NTNU
Kunnskap for en bedre verden

Department of Mathematical Sciences

MA2002 - Bachelor Thesis

Two-sided p-values for a non-symmetric distribution of test statistics

Author:
Mathias Dåsvand

May, 2022

Table of Contents

List of Figures i

List of Tables ii

1 Introduction 4

2 Preliminaries 5

2.1 Hypothesis testing and test statistics . . . 5

2.2 p-values . . . 6

2.3 Errors in hypothesis testing and statistical power . . . 7

2.3.1 Example of the power of a simple rejection rule . . . 7

3 Fisher's exact test 7

3.1 Fisher's exact test one-sided . . . 7

3.2 Constructing two-sided p-values for Fisher's exact test . . . 9

3.3 Simulation of power for the three methods . . . 11

3.3.1 Simulation of H_1: $p_x \neq p_y$ with 110 and 100 trials . . . 11

3.3.2 Simulation of $H_0 : p_x = p_y$ with 110 and 100 trials . . . 12

3.3.3 Simulation of H_1: $p_x \neq p_y$ with 220 and 55 trials . . . 12

3.3.4 Simulating of H_0: $p_x = p_y$ with 220 and 55 trials . . . 13

3.4 Exact calculation of power . . . 13

4 Discussion and conclusion 16

Bibliography 17

Appendix 18

A Proofs of validity of the 2-sided p-values . . . 18

A.1 The doubled one-sided p-value, p_1 . . . 18

A.2 The sum over all smaller probabilities, p_2 . . . 18

A.3 Sum of all probabilities with absolute larger deviations from the mean, p_3 . 18

List of Figures

1 The probability mass function of X . . . 9

2 The probability mass function of X of Example 2 . . . 10

3 The probability mass function of X . . . 11

4 The distribution of p-values for all 3 methods 12

List of Tables

1 Observed contingency table in Example 1 8
2 The probability mass function of X 8
3 2-sided p-values for different observed x 10
4 Number of times a p-value is strictly smaller than the other two 13
5 Simulated power for combinations of n_1, n_2, p_x, p_y and α 14
6 Exact power for combinations of n_1, n_2, p_x, p_y and α 15

Sammendrag

I denne oppgaven undersøkes ulike måter å beregne p-verdier for en tosidig hypotesetest hvor testobservatoren har en usymmetrisk sannsynlighetsfordeling. Først gis en kort introduksjon til statistiske begreper før Fishers eksakte ensidige test introduseres. Her vises det at testobservatoren X er hypergeometrisk fordelt og hvordan man da kan regne ut en p-verdi fra denne testobservatoren. Når man ønsker å teste en tosidig alternativ hypotese, så foreslår Agresti (1992) tre ulike måter å regne ut p-verdier på. Hovedfokuset i bacheloroppgaven er å beregne styrken til disse ulike p-verdiene på ulike forkastningsnivå. Styrken blir både simulert og regnet ut eksakt og blir sammenlignet mot hverandre. Til slutt vises det at disse tre p-verdiene er gyldige.

Forord

Dette arbeidet konkluderer min bacheloroppgave i statistikk. Jeg vil gjerne takke min veileder Øyvind Bakke for god veiledning gjennom hele bacheloroppgaven, og Mette Langaas for gode diskusjoner i starten av arbeidet med oppgaven. Jeg vil også takke Jon Arnt Kårstad for å ha lagd malen som jeg har brukt i oppgaven.

Summary

In this thesis we consider different methods of constructing p-values for a two-sided hypothesis test where the test statistic has an asymmetric probability distribution. First a short introduction to relevant statistical terms is given, whereafter Fisher's exact one-sided test is introduced. Here it is shown that the test statistic is hypergeometrically distributed when the null hypothesis is true, such that it is possible to calculate a one-sided p-value from this test statistic. For a two-sided hypothesis test, Agresti (1992) suggest three different methods of calculating p-values. The main focus of this thesis is to calculate the power of these three p-values in different situations. The power is both simulated and calculated exactly to compare the different methods. Finally, the p-values are also shown to be valid.

1 Introduction

The history of hypothesis testing can be said to start even as early as the formulation of what is known today as the scientific method. The scientific method in general can be said to proceed in 4 steps: Formulating a hypothesis, gathering data, testing the hypothesis based on the set of data and finally reaching a conclusion based on the test. As early as in the 11th century, Ibn al-Haytham performed experiments to falsify a hypothesis from Euclid about the function of the eyes (Rooney 2012).

Often in contemporary science one specifies two conflicting hypotheses, a null hypothesis and an alternative hypothesis. These hypotheses can be statements about a parameter in a distribution where both cannot be true simultaneously. In practical research, the test statistic generates a p-value which is used as a form of evidence to reject the null hypothesis (although often misunderstood to be the probability of the null hypothesis to be true).

Two-sided tests are often used, where the alternative hypothesis simply states that a parameter is either larger or smaller than some range of parameter values which are specified by the null hypothesis. Some journals such as the New England Journal of Medicine states in its guidelines: "Unless one-sided tests are required by study design, [...], all reported p-values should be two-sided" (*New Manuscripts* n.d.). Freedman (2008) notes that a possible reason for this guideline could be to prevent so-called post-hoc abuse where one specifies the hypothesis after the data is collected.

This further underlines a need for two-sided p-values. However, Kulinskaya (2008) notes a problem in the case where the test statistic is not symmetric: "two-sided statistical tests and p-values are well defined only when the test statistic in question has a symmetric distribution." Thus there is no commonly accepted method for how to assess the 2-sided p-value of asymmetric distributions, although different methods have been proposed for different situations (Agresti 1992).

This thesis will focus on different ways of constructing 2-sided p-values. These methods are then applied to Fisher's exact test, where the test statistic follows an asymmetric distribution when the null hypothesis is true. The p-values are shown to be valid and we compare the power for each p-value.

For data analysis and simulations we use the statistical software "R" (R Core Team 2022) with the package "ggplot2" (Wickham 2016).

2 Preliminaries

Suppose that an experiment is being conducted, which can be repeated and has a set of well-defined outcomes. Each outcome is defined as a *sample outcome* and the set of all possible outcomes is called the *sample space* S (Larsen and Marx 2018, p.17).

A *random variable* denoted X is a real-valued function which has as domain the sample space (Larsen and Marx 2018, p.121). If we denote s as a sample outcome, then we denote the probability of X having an outcome of s as $P(X = s)$. If the set of real values is countable or infinitely countable the random variable is called *discrete*, and if the the output set is uncountably infinite the random variable is called *continuous*.

If the random variable X is discrete we define a *probability mass function* to be a function of the sample space S satisfying two conditions

$$0 \leq P(X = s) \leq 1$$

$$\sum_{s \epsilon S} P(X = s) = 1$$

where s is any element of S (Larsen and Marx 2018, p.117).

Analogous to the discrete case, we define the *probability density function* of a continuous random variable Y to be a function f such that $\int_a^b f(t)dt = P(Y \in [a, b])$ and the two following conditions hold:

$$f(t) \geq 0, \; \forall \, t$$

$$\int_{-\infty}^{\infty} f(t) \, dt = 1$$

The *expected value* of a random variable X with sample space S is defined as

$$E(X) = \begin{cases} \sum_{s \in S} s \cdot P(X = s) & \text{if } X \text{ is discrete} \\ \\ \int_{-\infty}^{\infty} t \cdot f(t)dt & \text{if } X \text{ is continuous} \end{cases}$$

where the sum is over all $s \in S$.

We call the random variables $X_1, \ldots, X_n$ a *random sample* of size n if these random variables are independent and identically distributed (Casella and Berger 2002, p. 207). This means that $X_1 \ldots X_n$ all have the same probability density function or probability mass function.

A *parameter* is a numerical characteristic of a population or a model (Everitt and Skrondal 2010). Informally this means that the parameter gives information about the distribution of the population. In the following theoretical section, we use the symbol θ to refer to a population parameter.

2.1 Hypothesis testing and test statistics

A hypothesis is a statement about a population parameter (Casella and Berger 2002, p.373). This can for example be the mean of a population.

We denote two conflicting hypotheses by the *null hypothesis* and the *alternative hypothesis* by H_0 and H_1, respectively (Casella and Berger 2002, p.373). We denote by Θ the set of possible values of a parameter θ. Then we can state H_0 and H_1 as $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in (\Theta \cap \Theta_0^C)$ where C denotes the complementary set.

Three situations are common in defining the sets corresponding to H_0 and H_1. Either H_0 specifies that θ is either (1) less than, (2) equal to, or (3) larger than a specified point. Thus the three situations can be expressed mathematically in this way:

$$H_0 : \theta \leq \theta_0, \quad H_1 : \theta > \theta_0 \tag{1}$$

$$H_0 : \theta = \theta_0, \quad H_1 : \theta \neq \theta_0 \tag{2}$$

$$H_0 : \theta \geq \theta_0, \quad H_1 : \theta < \theta_0 \tag{3}$$

When we conduct a *hypothesis test* we are specifying a rule for which sample values H_0 will be rejected in favour of H_1 (Casella and Berger 2002, p.374). Note that since the hypotheses are complementary, acceptance of one implies rejection of the other.

We can define a *test statistic* to be a function of the sample that gives evidence either for or against the alternative hypothesis H_1 (Casella and Berger 2002, p.374). A test statistic will be denoted $T(\mathbf{X})$ where $\mathbf{X}$ is the sample $(X_1,\ldots,X_n)$.

Therefore a test statistic is the main tool to determine when H_0 is rejected. Note that there often exists multiple possible test statistics that is possible to use in the same situation (Bickel and Doksum 2002, p.223).

2.2 p-values

A p-value is a value in the interval $[0, 1]$ where small values gives evidence for rejecting H_0 for all possible samples (Casella and Berger 2002, p.397). Since the p-value is a function of the sample, the p-value itself is also a test statistic. Thus we can denote the p-value as $p(\mathbf{X})$ where $\mathbf{X}$ is the sample as in the case of the test statistic.

In the case of (2), where H_0 restricts itself to a single point and we use a test statistic, it is usual to define the p-value to be the probability of obtaining a certain value or more extreme in the particular test statistic *given H_0 is true* (Larsen and Marx 2018, p.351). When the distribution of a test statistic is asymmetric it is however not always clear how to define what constitutes ”more extreme”.

The p-value does not by itself specify when H_0 is rejected. In practical research one defines a *significance level* denoted α, where if the p-value is below the significance level, H_0 is rejected. The significance level is most often set at 0.05, but also 0.10 and 0.01 are common significance levels. If the distribution of a test statistic is known when H_0 is true, it is possible to specify for which samples a test statistic will have a p-value lower than the significance level. This sample region is denoted the *critical region* of a test statistic where H_0 is rejected (Larsen and Marx 2018, p.348).

Because of the general definition of a p-value, we also introduce the concept of a *valid* p-value. A p-value is said to be valid if the inequality

$$P(p(\mathbf{X}) \leq \alpha) \leq \alpha \tag{4}$$

holds for $0 \leq \alpha \leq 1$ and all θ which are elements of Θ_0 (Casella and Berger 2002, p.397). Informally, this means that the probability of the p-value being lower than α has to be lower than α for all θ_0 in order to be valid. If $P(p(\mathbf{X}) \leq \alpha) = \alpha$ for $\alpha \in [0, 1]$, then the valid p-value is called *exact*.

The most common way of defining a valid p-value is the following theorem.

Theorem 1. *If $W(\boldsymbol{X})$ is a test statistic where larger values of $W(\boldsymbol{X})$ is evidence for H_1, then*

$$p(\boldsymbol{x}) = \sup_{\theta \in \Theta_0} P(W(\boldsymbol{X}) \geq W(\boldsymbol{x}))$$

defines a valid p-value. (Casella and Berger 2002, p.397)

In the conditional setting that we will be considering, the null distribution of $W(\mathbf{X})$ does not depend on $\theta \in \Theta_0$, in which case the supremum is superfluous.

2.3 Errors in hypothesis testing and statistical power

If we use p-values to perform a hypothesis test, which means we reject H_0 when the p-value is below a given α, then a small p-value provides evidence against H_0. However this evidence is not definitive unless the p-value is equal to 0. This means that when we reject one of the hypotheses based on the p-value or any other test statistic, there is a chance that we are making an error. We denote rejecting H_0 when H_0 is true a type I error, and we denote accepting H_0 when H_0 is false a type II error (Casella and Berger 2002). We can express the four possibilities in a 2×2 table:

	H_0 Rejected	H_1 Rejected
H_0 True	Type I error	Correct decision
H_1 True	Correct decision	Type II error

If we specify a significance level, it is useful to know the probability of a type I or type II error will be made. For this purpose we define a *power function* $\beta(\theta) = P(\mathbf{X} \in R)$ where R denotes the critical region. If $\theta \in \Theta_0$ then the probability of the sample being in the critical region is the chance of committing a type I error. If $\theta \in \Theta \cap \Theta_0^C$, then the probability of the sample being in the critical region is 1 minus the probability of a type II error.

We want the probability of committing either type of error as low as possible. This means that we want $\beta(\theta)$ for $\theta \in \Theta_0$ as close to 0 as possible, and $\beta(\theta)$ for $\theta \in \Theta_0^C$ as close to 1 as possible.

2.3.1 Example of the power of a simple rejection rule

Let Y be Poisson distributed with parameter λ. We specify H_0: $\lambda \leq 2$ and H_1: $\lambda > 2$, and assume we have a single observation from the distribution. If we reject H_0 when the observation is equal to 4 or above, then we can construct our power function. For any $\lambda \in \Theta_0$, $P(Y \geq 4) \leq 1 - \sum_{i=0}^{3} \frac{2^i \exp(-2)}{i!} = 0.143$ which means that $0 \leq \beta(\lambda) \leq 0.143$ for a λ between 0 and 2. For a λ over 2, $P(Y \geq 4) \geq 1 - \sum_{i=0}^{3} \frac{2^i \exp(-2)}{i!} = 0.143$. This implies that $0.143 \leq \beta(\lambda) \leq 1$ for a λ greater than 2. Thus from our power function we know that the probability of committing a type I error when $\lambda \leq 2$ is between 0 and 0.143, and the probability of committing a type II error when $\lambda > 2$ is between 0 and 0.857.

Using different test statistics for the same hypothesis test yields different power functions. Typically using any test statistic in a hypothesis test, there is a tradeoff between the probability of committing a type I and type II error. If the probability a type I error is low, then the probability of a type II error is usually high and vice versa. With this in mind, for $0 \leq \alpha \leq 1$, we call a test statistic a *size α test* if $\sup_{\theta \epsilon \Theta_0} \beta(\theta) = \alpha$ and we call it a *level* α test if $\sup_{\theta \epsilon \Theta_0} \beta(\theta) \leq \alpha$ (Casella and Berger 2002, p.385).

3 Fisher's exact test

3.1 Fisher's exact test one-sided

Let X and Y be independent binomially distributed with parameters p_x, p_y and n_1, n_2, respectively. We wish to test the null hypothesis H_0: $p_x = p_y$ against the alternative hypothesis H_1: $p_x > p_y$. We can represent a single realization in a 2×2 contingency table:

x	y	$x + y$
$n_1 - x$	$n_2 - y$	$n_1 + n_2 - x - y$
n_1	n_2	$n_1 + n_2$

Here x and y are the respective number of successes and n_1, n_2 are the respective fixed trials in each experiment. For brevity, $x + y$ is denoted as c subsequently.

Fisher's exact test is a popular test for testing whether the probability parameters are equal in two different binomial distributions with fixed n_1, n_2. Fisher showed that if one can assume the number of trials for both variables to be known as well as the combined amount of successes, then the probability of successes in one of the variables follows a hypergeometric distribution when the null hypothesis is true (Agresti 1992, p.134). Let X and Y be binomially distributed as before. This can be shown by the definition of conditional probability:

$$P(X = x \mid X + Y = c) = \frac{P(X = x \cap Y + X = c)}{P(X + Y = c)}$$

Now the intersection determines that $Y + X = c$ is the same as $Y = c - x$ and $X + Y$ follows a binomial distribution with parameters (n_1+n_2, p_x) when H_0 is true (Larsen and Marx 2018, p.177). Thus when H_0 is true the expression simplifies to

$$P(X = x|X + Y = c) = \frac{\binom{n_1}{x} p_x^x (1 - p_x)^{n_1 - x} \binom{n_2}{c-x} p_x^{c-x} (1 - p_x)^{n_2 - c + x}}{\binom{n_1+n_2}{x+y} p_x^{x+y} (1 - p_x)^{n_1 + n_2 - x - y}} = \frac{\binom{n_1}{x}\binom{n_2}{c-x}}{\binom{n_1+n_2}{c}} \tag{5}$$

which is the probability mass function of a hypergeometric distribution with parameters ($n_1 + n_2$, n_1, c). Continuing in this section, we will refer to $P(X = x \mid X + Y = c)$ as $f(x \mid c)$ (as H_0 is assumed to be true in the following calculations).

Example 1 Consider the observation of the following table:

15	7	22
4	4	8
19	11	30

Table 1: Observations in Example 1

In Table 2, the probability mass function $f(x \mid 22)$ as in (5) is shown for any observed x. Thus under H_0 the probability of obtaining Table 1 is 0.218.

Since the conditional distribution of X is known when H_0 is true, using X itself as a test statistic is sensible. The sum of the probabilities of obtaining an x that is equal to or larger than X when H_0 is true is then a valid p-value. This p-value can be expressed mathematically by

$$p(\mathbf{x}) = \sum_{j=x}^{\min(n_1,c)} f(j \mid c) \tag{6}$$

for any sample point $\mathbf{x}$.

Consider testing $H_0 : p_x = p_y$ against $H_1 : p_x > p_y$ conditional on observing $c = 22$. For this

x	$f(x)$
11	0.012
12	0.094
13	0.254
14	0.328
15	0.218
16	0.076
17	0.001
18	0.000
19	0.000

Table 2: The probability mass function of X for every possible x in Example 1

example the p-value becomes:

$$p(\mathbf{x}) = \sum_{j=15}^{\min(19,22)} f(j \mid 22) = \sum_{j=15}^{19} f(j \mid 22) = 0.31$$

(see Figure 1 and Table 2).

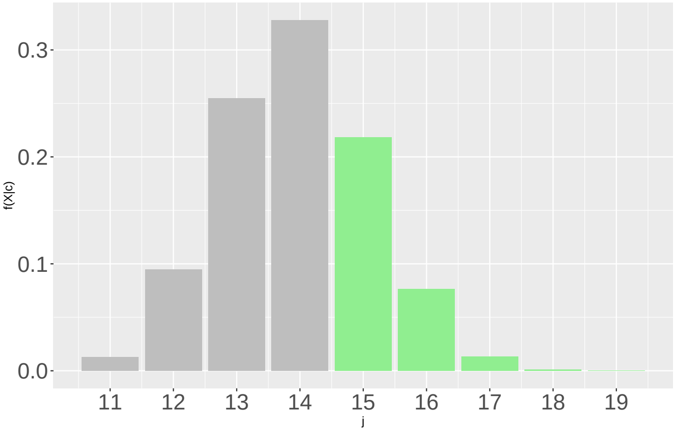

Figure 1: The probability mass function of X of Example 1. The green bars indicate the sum of $f(j \mid c)$ for $j = 15, \ldots, 19$ that gives the p-value for Fisher's 1-sided exact test.

3.2 Constructing two-sided p-values for Fisher's exact test

In the preceding section we tested H_0: $p_x = p_y$ against H_1: $p_x > p_y$ and the p-value was the conditional probability of obtaining the observed value or more extreme when H_0 was true. This one-sided p-value can be extended to a two-sided p-value in the case where $H_0 : p_x = p_y$ is tested against $H_1 : p_x \neq p_y$.

We look at three possible ways of constructing valid p-values for Fisher's exact test, that is, when we are able to condition on c. The three p-values are shown to be valid in Appendix A, and will henceforth be denoted p_1, p_2 and p_3. Agresti (1992) notes the three possible ways to extend the conditional one-sided p-value to the two-sided case:

p_1) **Doubling the smallest one-sided conditional p-value**. Let P_r denote the p-value for the one-sided test H_0: $p_x = p_y$ against H_1: $p_x > p_y$, and let P_l denote the test for H_1: $p_x < p_y$ against $H_0 : p_x = p_y$. P_l is calculated as in equation 6. P_r is also calculated as in (6) when substituting the distribution of X with Y and using $Y = c - x$ as the sample point. Then the two-sided p-value is defined as

$$p_1(\mathbf{X}) = 2 \min\left(P_l, P_r, \frac{1}{2}\right). \tag{7}$$

p_2) **Using $f(x \mid c)$ as a test statistic.** For this method we sum all probabilities of outcomes having probabilities less than or equal to the probability of the observed x, that is,

$$p_2(\mathbf{X}) = \sum_j f(j|c) \tag{8}$$

where the sum is over all j such that $f(j \mid c) \leq f(x \mid c)$.

p_3) **Using the absolute value of the deviation of the sample point from the expected value as a test statistic.** This means summing all distributions further from the expected value

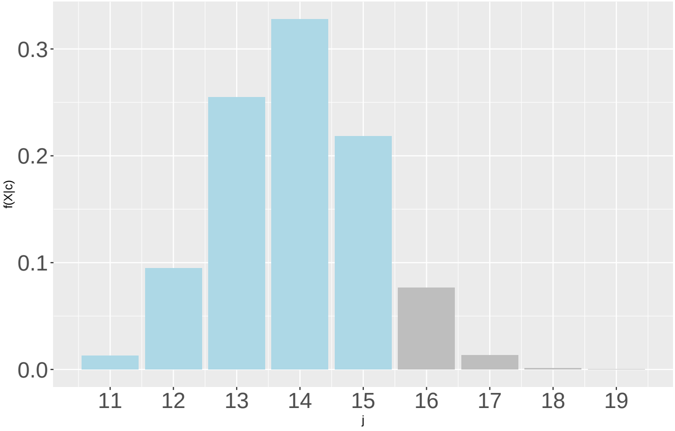

Figure 2: The probability mass function of X of Example 1. The blue bars indicate the sum of $f(j \mid c)$ for $j = 11, \ldots, 15$ that gives the value of P_l. Since $P_l > P_r$ (see Figure 1), $p_1(15 \mid 22) = 2P_l = 0.62$.

x	$p_1(x \mid 22)$	$p_2(x \mid 22)$	$p_3(x \mid 22)$
11	0.026	0.014	0.028
12	0.215	0.199	0.199
13	0.725	0.672	0.672
14	1	1	1
15	0.619	0.417	0.417
16	0.182	0.104	0.104
17	0.029	0.027	0.015
18	0.002	0.001	0.001
19	<0.001	<0.001	<0.001

Table 3: 2-sided p-values for each method for different observed $\mathbf{x}$ in Example 1.

than the observed x. The expected value of X is given by $c\, n_1/n$ when H_0 is true (Larsen and Marx 2018). This p-value can be expressed mathematically as

$$p_3(\mathbf{X}) = \sum_j f(j|c) \tag{9}$$

summed over all j where $|j - E(X)| \geq |x - E(X)|$.

It is shown in Appendix A that all three p-values are valid.

Example 1 (continued)

We continue to use the same example as in the previous section to calculate the two-sided p-values. In our example this gives p_1 to be 0.62, as our previous one-sided p-value is the smallest of P_r, P_l and $\frac{1}{2}$ (see Figure 1 and 2).

p_2 is the sum of all probabilities lower than or equal to the observed x. p_2 is thus 0.42 (see Figure 3).

Note that under H_0, $E(X)$ is a fixed constant as all values are given. In our example the expected value of X is $22{\cdot}19/30 = 13.93$. The observed x is equal to 15. Thus the sum is over all x which are less than or equal to 12 or greater than or equal to 15. For our example, this method gives the same p-value as in method 2 of 0.42 (see Figure 3).

Table 3 shows the possible observations of x and the respective p-values generated for each point x.

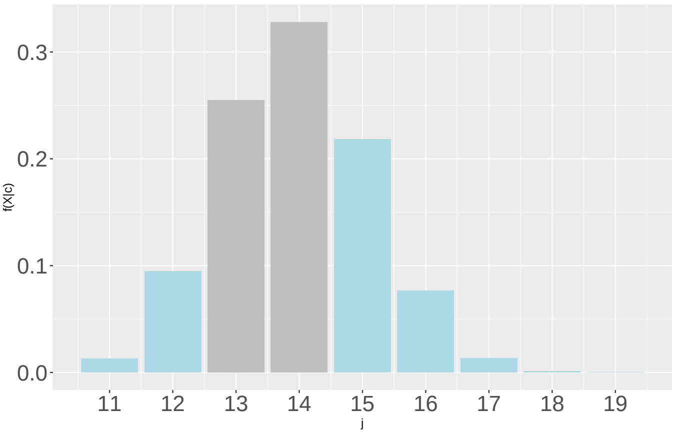

Figure 3: The probability mass function of X of Example 1. The blue bars indicate the sum of the probabilities $f(j \mid c)$ for $p_2(15 \mid 22)$ and $p_3(15 \mid 22)$ which gives a value of 0.42.

In the following sections we show that these 2-sided p-values are valid and we investigate the statistical power of the three different methods.

3.3 Simulation of power for the three methods

We wish to estimate the power by simulation for each of the three methods for various combinations of n_1, n_2, p_x, p_y and α. The results are given in Tables 3 and 5. A detailed description of four cases is given below.

We denote $\hat{\gamma}_1, \hat{\gamma}_2, \hat{\gamma}_3$ to be the estimated power for a significance level of α for the respective 2-sided p-values. The variables γ_i are all binomially distributed with an unknown probability of rejecting H_0. We provide a 95% Clopper-Pearson confidence interval for each γ_i based on each $\hat{\gamma}_i$.

3.3.1 Simulation of H_1: $p_x \neq p_y$ with 110 and 100 trials

Let X and Y be independent and binomially distributed with 110 and 100 trials respectively, and with probabilities of success $p_x = 0.2$, $p_y = 0.1$. Thus $H_1 : p_x \neq p_y$ is true, and we wish to see how often H_0 is rejected for the different methods. Here we let $\alpha = 0.05$.

Using 1000 simulations with an α of 0.05, the null hypothesis is rejected 448, 476 and 476 times, respectively (see Figure 4, first column). The respective 95% Clopper-Pearson confidence intervals are $[0.417, 0.479]$, $[0.445, 0.507]$ and $[0.445, 0.507]$.

The mean of the simulated p-values for the three methods are 0.169, 0.151 and 0.151 for the simulation, respectively. In only 30 out of 1000 simulations was one of the p-values strictly smaller than both others. Method 1 was smallest in 9 simulations, method 2 was smallest in 21 simulations, and method 3 never gave a strictly lowest p-value of the three. Method 3 and 2 gave equal p-values in 979 simulations, and in one of these cases all 3 p-values were equal. Otherwise, they were dissimilar.

In all 1000 simulations, method 2 provided p-values smaller than or equal to method 3. Thus it seems plausible that this method has more power than method 3 in this case. However, note that all differences between p_2 and p_3 did not imply any difference in the rejection of H_0 (see Table 4). In the previous section by Table 3, $p_2(17 \mid 22) > p_3(17 \mid 22)$, so the difference in power may vary to a great extent based on the number of trials in the distributions of X and Y.

In all cases where method 1 gave lower p-values than method 2, the difference between the two was smaller than 0.001. When method 2 gave smaller p-values than method 1, the difference was as large as 0.131.

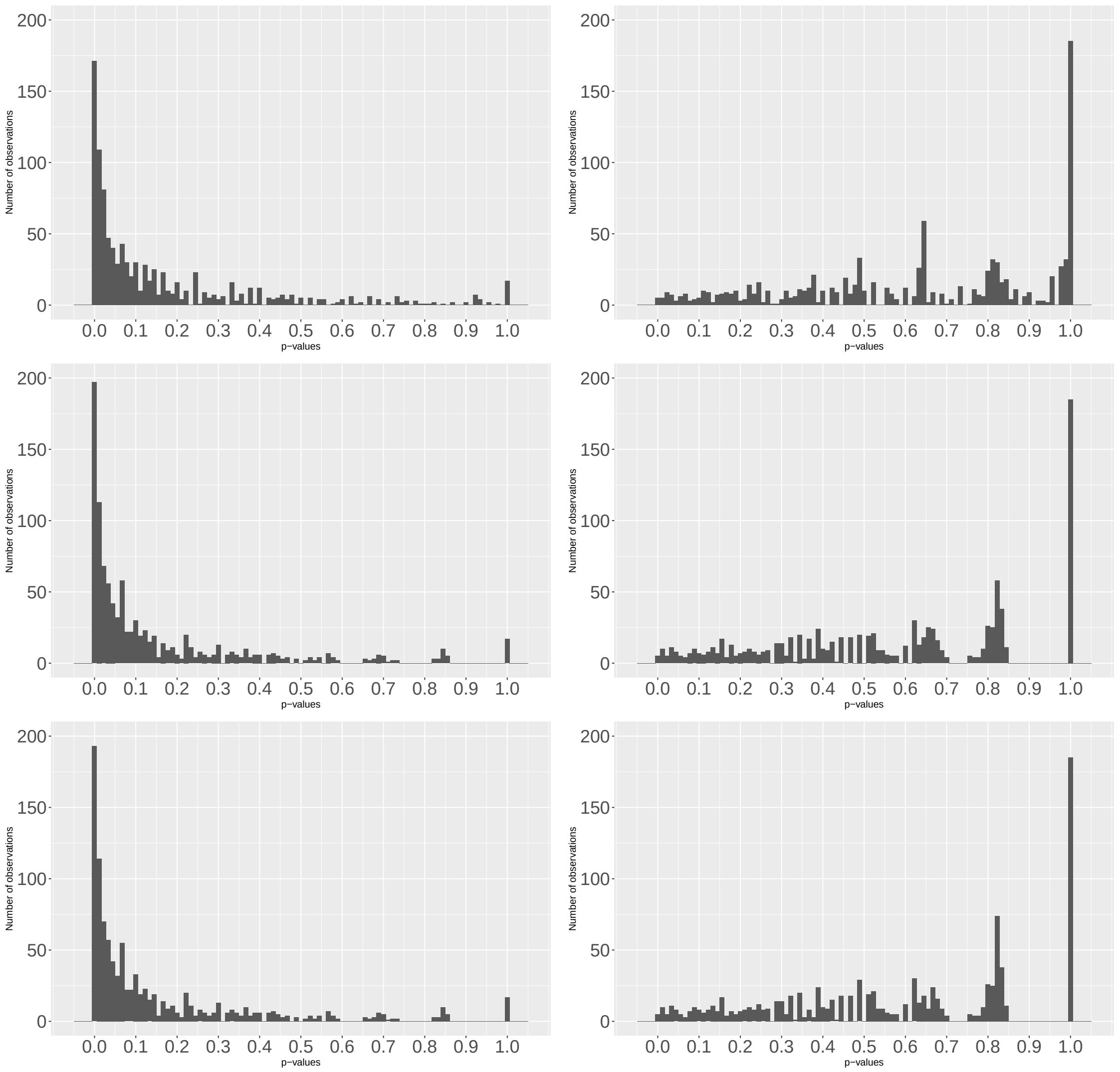

Figure 4: The distribution of the simulated p-values p_1, p_2 and p_3 when $p_x = 0.2$ $p_y = 0.1$ (first column) and $p_x = p_y = 0.1$ (second column). The first row is of p_1, the second row of p_2 and the third row is of p_3, all for the same 1000 simulations. $n_1 = 110$, $n_2 = 100$

3.3.2 Simulation of $H_0 : p_x = p_y$ with 110 and 100 trials

We again consider 1000 simulations, with the only difference being that $p_x = p_y = 0.1$ in this case. This means that we are investigating the probability of making a type I error using the different p-values and $\alpha = 0.05$ for rejection of H_0.

Of the 1000 simulations the p-values are below 0.05: 29, 39 and 39 times for each method respectively. Similar to the previous simulation, method 2 and 3 give equal p-values in 970 out of 1000 simulations, with 19 giving equal p-values for all three methods (see Figure 4, right column).

3.3.3 Simulation of H_1: $p_x \neq p_y$ with 220 and 55 trials

We saw in the previous section that the difference between the methods were not large. Out of 1000 simulations when p_x=0.2 and p_y=0.1 when $\alpha = 0.05$, H_0 was rejected 28 times more using either p_2 or p_3 instead of p_1. However simulating under H_0, p_2 and p_3 led to 10 more type I errors than p_1 using a 0.05 rule. Thus p_1 seems to be more conservative than the others, meaning less likely to reject H_0 in general.

	n_1/n_2	$p_1 < p_2, p_3$	$p_2 < p_1, p_3$	$p_3 < p_1, p_2$
H_0	220/200	11	25	0
	110/100	11	32	0
	220/55	18	224	141
	440/55	11	223	209
	300/75	20	182	120
$H_1 >$	220/200	23	111	0
	110/100	9	21	0
	220/55	133	264	42
	440/55	177	297	71
	300/75	166	353	41
$H_1 <$	220/200	0	8	62
	110/100	2	8	11
	220/55	0	133	225
	440/55	0	126	285
	300/75	0	76	340

Table 4: Number of times a p-value is strictly smaller than the other two when H_0 and H_1 is true. “H_0” denotes $p_x = p_y = 0.1$, “$H_1 >$ ” denotes $p_x = 0.2$, $p_y = 0.1$ and “$H_1 <$ ” denotes $p_x = 0.1$, $p_y = 0.2$.

To highlight the possible differences we draw 1000 observations from X and Y, both independent and binomially distributed with 220 and 55 trials and p_x=0.2, p_y=0.1. n_1 has been doubled and n_2 has been halved. The significance level is set at $\alpha = 0.05$.

Using a 0.05 rule, H_0 is rejected 350, 358 and 341 times by each respective p-value (see Table 4). The respective 95% confidence intervals are [0.320, 0.380], [0.328, 0.389] and [0.312, 0.371]. While the total number of trials has been increased, the power has decreased.

A p-value for each method was strictly smaller than the others 133, 264 and 42 times respectively. All three perform similarly, rejecting the null hypothesis roughly 35% of the time.

3.3.4 Simulating of H_0: $p_x = p_y$ with 220 and 55 trials

We consider the same example of $n_1 = 55$, $n_2 = 220$, $\alpha = 0.05$, but change p_x and p_y to be $p_x = p_y = 0.1$.

H_0 is rejected 0.031, 0.038 and 0.030 by the respective methods, with 95% Clopper-Pearson intervals [0.021, 0.044], [0.027, 0.052] and [0.020, 0.043] respectively. p_1 was strictly smallest 18 times, p_2 was strictly smallest 224 times and p_3 was strictly smallest 141 times. In 739 simulations p_2 was equal to p_3, with 23 simulations of these giving equal p-values for all three.

3.4 Exact calculation of power

It is also possible to exactly calculate the power of rejection based on each p-value. Let n_1, n_2, p_x, p_y be the number of trials and the success parameters, respectively. Then the power γ_i for each p-value p_i is the probability of obtaining any outcome in X and Y such that the p-value is below or equal to α, that is,

$$\sum_i \sum_j P(X = i) \cdot P(Y = j),$$

where the sum is taken over all pairs (i, j) such that $p_i \leq \alpha$, $0 \leq i \leq n_1$, $0 \leq j \leq n_2$. The exact power of each method is given in Table 6.

$\alpha = 0.05$		n_1/n_2	$\hat{\gamma_1}$	$\hat{\gamma_2}$	$\hat{\gamma_3}$
	H_0	220/200	0.030	0.034	0.034
		110/100	0.029	0.039	0.039
		220/55	0.031	0.038	0.030
		440/55	0.020	0.027	0.025
		300/75	0.032	0.041	0.043
	$H_1 >$	220/200	0.788	0.808	0.808
		110/100	0.448	0.476	0.476
		220/55	0.350	0.358	0.341
		440/55	0.415	0.432	0.412
		300/75	0.518	0.546	0.522
	$H_1 <$	220/200	0.790	0.807	0.807
		110/100	0.469	0.501	0.501
		220/55	0.400	0.472	0.482
		440/55	0.452	0.507	0.521
		300/75	0.543	0.603	0.603
$\alpha = 0.1$					
	H_0	220/200	0.065	0.084	0.084
		110/100	0.051	0.068	0.067
		220/55	0.052	0.071	0.069
		440/55	0.039	0.052	0.049
		300/75	0.065	0.080	0.079
	$H_1 >$	220/200	0.870	0.885	0.885
		110/100	0.585	0.623	0.620
		220/55	0.475	0.500	0.499
		440/55	0.560	0.573	0.570
		300/75	0.651	0.671	0.664
	$H_1 <$	220/200	0.855	0.871	0.871
		110/100	0.597	0.623	0.621
		220/55	0.535	0.593	0.588
		440/55	0.575	0.629	0.627
		300/75	0.673	0.718	0.711

Table 5: Simulated $\hat{\gamma}_1$, $\hat{\gamma}_2$, $\hat{\gamma}_3$, using the p-value p_1, p_2, p_3, respectively, for combinations of n_1, n_2, p_x and p_y when $\alpha = 0.05$ and $\alpha = 0.10$. "H_0" denotes $p_x = p_y = 0.1$, "$H_1 >$ " denotes $p_x = 0.2$, $p_y = 0.1$ and "$H_1 <$ " denotes $p_x = 0.1$, $p_y = 0.2$. n_1 and n_2 are given in column n_1/n_2.

$\alpha = 0.05$		n_1/n_2	γ_1	γ_2	γ_3
	H_0	220/200	0.034	0.041	0.041
		110/100	0.029	0.037	0.037
		220/55	0.028	0.036	0.035
		440/55	0.028	0.039	0.038
		300/75	0.029	0.038	0.036
	$H_1 >$	220/200	0.787	0.807	0.807
		110/100	0.459	0.491	0.489
		220/55	0.357	0.365	0.351
		440/55	0.392	0.409	0.393
		300/75	0.493	0.523	0.500
	$H_1 <$	220/200	0.790	0.808	0.808
		110/100	0.465	0.496	0.496
		220/55	0.416	0.474	0.486
		440/55	0.464	0.521	0.537
		300/75	0.545	0.597	0.598
$\alpha = 0.1$					
	H_0	220/200	0.071	0.085	0.085
		110/100	0.060	0.077	0.073
		220/55	0.056	0.076	0.074
		440/55	0.060	0.075	0.073
		300/75	0.062	0.082	0.077
	$H_1 >$	220/200	0.870	0.881	0.881
		110/100	0.583	0.615	0.614
		220/55	0.485	0.509	0.507
		440/55	0.528	0.543	0.536
		300/75	0.628	0.647	0.641
	$H_1 <$	220/200	0.871	0.883	0.883
		110/100	0.582	0.622	0.621
		220/55	0.535	0.587	0.582
		440/55	0.584	0.638	0.635
		300/75	0.661	0.707	0.701

Table 6: Exact power, γ_1, γ_2, γ_3, using the p-value p_1, p_2, p_3, respectively, for combinations of n_1, n_2, p_x and p_y when $\alpha = 0.05$ or $\alpha = 0.10$. "H_0" denotes $p_x = p_y = 0.1$, "$H_1 >$ " denotes $p_x = 0.2$, $p_y = 0.1$ and "$H_1 <$ " denotes $p_x = 0.1$, $p_y = 0.2$. n_1 and n_2 are given in column n_1/n_2.

4 Discussion and conclusion

Having shown that the three p-values are valid in general for the null hypothesis $H_0 : p_x = p_y$ (see Appendix A), we consider the influence of the parameters n_1, n_2 and α on the power of the different methods.

When n_1 and n_2 are both over 200, all p-values often reject H_0 when H_0 is false. From Table 6 we see that it is most important to have a sufficient number of trials in both X and Y for the power of the different methods. This matters more for the power than increasing the number of trials in only one of the random variables. Having both n_1 and n_2 above 200 seems ideal for the cases considered in Chapter 3.3, as having fewer trials than this leads to more type II errors.

From Table 5 and 6, we see that changing α from 0.05 to 0.10 leads to roughly 35-45 more type I errors and 75-150 less type II errors. In practice this tradeoff has to be evaluated by considering the potential consequences of committing either error.

The exact calculations in Table 5 are within 95% Clopper-Pearson confidence intervals constructed from each $\hat{\gamma_i}$ with one exception when $\alpha = 0.1$, H_0 is true and with trials 440/55. These exact γ_i are not contained in 95% Clopper-Pearson confidence intervals of γ_i constructed from each $\hat{\gamma_i}$, while they are contained in a 99.9% confidence interval.

This outlier is probably due to randomness, as there is little difference in the sum of rejections between the simulated p-values and the exact calculations when $\alpha = 0.05$. Consider the upper half of Table 4, where each row corresponds to 1000 simulations. For the 15 000 simulations, the difference between the exact and simulated rejections for each p_i is less than 20 for all three p-values.

From Table 6 we see that p_1 is overall more conservative in rejecting H_0 than p_2 and p_3 both when H_0 is true and when H_1 is true. While p_1 is strictly smaller than p_2 and p_3 in some cases when H_1 is true, this does not lead to more rejections than p_2 (see Tables 4 and 6). This pattern holds for both $\alpha = 0.05$ and $\alpha = 0.10$.

p_2 seems in general to have the highest power of the three, with an exception being made when $H_1 : p_x = 0.1 < p_y = 0.2$ where p_3 seems to have higher power when $\alpha = 0.05$ (see Tables 5 and 6). However when $\alpha = 0.10$, p_2 has higher power than p_3. Thus there are a select few scenarios where p_3 performs better (see Table 6).

Overall, p_2 seems to have the highest power of the three methods. This finding is not too surprising as p_2 is used as the standard method of calculating p-values for the fisher.test function in the statistical software R (R Core Team 2022).

Bibliography

Agresti, Alan (1992). 'A Survey of Exact Inference for Contingency Tables'. In: *Statistical Science* 7.1, pp. 131–153. DOI: 10.1214/ss/1177011454. URL: https://doi.org/10.1214/ss/1177011454.

Bickel, Peter J. and Kjell A. Doksum (2002). *Mathematical Statistics: Basic Ideas and Selected Topics*. 2nd ed. Vol. 1. Prentice Hall.

Casella, George and Roger L. Berger (2002). *Statistical Inference*. 2nd ed. Thomson Learning.

Everitt, B.S. and Anders Skrondal (2010). *The Cambridge Dictionary of Statistics*. 3rd ed. Cambridge University Press.

Freedman, Laurence S. (Dec. 2008). 'An analysis of the controversy over classical one-sided tests'. In: *Sage Journals* 5, pp. 635–640.

Kulinskaya, Elena (2008). *On two-sided p-values for non-symmetric distributions*. DOI: 10.48550/ARXIV.0810.2124. URL: https://arxiv.org/abs/0810.2124.

Larsen, Richard J. and Morris L. Marx (2018). *An introduction to Mathematical Statistics and its Applications*. 6th ed. Pearson Education.

New Manuscripts (n.d.). https://www.nejm.org/author-center/new-manuscripts. Accessed: 2022-05-27.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Rooney, Anne (2012). *The History of Physics*. Rosen Publishing.

Wickham, Hadley (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org.

Appendix

A Proofs of validity of the 2-sided p-values

A.1 The doubled one-sided p-value, p_1

Let $p(\mathbf{X}) = 2\ \min(P_r, P_l, \frac{1}{2})$ where P_r and P_l are the one-sided p-values corresponding to Fisher's one-sided test as in (7). Then by (4), for $\alpha \in [0, 1)$, when H_0 is true we see that

$$P(p(\mathbf{X}) \leq \alpha) = P\left(2\min\left(P_r, P_l, \frac{1}{2}\right) \leq \alpha\right) = P\left(\min\left(P_r, P_l, \frac{1}{2}\right) \leq \frac{\alpha}{2}\right)$$
$$\leq P\left(P_r \leq \frac{\alpha}{2}\right) + P\left(P_l \leq \frac{\alpha}{2}\right) + P\left(\frac{1}{2} \leq \frac{\alpha}{2}\right) = \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha$$

and $P(p(\mathbf{X}) \leq \alpha) \leq \alpha$ by definition for $\alpha = 1$. This means that the p-value is valid.

A.2 The sum over all smaller probabilities, p_2

Let $p(x) = \sum_i f(i \mid c)$ be the sum of all $f(i \mid c)$ less than or equal to $f(x \mid c)$ as in (8). Denote by x_1 the x to give the largest $p(x)$ which is smaller than or equal to α. Denote by x_2 the x to give the second largest $p(x)$, and so on until x_n), the giving the smallest nonzero p-value. Then by (4),

$$P(p(x) \leq \alpha) = P(X = x_1 \cup X = x_2 \cup \cdots \cup X = x_n) = \sum_{i=1}^{n} f(x_i \mid c) = p(x_1) \leq \alpha$$

which means that $p(\mathbf{x})$ is a valid p-value.

A.3 Sum of all probabilities with absolute larger deviations from the mean, p_3

Let $W(\mathbf{X}) = |X - E(X)|$. Then

$$P(W(\mathbf{X}) \geq W(\mathbf{x})) = P(|X - E(X)| \geq |x - E(X)|) = \sum_j f(j \mid c)$$

where the sum is over all j such that $|j - E(X)| \geq |x - E(X)|$. Thus by Theorem 1, the p-value is valid.