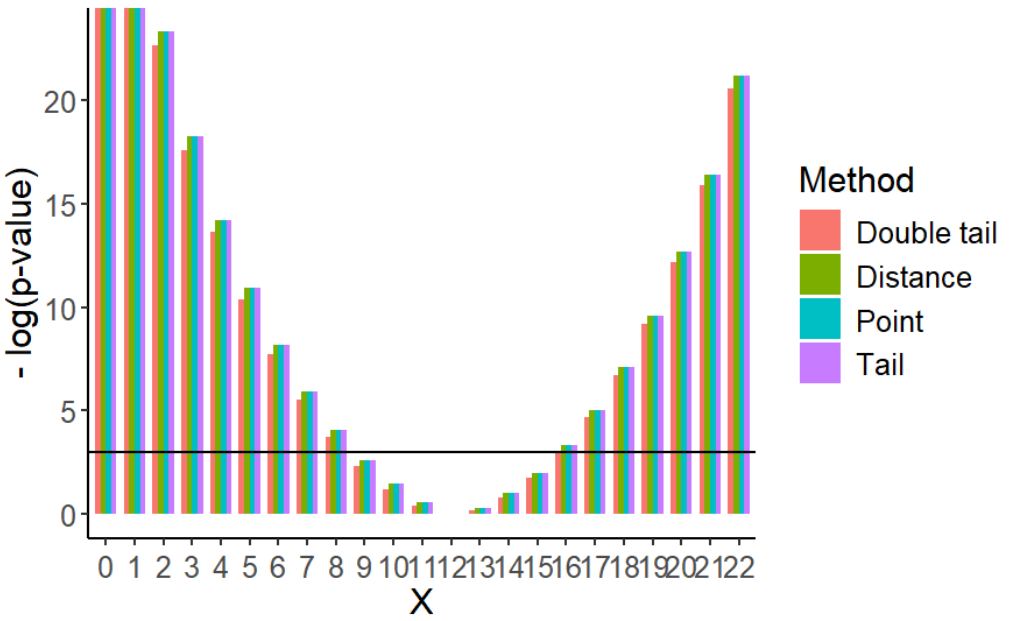


(b) - log P-value

Figure D.1.10: P-values as functions of realized values **X** for Fishers exact test with $n = 25/20,\ c = 15$. All methods

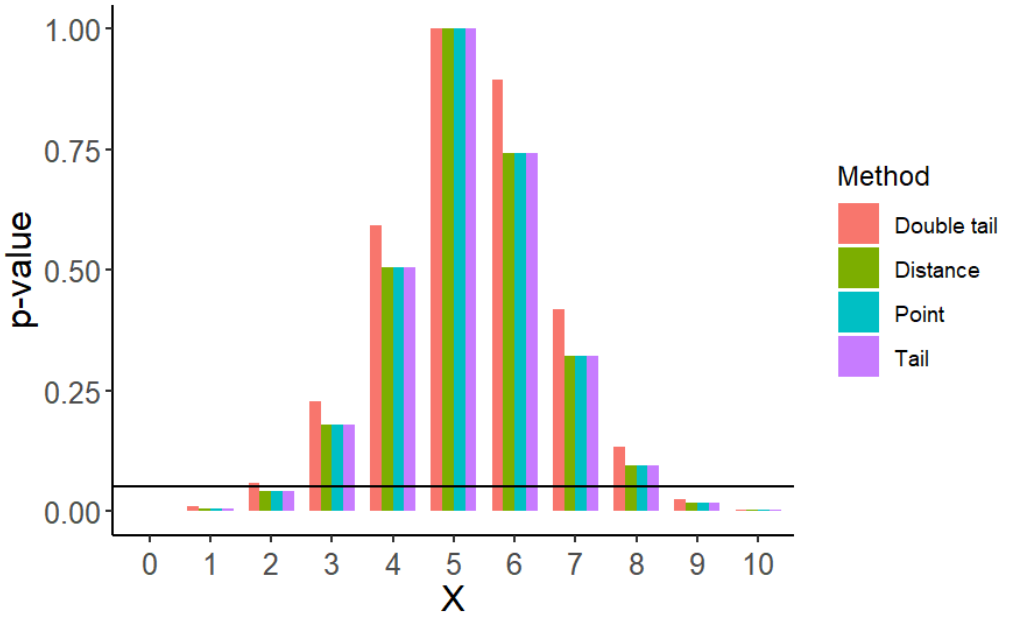


(a) P-value

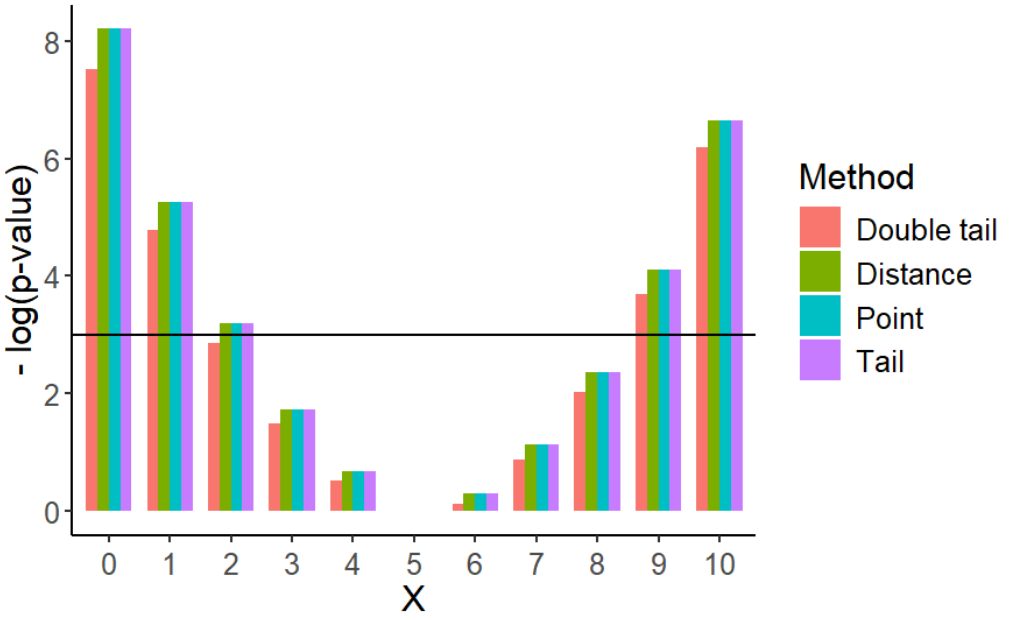


(b) - log P-value

Figure D.1.11: P-values as functions of realized values **X** for Fishers exact test with $n = 25/20,\ c = 22$. All methods

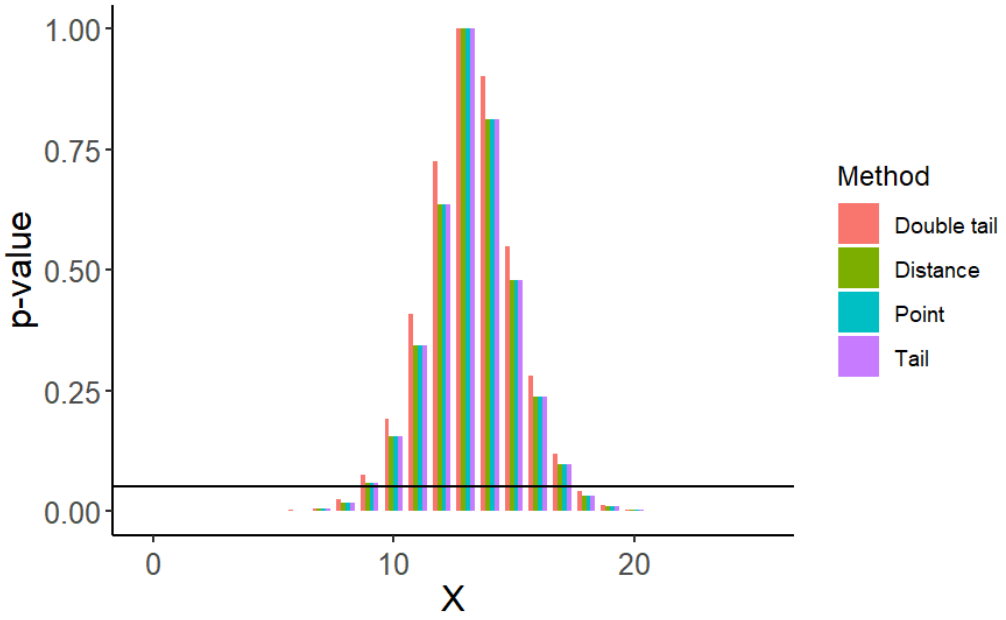


(a) P-value

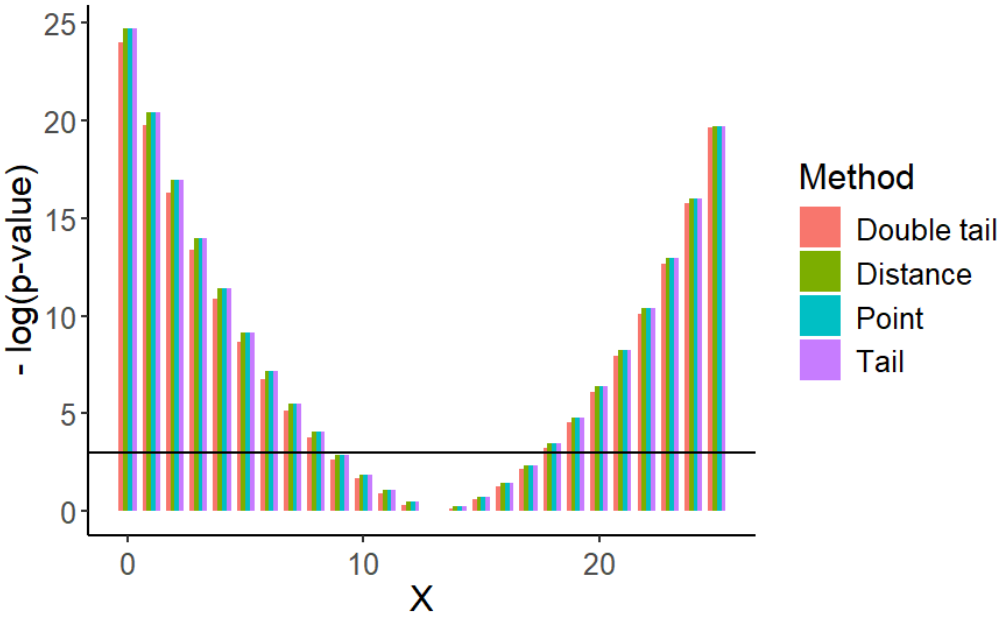


(b) - log P-value

Figure D.1.12: P-values as functions of realized values **X** for Fishers exact test with $n = 45/40,\ c = 10$. All methods

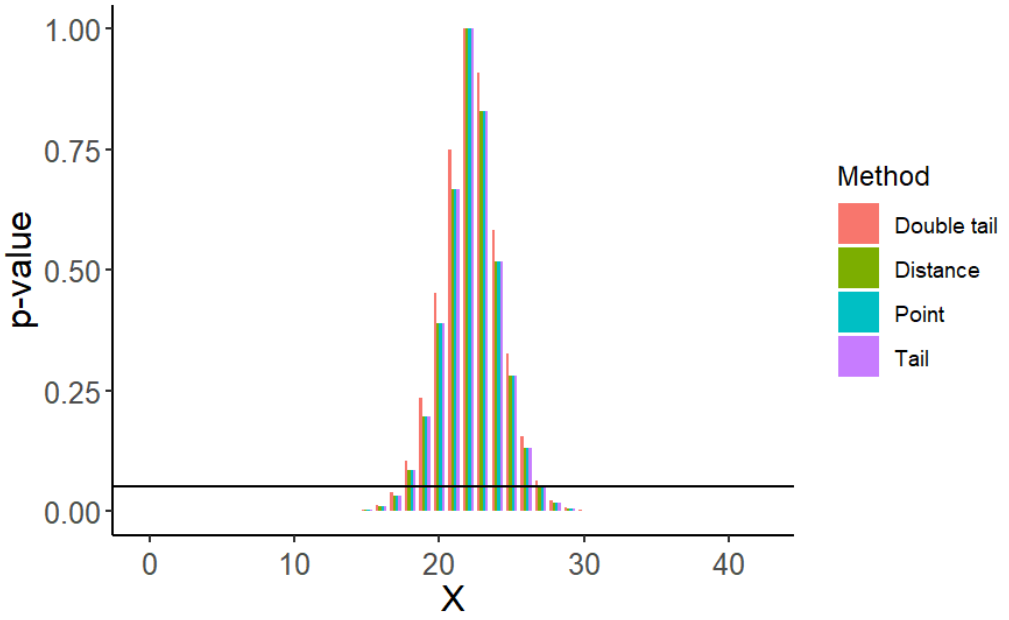


(a) P-value

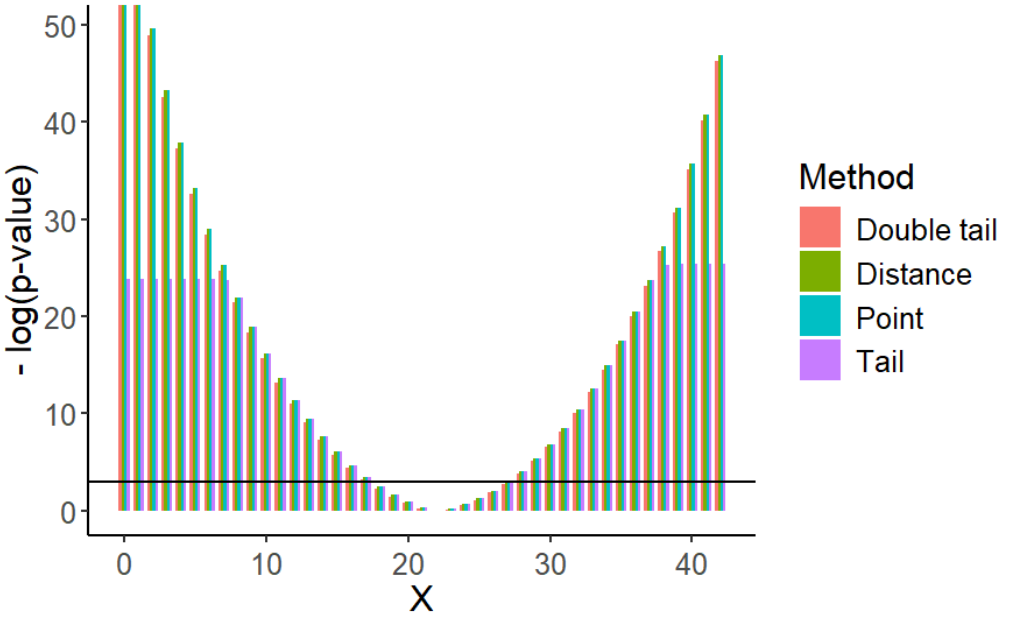


(b) - log P-value

Figure D.1.13: P-values as functions of realized values **X** for Fishers exact test with $n = 45/40,\ c = 25$. All methods

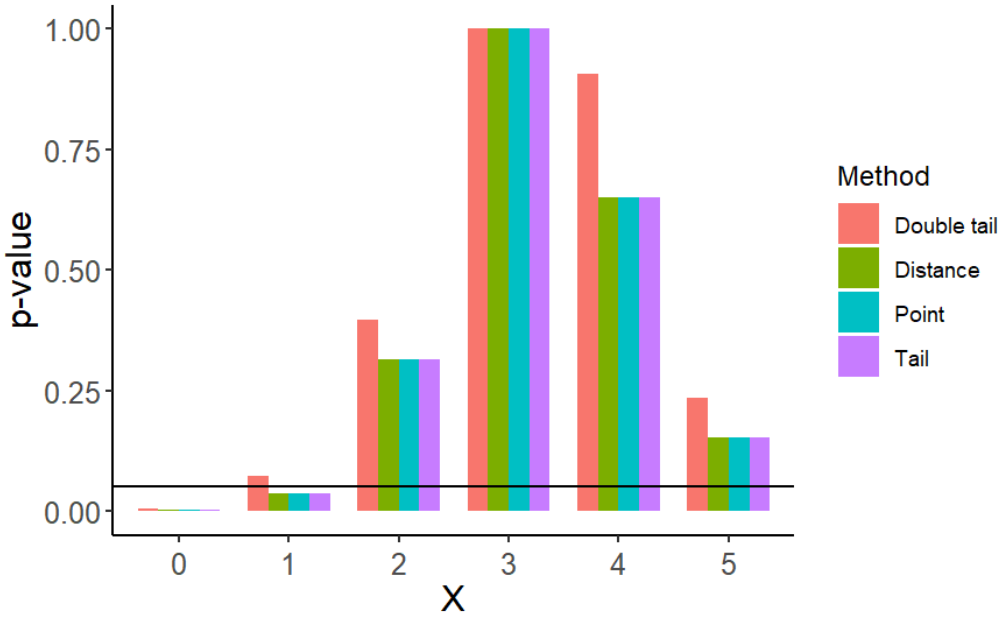


(a) P-value

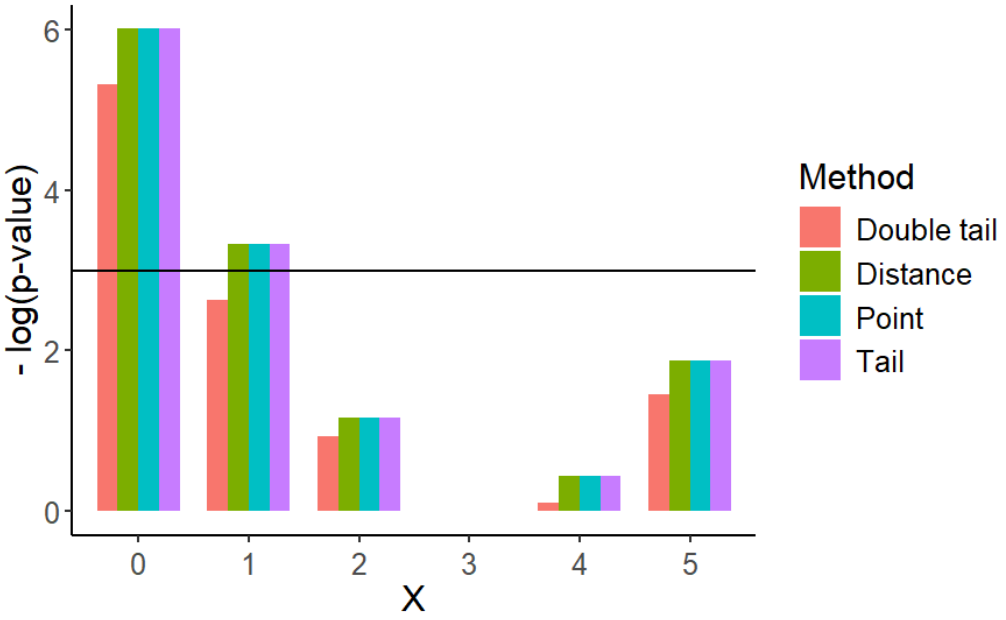


(b) - log P-value

Figure D.1.14: P-values as functions of realized values **X** for Fishers exact test with $n = 45/40,\ c = 42$. All methods

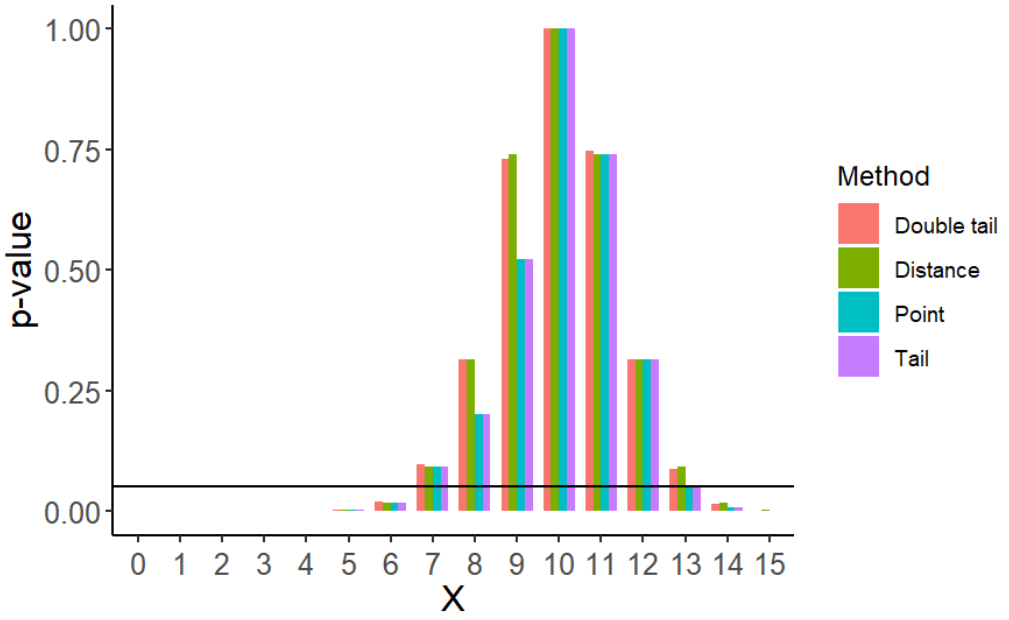


(a) P-value

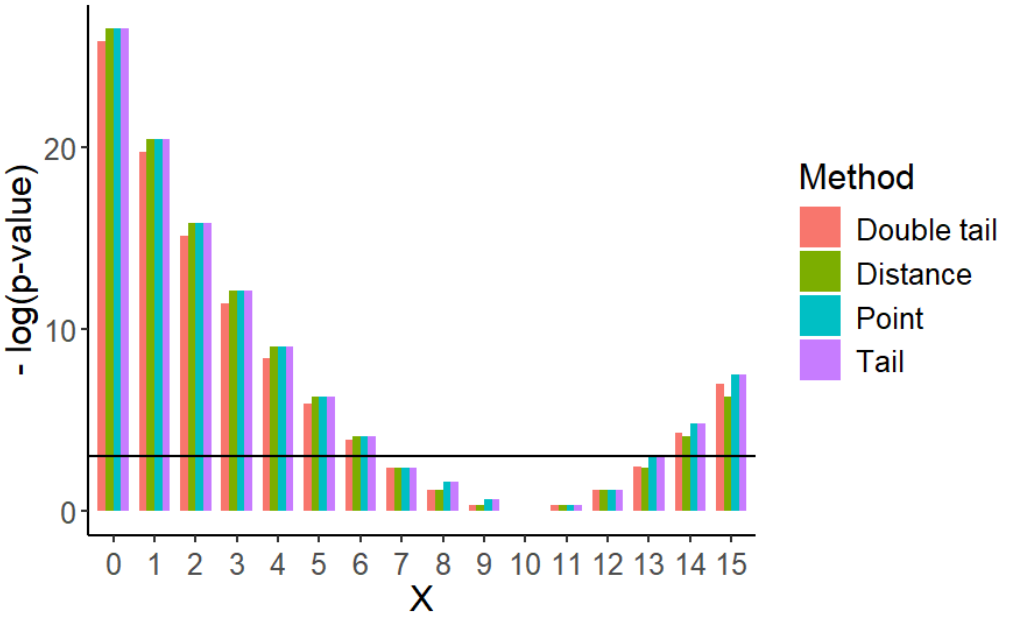


(b) - log P-value

Figure D.1.15: P-values as functions of realized values **X** for Fishers exact test with $n = 30/15,\ c = 5$. All methods

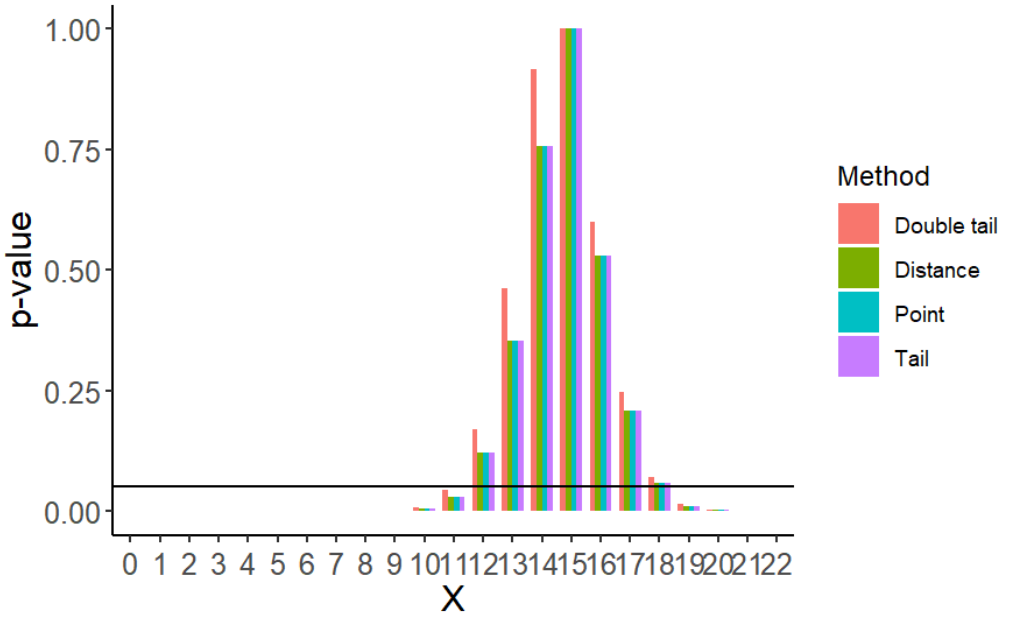


(a) P-value

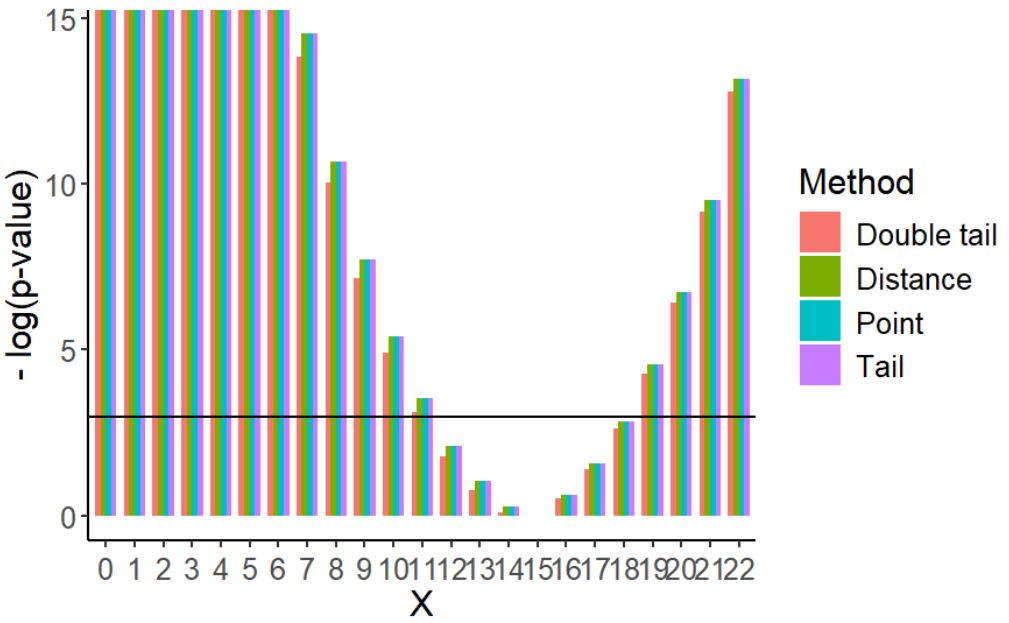


(b) - log P-value

Figure D.1.16: P-values as functions of realized values **X** for Fishers exact test with $n = 30/15,\ c = 15$. All methods

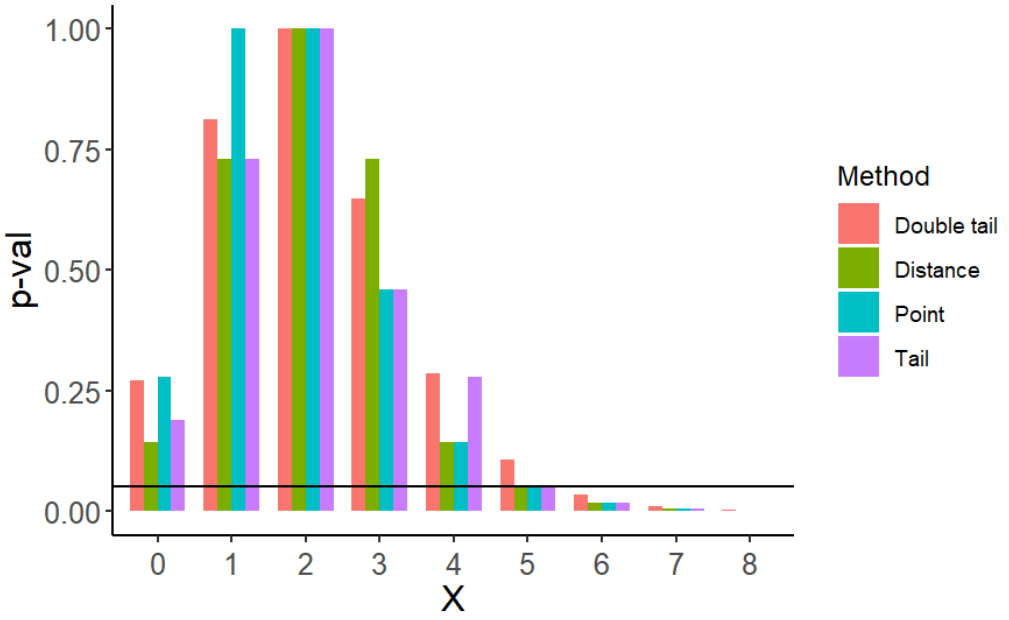


(a) P-value

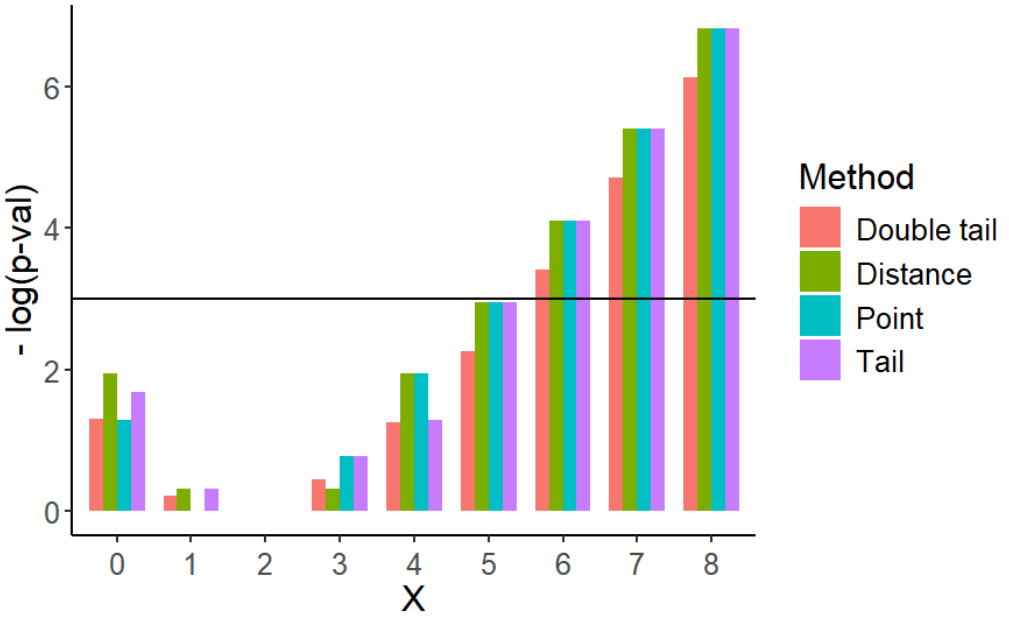


(b) - log P-value

Figure D.1.17: P-values as functions of realized values **X** for Fishers exact test with $n = 30/15,\ c = 22$. All methods

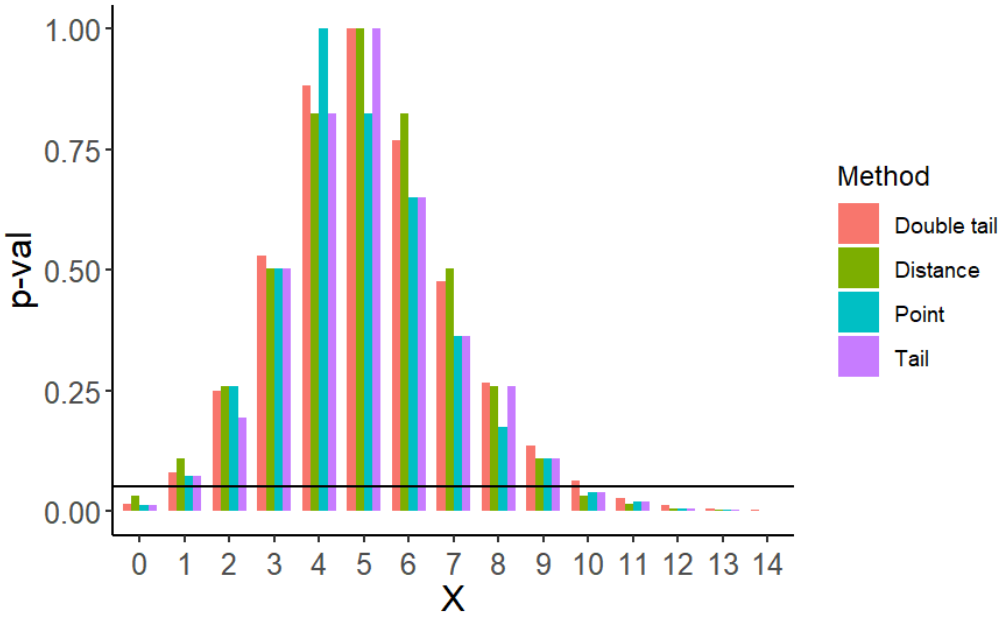


(a) P-value

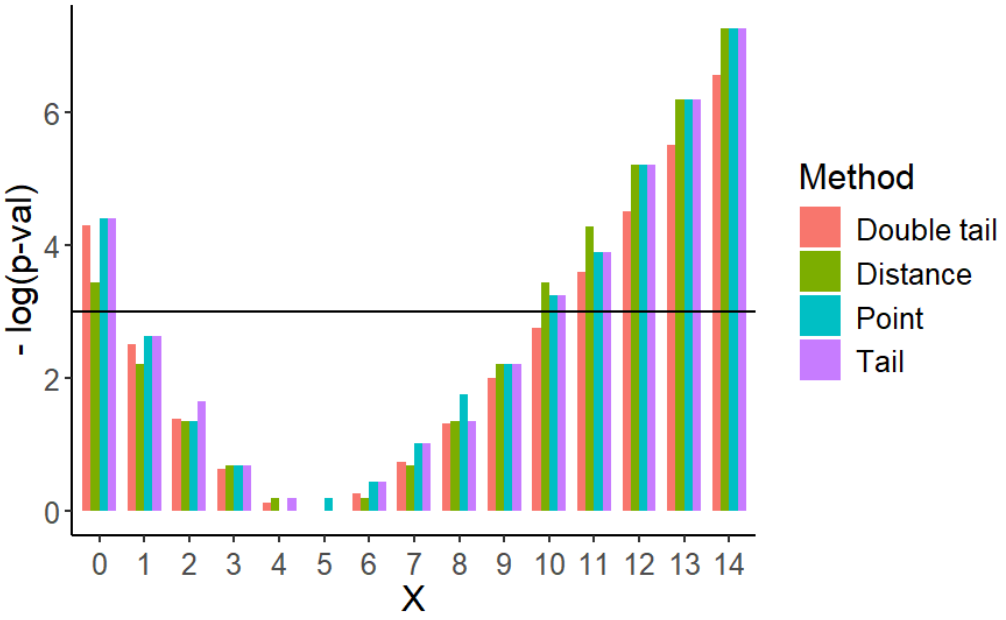


(b) - log P-value

Figure D.1.18: P-values as functions of realized values **X** for a Poisson dsitribution with $\mu_0 = 2$. All methods

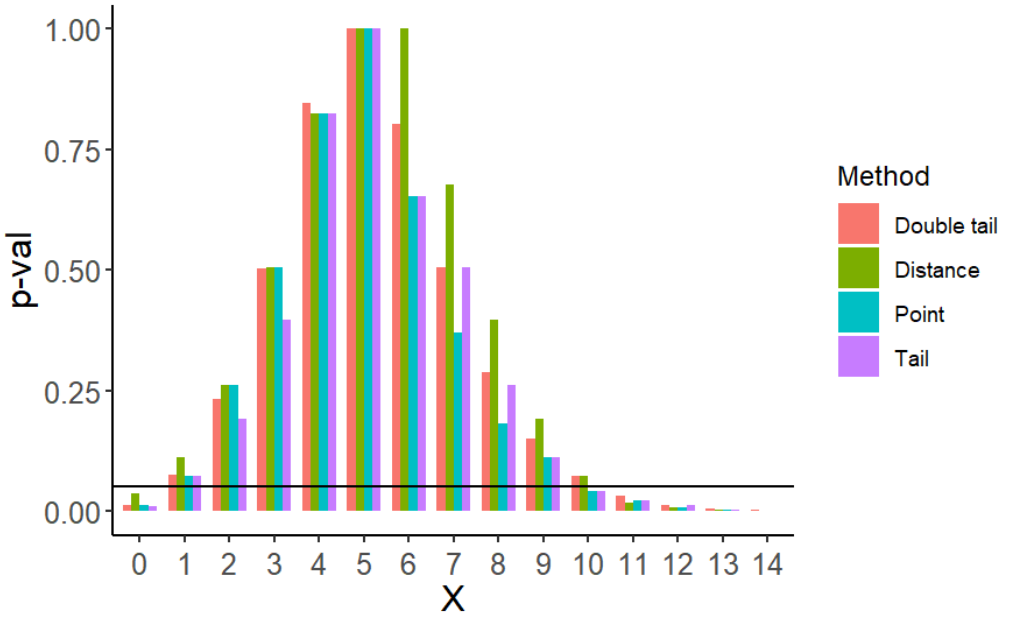


(a) P-value

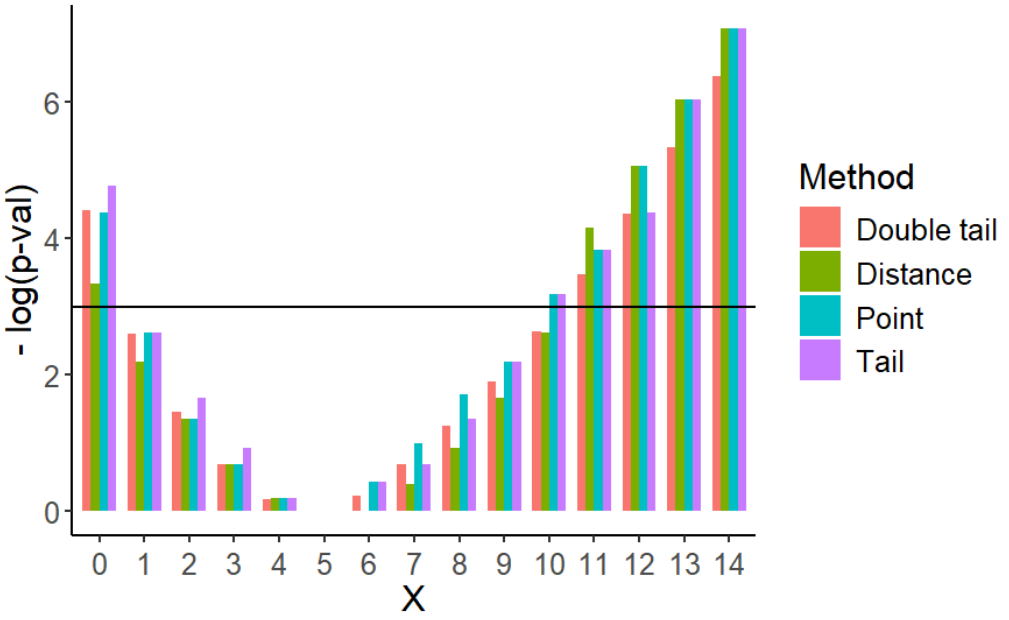


(b) - log P-value

Figure D.1.19: P-values as functions of realized values **X** for a Poisson dsitribution with $\mu_0 = 5$. All methods

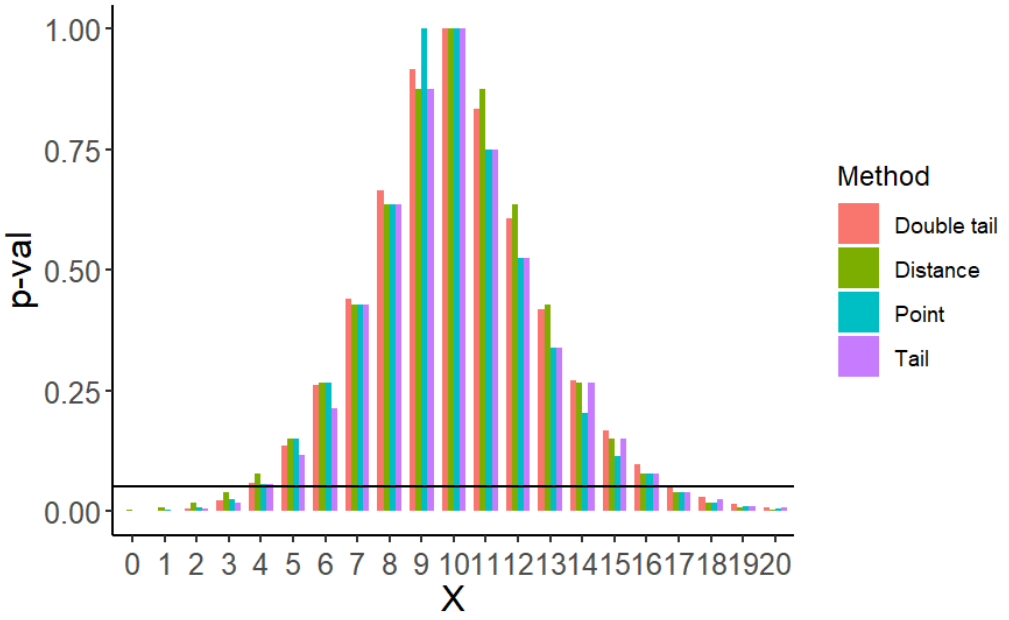


(a) P-value

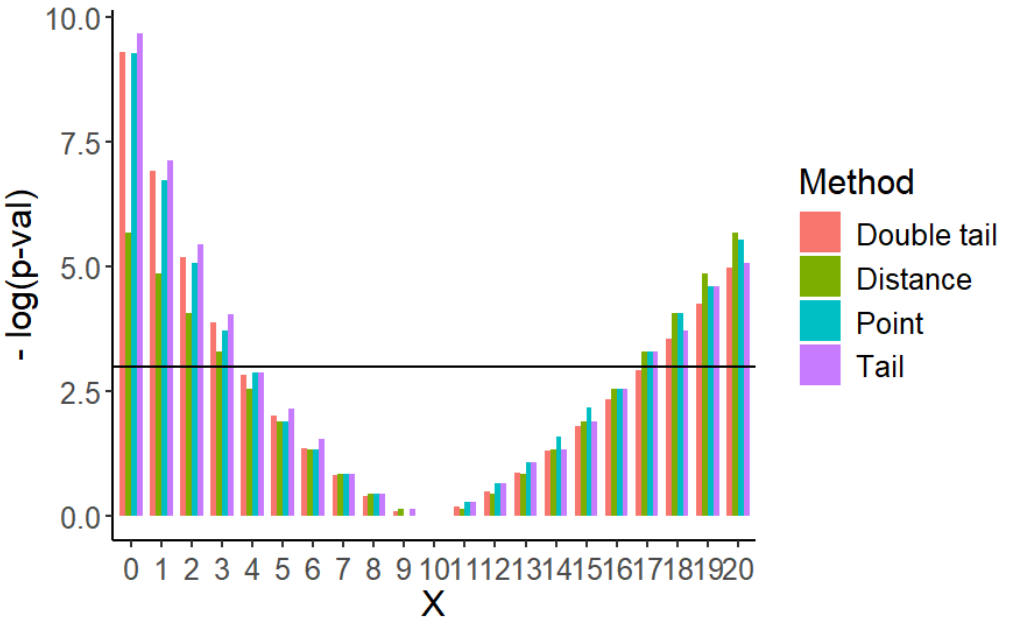


(b) - log P-value

Figure D.1.20: P-values as functions of realized values **X** for a Poisson dsitribution with $\mu_0 = 5.1$. All methods

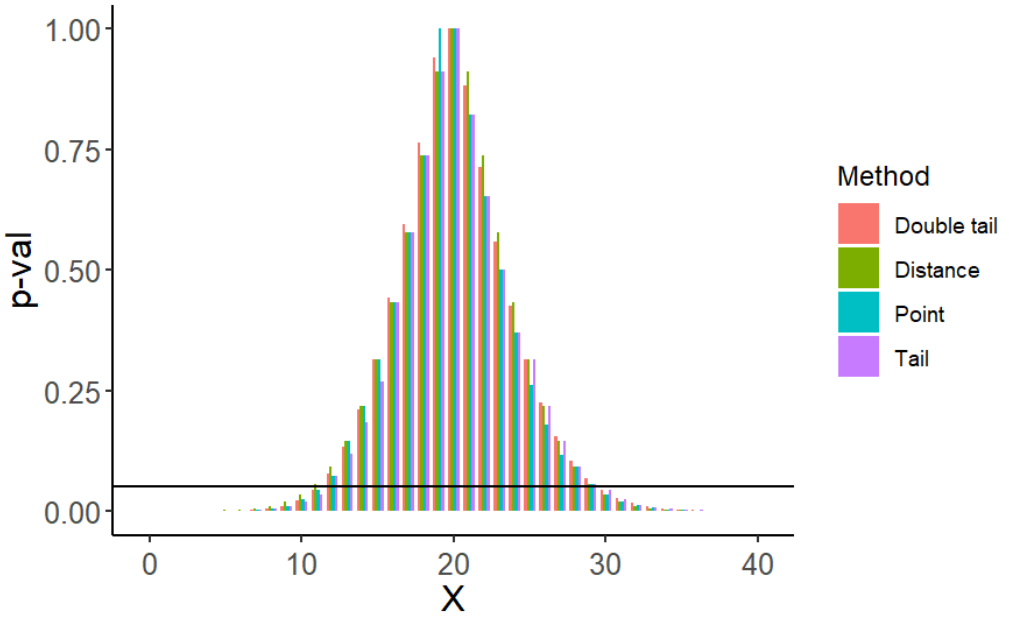


(a) P-value

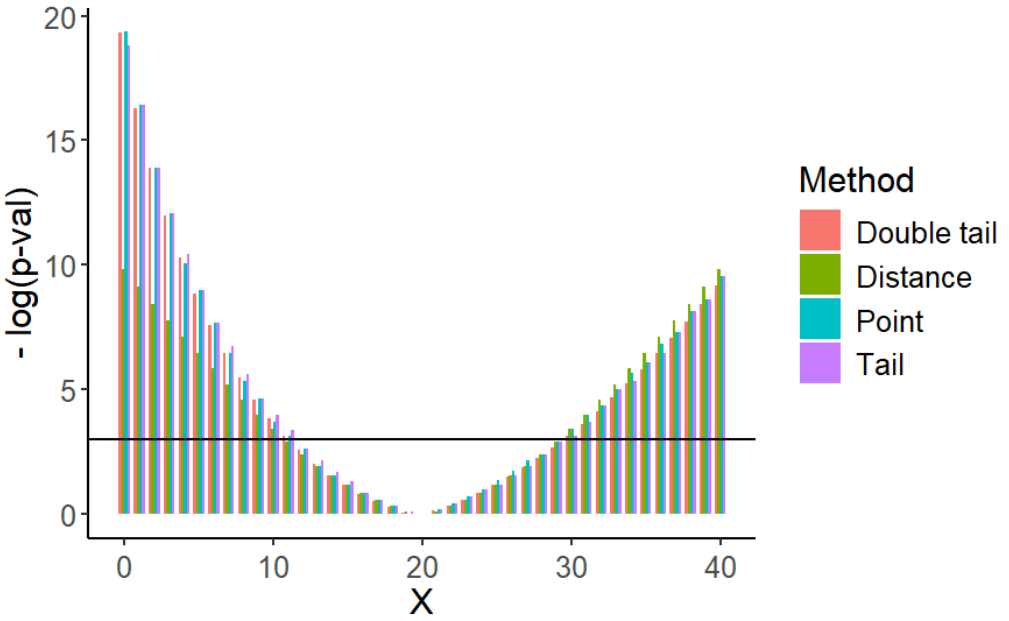


(b) - log P-value

Figure D.1.21: P-values as functions of realized values **X** for a Poisson dsitribution with $\mu_0 = 10$. All methods



(a) P-value



(b) - log P-value

Figure D.1.22: P-values as functions of realized values **X** for a Poisson dsitribution with $\mu_0 = 20$. All methods

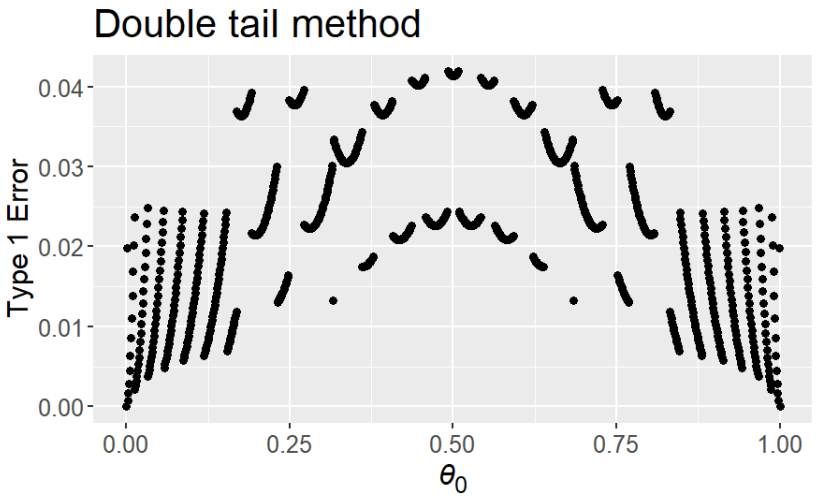


Figure D.2.1: Type I Error for different $H_0 : \theta = \theta_0$ where θ is the binomial parameter. Double Tail method

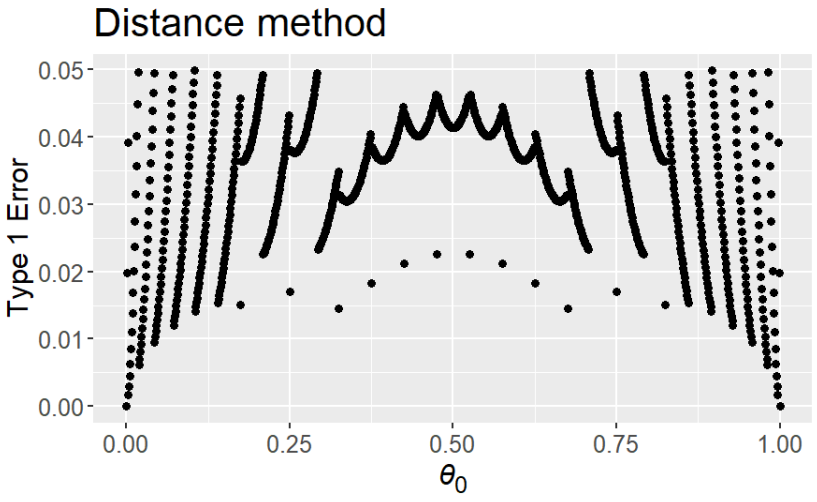


Figure D.2.2: Type I Error for different $H_0 : \theta = \theta_0$ where θ is the binomial parameter. Distance method

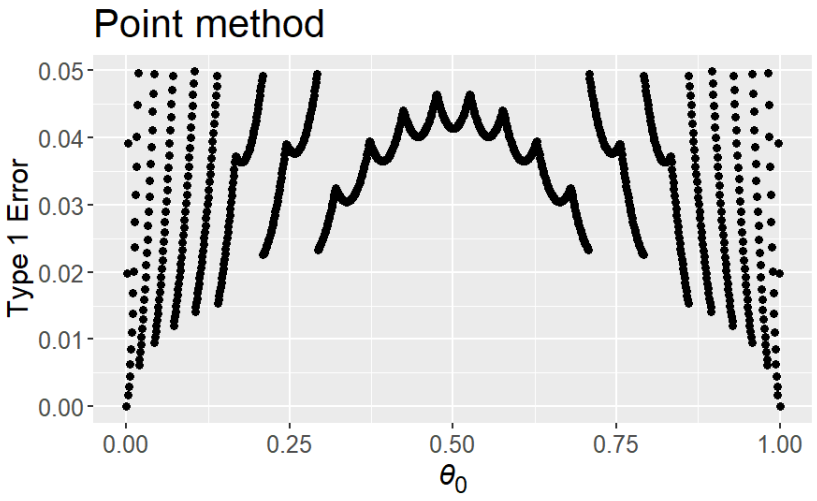


Figure D.2.3: Type I Error for different $H_0 : \theta = \theta_0$ where θ is the binomial parameter. Point method

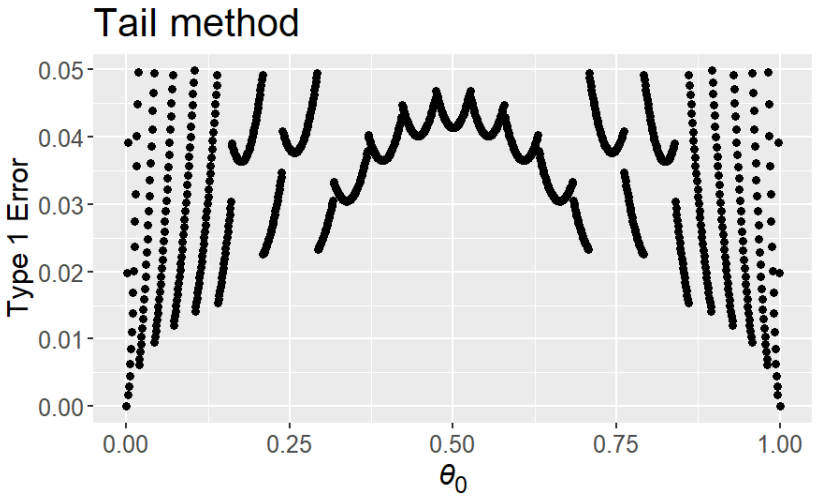


Figure D.2.4: Type I Error for different $H_0 : \theta = \theta_0$ where θ is the binomial parameter. Tail method

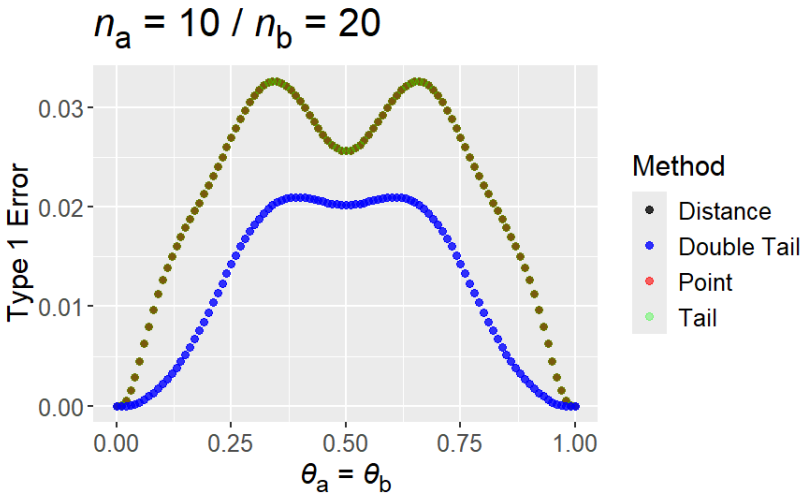


Figure D.2.5: Type I Error for a Fischer's exact test with $n = 10/20$. All methods

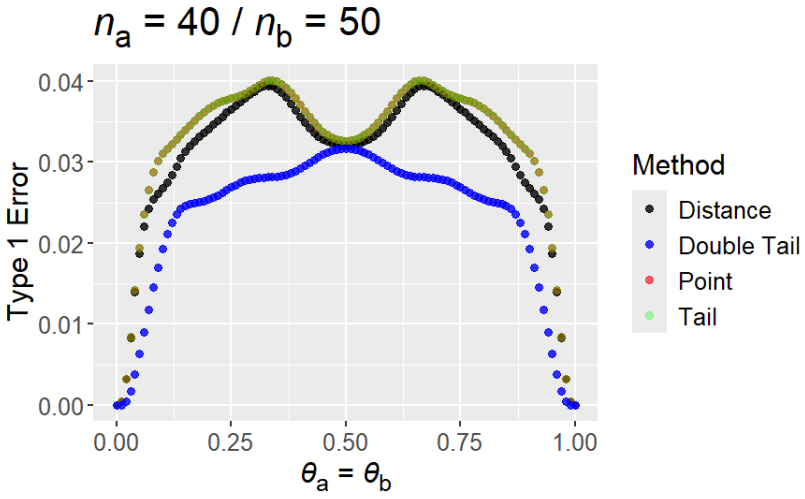


Figure D.2.6: Type I Error for a Fischer's exact test with $n = 10/20$

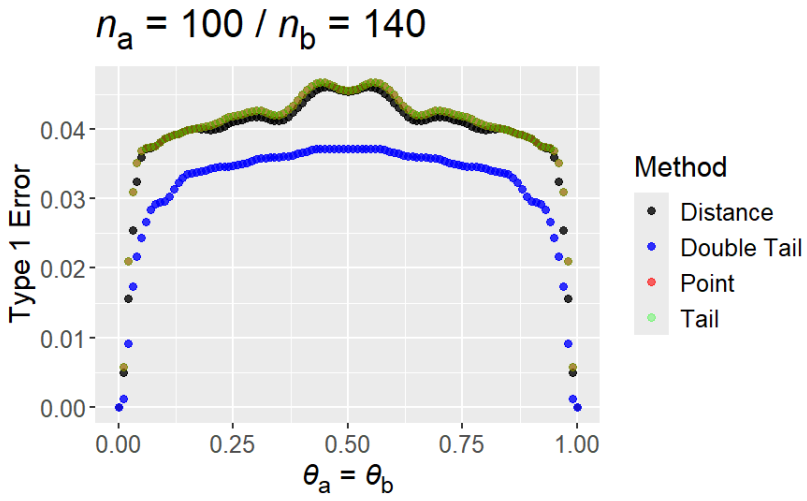


Figure D.2.7: Type I Error for a Fischer's exact test with $n = 100/140$

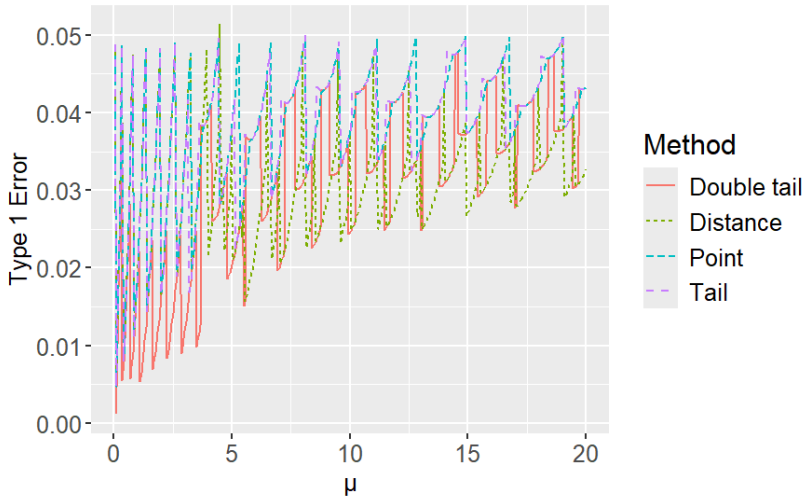


Figure D.2.8: Type I Error for the all methods. Poisson experiment.

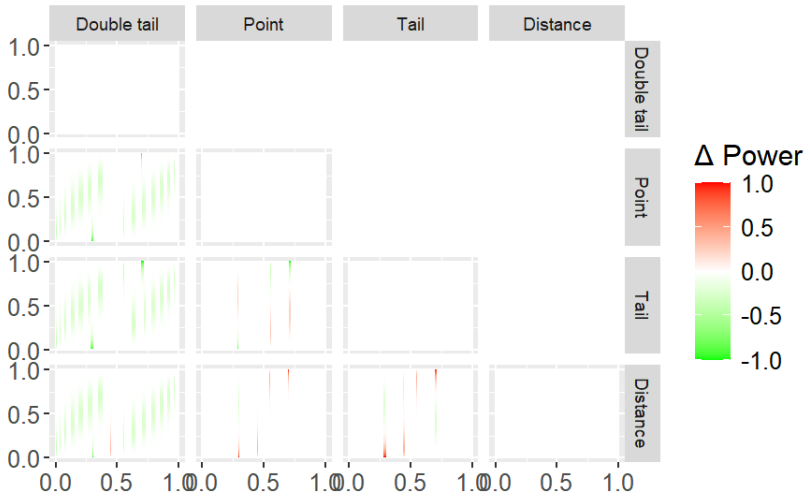


Figure D.3.1: Power comparison between all methods for varying θ_0 and θ in a test for the probability parameter in a binomial dsitribution

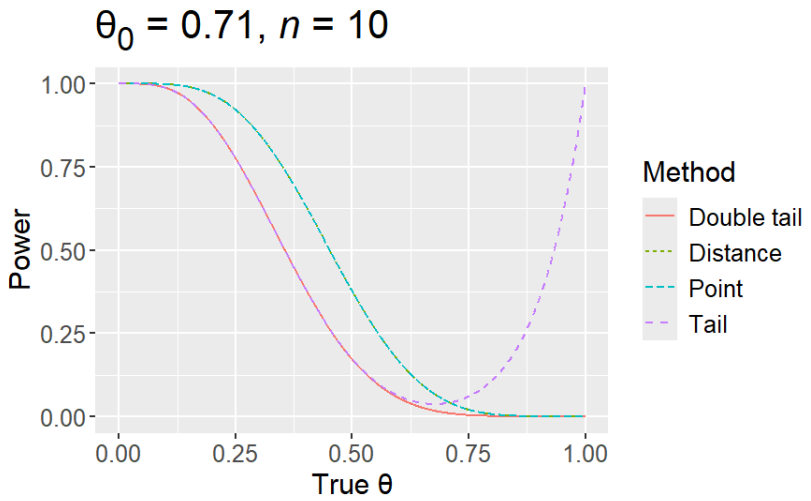


Figure D.3.2: Power function for all methods for the interesting case of $\theta_0 = 0.71$

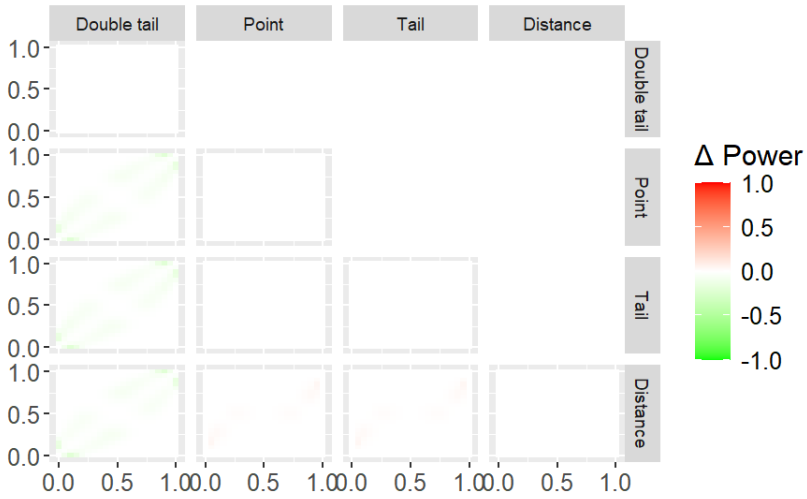


Figure D.3.3: Power comparison between all methods for varying p_1 and p_2 in a Fisher's exact test

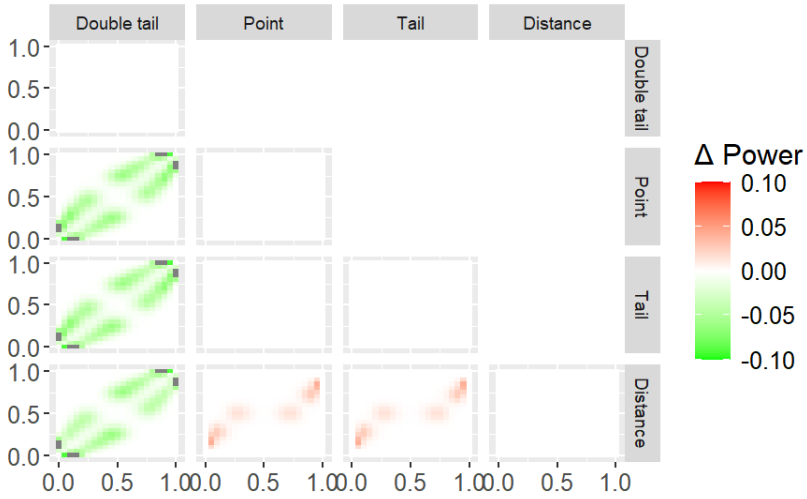
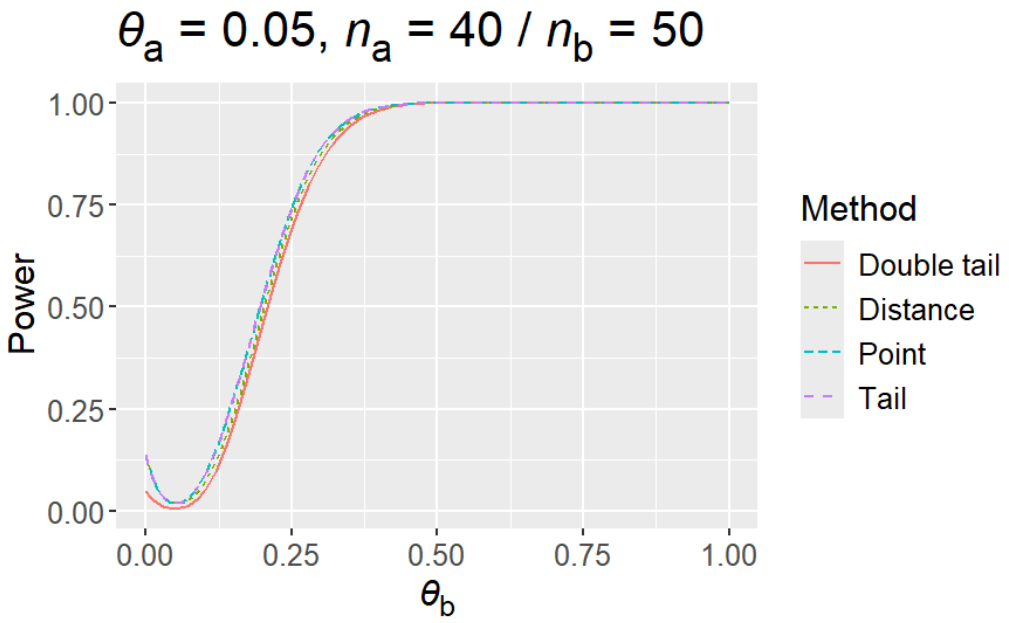
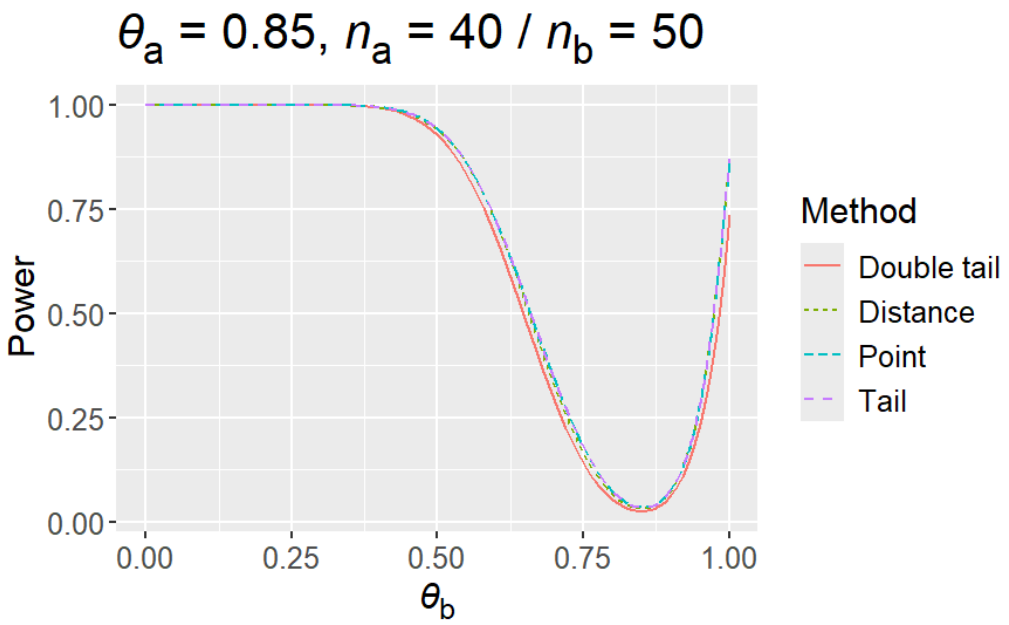


Figure D.3.4: higher contrast power comparison between all methods for varying p_1 and p_2 in a Fisher's exact test. Black areas indicate that the difference is out of bounds



(a) $p_1 = 0.05$



(b) $p_1 = 0.85$

Figure D.3.5: Power functions for some interesting values of p_1

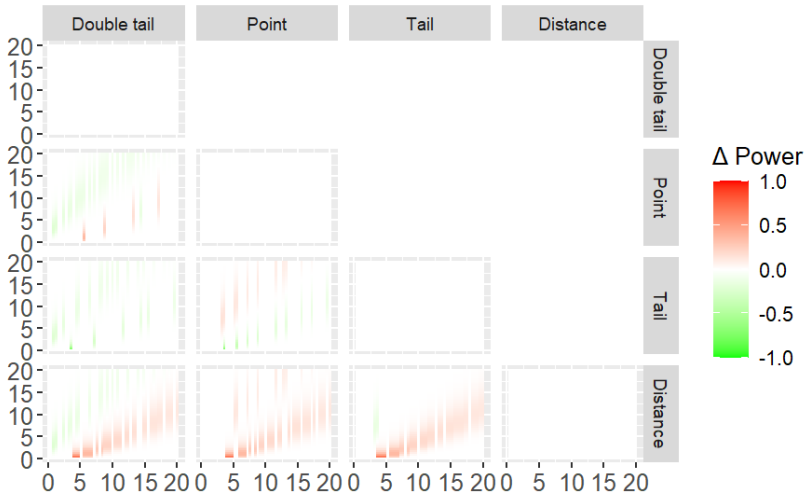


Figure D.3.6: Power matrix of all methods compared

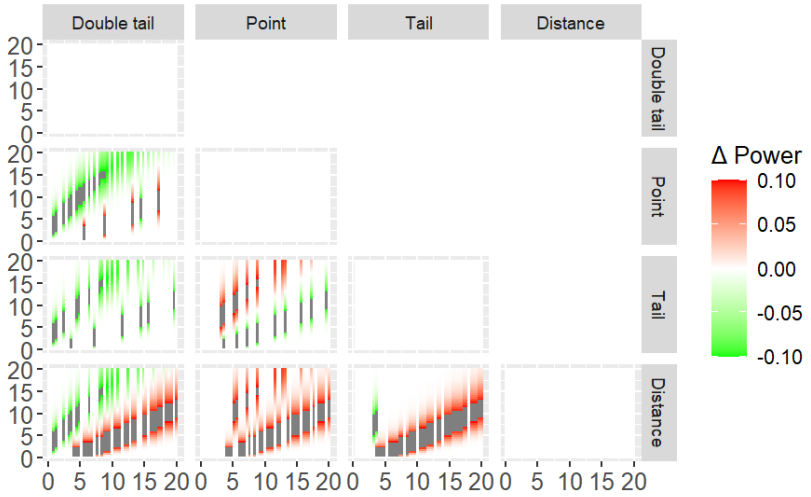


Figure D.3.7: Power matrix of all methods compared with contrast, values ranging from ± 0.1. Black areas means that the difference in power is out of bounds.

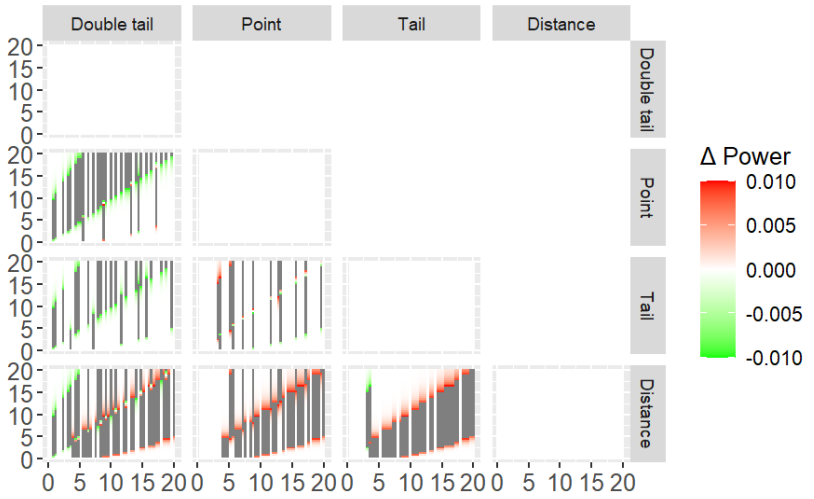


Figure D.3.8: Power matrix of all methods compared with even more contrast, values ranging from ±0.01. Black areas means that the difference in power is out of bounds.

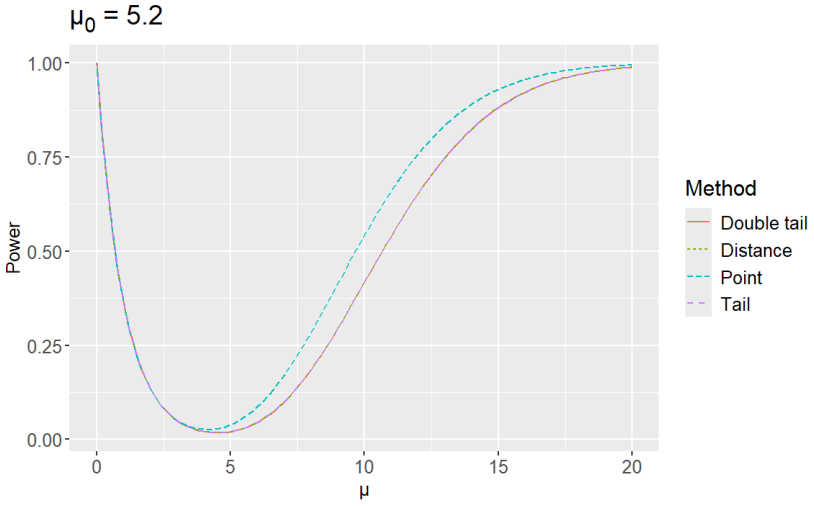


Figure D.3.9: Power function for all methods for $\mu_0 = 5.2$

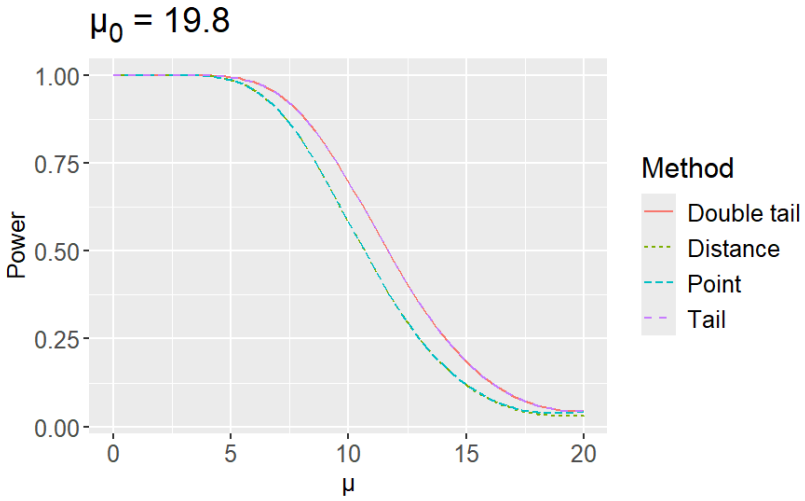


Figure D.3.10: Power function for all methods for $\mu_0 = 19.8$

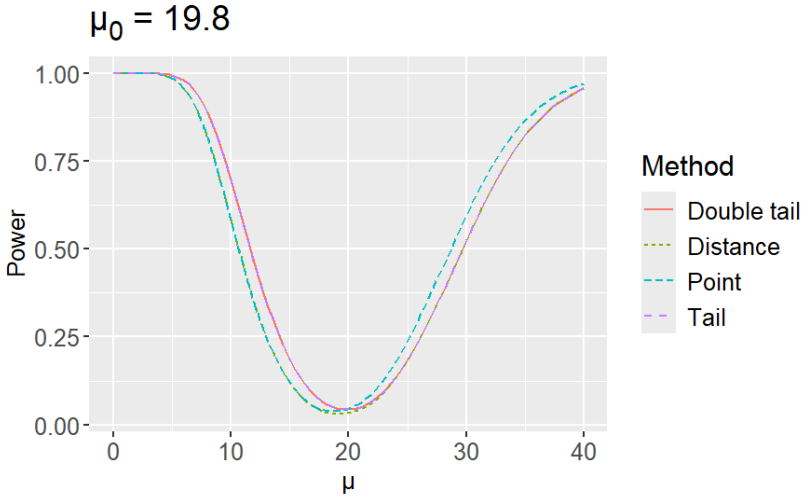


Figure D.3.11: Power function for all methods for $\mu_0 = 19.8$