UNIVERZA NA PRIMORSKEM
FAKULTETA ZA MATEMATIKO, NARAVOSLOVJE IN
INFORMACIJSKE TEHNOLOGIJE

Zaključna naloga
(Final project paper)

O neeksaknotsti eksaktnega Fisherjevega testa za dva neodvisna deleža

(On the inexactness of Fisher's exact test for testing the equality of two independent proportions)

Ime in priimek: Marija Tepegjozova
Študijski program: Matematika
Mentor: doc. dr. Rok Blagus

Koper, avgust 2017

Ključna dokumentacijska informacija

Ime in PRIIMEK: Marija TEPEGJOZOVA

Naslov zaključne naloge: O neeksaknotsti eksaktnega Fisherjevega testa za dva neodvisna deleža

Kraj: Koper

Leto: 2017

Število listov: 37 Število slik: 5 Število tabel: 12

Število prilog: 1 Število strani prilog: 2 Število referenc: 18

Mentor: doc. dr. Rok Blagus

Ključne besede: Testiranje hipotez, konzervativen test, χ^2 test, $z-$test, Fisherjev eksaktni test, simulacije.

Math. Subj. Class. (2010): 62F03, 62G10, 62H17

Izvleček:
Fisherjev eksaktni test je ena od najpogosteje uporabljenih metod za preverjanje povezave med dvema kategorialnima spremenljivkama. Temelji na hipergeometrični diskretni porazdelitvi vnaprej določene testne statistike. Test je asimptotično enakovreden testu χ^2, ki je ekvivalenten $z-$testu za enakost dveh neodvisnih deležev kadar imata obe spremenljivki le dve vrednosti. Medtem ko sta zadnja dva testa le približna testa, se pravi da, je njuna velikost enaka nominalni ravni, ko gre n v neskončnost, tukaj pokažemo, da je tudi Fisherjev točni test netočen pri majhnih vzorcih z uporabo 2×2 preglednih tabel. Fisherjev točni test velja za točen test, kar pomeni, da je pri njem nominalna raven enaka efektivni napaki prvega tipa, vendar temu ni tako. Pravzaprav je konzervativen test, kar pomeni, da resnična verjetnost nepravilne zavrnitve ničelne hipoteze nikoli ni večja ali enaka nominalni ravni.

Key words documentation

Name and SURNAME: Marija TEPEGJOZOVA

Title of final project paper: On the inexactness of Fisher's exact test for testing the equality of two independent proportions

Place: Koper

Year: 2017

Number of pages: 37 Number of figures: 5 Number of tables: 12

Number of appendices: 1 Number of appendix pages: 2 Number of references: 18

Mentor: Assist. Prof. Rok Blagus, PhD

Keywords: Hypothesis testing, conservative test, χ^2 test, $z-$test, Fisher's exact test, simulations.

Math. Subj. Class. (2010): 62F03, 62G10, 62H17

Abstract:
Fisher's exact test is one of the most commonly used methods for testing the association between two categorical variables. It is based on the hypergeometric discrete distribution of the predefined test statistic. The test is asymptotically equivalent to the χ^2 test, which when both variables have only two levels, is equivalent as the $z-$test for the equality of two independent proportions. While the latter two tests are known to be only approximate tests, i.e. their size is equal to the nominal level when n goes to infinity, we show that also the Fisher's exact test is inexact with small samples using 2×2 contingency tables. The Fisher's exact test is said to be an exact test, and a statistical test for which the nominal level is equal to the effective type I error, but that is not the case. It is actually a conservative test, meaning that the true probability of incorrectly rejecting the null hypothesis is never greater or equal to the nominal level.

Acknowledgement

I would like to express my deepest gratitude to my advisor, Assist. Prof. Rok Blagus, PhD, for his support and guidance throughout the research. His continued support led me to the right way.

Also, I would like to thank Aljaž Ule and Sladjana Babič for helping me with the English-Slovenian translation and Anes Valentić for the technical help.

Finally, I would like to thank my parents and my brother, for the unconditional love and support.

Contents

List of Tables

List of Figures

Appendices

A Tables for comparison of α_e for Fisher's and χ^2 test

List of Abbreviations

i.e. that is

a.e. almost everywhere

1 Introduction

1.1 Lady Tasting Tea

In the early XX-th century the very famous statistician and biologist Ronald Fisher during a conversation with a friend of his, Dr. Muriel Bristol, came up with a very profound idea in statistics. Namely, during drinking their teas, Dr. Bristol claimed that she could distinguish whether the tea or the milk was poured first in her cup of tea. Fisher doubted her claim, and wanted to test her claim. Therefore, he designed an experiment as follows. He provided 8 cups of tea from which in 4 of then the tea was poured first, and the other 4 the milk was poured first. He randomly ordered the cups, explained that there are 4 cups in which the tea was poured first, and 4 were with the milk poured first and asked Bristol to taste them and choose four of them of one type.

This randomized experiment, the lady tasting tea, was elaborately explained in his book entitled 'The Design of Experiments' [7]. It was then when he firstly introduced the notion of null hypothesis,

"...the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis."

In this experiment the null hypothesis, or the hypothesis which we want to test, was the claim that the lady is not able to distinguish whether milk or tea is poured first. The proposed test statistic, or the value needed for comparison was the number of successful selections in the 4 cups the lady choose. The null hypothesis distribution was calculated by the number of permutations. Thus, given 8 cups and choosing 4 of them, gives us

$$\frac{8!}{4!(8-4)!} = 70$$

possible combinations.

The critical region, or the region in which we will reject the validity of the null hypothesis was the single case when the lady successfully guessed all 4 cups she chose. The probability of doing so is 1 in 70 which is 0.014, giving a significance level of 1.4% .

This simple experiment is one of the supporting pillars of the topic of hypothesis testing and randomization of experimental data [6].

Bellow are all the possible outcomes of the experiment, where with empty dots we denote a successful guess and the x's are faulty guesses.

Success count	Permutations of selection	Number of permutations
0	oooo	$1 \ \ 1 = 1$
1	oox, ooxo, oxoo, xooo	$4 \ \ 4 = 16$
2	ooxx, oxox, oxxo, xoxo, xxoo, xoox	$6 \ \ 6 = 36$
3	oxxx, xoxx, xxox, xxxo	$4 \ \ 4 = 16$
4	xxxx	$1 \ \ 1 = 1$
Total		70

Table 1: Lady Tasting Tea experiment

1.2 Hypothesis Testing

In statistics a *hypothesis* is an assumption about a population parameter which may or may not be true. Hypothesis testing is the process of deciding from a sample whether some stated hypothesis is correct. The decision is to be made between accepting or rejection the hypothesis stated. A decision procedure for such a problem is called *a test of the hypothesis*. A *test statistic* is a value calculated from the sample, in a way that it summarizes the sample for comparison purposes.

During a statistical investigation we need to carefully define the population, then we need to randomly select a sample from the population which will be our *set of observations* i.e. the values our chosen random variable X takes. We also need to define an assumption about the parameter θ which will label X and its distribution P_θ, this will be our hypothesis. The set of all possible values the parameter θ can take is called a *parameter space*, and is denoted with Ω.

The decision in hypothesis testing is made based on the outcome of a certain random variable X and the distribution P_θ which belongs to a distribution class $P = \{P_\theta, \theta \in \Omega\}$. The distributions of P can be classified into two mutually exclusive classes, one for which the hypothesis is true and the other for which the hypothesis is false. We will denote them by H_0 and H_a, respectively. The parameter space will be also similarly divided into Ω_0 and Ω_a, each consisting of parameters for which the hypothesis is true or false, respectively. Also, note that $H_0 \cup H_a = P$ and $\Omega_0 \cup \Omega_a = \Omega$. Mathematically,

whether the hypothesis is true is equivalent to whether P_θ in an element of H_0. Therefore, it is convenient to identify the hypothesis with the above statement and denote the testing hypothesis with H_0, usually stated as *null hypothesis*. The distributions that are in H_a we call *alternatives*, and we say that H_a is a class of alternatives. Now let us define a *decision function* δ whose domain is X and its range is $\{d_0, d_a\}$ where the decision of accepting the null hypothesis is assigned the value d_0 and the decision of rejecting the null hypothesis is assigned the value d_a.

A nonrandomized test procedure assigns to each possible value x of X one of these two decisions, accept or reject, and divides the sample space into two complementary regions S_0 and S_a. If X falls into S_0, the hypothesis is accepted, otherwise it is rejected. Therefore, the set S_0 is called *region of acceptance*, and the set S_a the region of rejection or *critical region*.

When we perform a test we may chose the correct decision or make one of the two possible mistakes. The first one is rejecting the null hypothesis when it is true, which is called *type I error* and is denoted by α

$$\alpha = P(\delta(X) = d_a | \theta \in \Omega_0) = P(X \in S_a | \theta \in \Omega_0).$$

The other one is accepting the null hypothesis when it is false, *type II error* denoted by $1 - \beta(\theta)$, where $\beta(\theta)$ is the *power function* of the test,

$$\beta(\theta) = P(X \in S_0 | \theta \in \Omega_a)$$

$$1 - \beta(\theta) = P(X \in S_a | \theta \in \Omega_a).$$

The consequences done by these mistakes vary. For example, if we test the presence of some virus, and our test leads us to a conclusion that the virus is not present, but the patient in reality is infected by the virus, such diagnosis may lead to a death of the patient. Therefore, we need to minimize the probability of the errors occurring. However, we can not control both those probabilities simultaneously . Therefore, it is customary to bound the probability of incorrectly rejecting H_0 when it is true and try to minimize the other probability. Thus, we bound $\alpha \in (0, 1)$ and we also call it *level of significance*. Usually α is somewhat arbitrary, but usage of conventional levels of 0.01 and 0.05 is the most common procedure. These values were originally chosen to reduce the number of tables needed to carry out various computations in different tests. Later they were adapted mainly due to habit and due to convenience of standardization in providing a common frame of reference in different tests [11].

1.3 Contingency Tables

A categorical variable is one that has a measurement scale consisting of a set of categories. In our case we are going to focus on variables having only two categories. For example, gender is a categorical variables having two categories (male and female) or whether a treatment is successful or not in biomedical statistics. Usually, we distinguish between two types of categorical variable, response or dependent variables. A response variable is the particular quantity for which we ask a question in a study and an explanatory variable is any influence or factor that can influence the response variable. As an example we may consider testing whether the number of hours spent doing homework has an effect on the grade a student earns on an exam. In such a case, we are having a variable and we would like to know how it affects another variable. This means that the variable representing the number of hours studied is an explanatory variable and the score on the test is a response variable. Now we can define a contingency table.

Let X and Y be two categorical response variables, such that X has I possible categories and Y has J categories. Classifications of subjects on both variables (X, Y) have $I \times J$ possible combinations. Such responses (X, Y) from a sample have a probability distribution. A rectangular $I \times J$ table, that has I rows for categories of X and J columns for categories of Y, represents this distribution. The cells of such table represent each of the $I \times J$ possible outcomes. When the cells contain frequency counts of outcomes for a sample it is called a contingency table or a cross-classification table. A contingency table with I rows and J columns is called an $I \times J$ table [1].

In our case we will only only consider 2×2 contingency tables, with two discrete response variables.

1.4 The Neyman-Pearson Fundamental Lemma

Definition 1.1. A uniformly most powerful (UMP) test is a hypothesis test that has the greatest power $\beta(\theta)$ among all possible tests of a given size α.

If a class of distributions contains a single distribution then it is called *simple*, otherwise it is said to be *composite*. The problem of hypothesis testing is said to be completely specified when H_a is simple. Its solution can be given explicitly when the same is true for H_0.

Now, let the distributions under a simple hypothesis H_0 and alternative H_a be P_0 and P_a, respectively and suppose that these distributions are discrete with $P_i(X = x) = P_i(x)$ for $i = 0, a$.

Theorem 1.2. *Let P_0 and P_a be probability distributions possessing densities p_0 and p_1 respectively with respect to a measure .*
(i) Existence. For testing $H_0 : p_0$ against the alternative $H_a : p_a$ there exists a test ϕ and a constant k such that

$$E_0\phi(X) = \alpha \tag{1.1}$$

and

$$\phi(x) = \begin{cases} 1 & when \quad p_a(x) > kp_0(x) \\ 0 & when \quad p_a(x) < kp_a(x). \end{cases} \tag{1.2}$$

(ii) Sufficiency condition for a most powerful test. If a test satisfies 1.1 and 1.2 for some k, then it is most powerful for testing p_0 against p_a at level α.
(iii) Necessary condition for a most powerful test. If ϕ is most powerful at level α for testing p_0 against p_a, then for some k it satisfies 1.2 a.e. μ. It also satisfies 1.1 unless there exists a test of size $< \alpha$ and with power 1.

Proof. We omit the proof. It can be read from here reference. □

Corollary 1.3. *Let β denote the power of the most powerful level$-\alpha$ test $(0 < \alpha < 1)$ for testing P_0 against P_a. Then $\alpha < \beta$ unless $P_0 = P_a$.*

Proof. Since the level$-\alpha$ test given by $\phi(x) \equiv \alpha$ has power α, it is seen that $\alpha \leq \beta$. If $\alpha = \beta < 1$, the test $\phi(x) \equiv \alpha$ is most powerful and by Theorem 1.2 (iii) must satisfy 1.2. Then $p_0(x) = p_a(x)$ a.e. μ and hence $P_0 = P_a$. □

1.5 $p-$value

The $p-$value is one another important concept in hypothesis testing. It is defined as the smallest significance level, at which the null hypothesis would be rejected for the given observation. In other words, it is the probability of a result as or more extreme then the actually observed one if the null hypothesis is true. Mathematically, we define it as the value of the test statistic T on data t

$$p = \sup_{\theta \in \Omega_0} P(T \geq t).$$

The smaller the $p-$values is, we have stronger evidence against the null hypothesis.

We have statistically significant results only when $p < \alpha$, otherwise the results are disregarded.

Having said all of this, the question is what happens if we increase or decrease α, our predefined significance level, to an already obtained set of data. That is, one can ask what happens in the situation if you reject the null hypothesis at a level α_1, will you still reject the null hypothesis at another significance level α_2 for which holds that $\alpha_1 \leq \alpha_2$. The question can be partially answered if we are using at both significance levels the most powerful nonrandomized α test under the assumption that the null hypothesis holds. Using the most powerful nonrandomized α test we get the rejection regions nested, that is

$$S_{\alpha_1} \subset S_{\alpha_2} \qquad if\ \alpha_1 < \alpha_2, \tag{1.3}$$

where S_α denotes the rejection region for the significance level α. We also have to note that if we are not using the most powerful nonrandomized α test this does not need to be the case. But when this is the case, we have the rejection regions nested, and then we define the $p-$value as exactly the smallest significance level at which the null hypothesis would be rejected for the given observation. More formally written we have the $p-$value defined as

$$\hat{p}\,(X) = \inf\{\alpha \ :\ X \in S_\alpha\}\,.$$

Returning the $p-$value gives an idea of how strongly a certain observation contradicts the null hypothesis. Now we can go on to investigating some general properties of the $p-$values with the following theorem:

Theorem 1.4. *Suppose X has distribution P_θ for some $\theta \subset \Omega$, and the null hypothesis H specifies $\theta \in \Omega_0$. Assume further the rejection regions satisfy the nesting property 1.3 then we have*
(i) if

$$\sup_{\theta \in \Omega_0} P_\theta \{X \in S_\alpha\} \leq \alpha \quad for\ all\ 0 < \alpha < 1$$

then the distribution of $\hat{p}$ under $\theta \in \Omega_0$ satisfies

$$P_\theta \{\hat{p} \leq u\} \leq u \quad for\ all\ 0 \leq u \leq 1$$

(ii) if, for $\theta \in \Omega_0$,

$$P_\theta \{X \in S_\alpha\} = \alpha \quad for\ all\ 0 < \alpha < 1$$

then

$$P_\theta \{\hat{p} \leq u\} = u \quad for\ all\ 0 \leq u \leq 1$$

that is $\hat{p}$ is uniformly distributed over $(0, 1)$.

Proof. (i) If $\theta \in \Omega_0$ then because of the nesting property of rejection regions $\{\hat{p} \leq u\}$ implies $\{X \in S_v\}$ for all $u < v$. Now from the assumption we have that

$$P_\theta \{X \in S_v\} \leq v \quad for\ all\ 0 \leq v \leq 1$$

since $\{\hat{p} \leq u\}$ implies $\{X \in S_v\}$ for all $u < v$ we will have that

$$P_\theta \{\hat{p} \leq u\} \leq P_\theta \{X \in S_v\} \leq v \quad for\ all\ 0 \leq v \leq 1$$

taking the limit as $v \longrightarrow u$ we obtain the following

$$\begin{aligned} \lim_{v \longrightarrow u} P_\theta \{\hat{p} \leq u\} = P_\theta \{\hat{p} \leq u\} &\leq \lim_{v \longrightarrow u} P_\theta \{X \in S_v\} \\ &= P_\theta \{X \in S_u\} \\ &\leq u \end{aligned}$$

so we obtain

$$P_\theta \{\hat{p} \leq u\} \leq u \quad for\ all\ 0 \leq u \leq 1$$

which is exactly what we needed.
(ii) Since we have that the event $\{X \in S_u\}$ implies $\{\hat{p} \leq u\}$ which follows directly from the definition of the $p-$value and since the first event implies the second we have that

$$P_\theta \{\hat{p} \leq u\} \geq P_\theta \{X \in S_u\}$$

now if the assumption from part (ii) holds (which implies the assumption from part (i)) we have that

$$P_\theta \{\hat{p} \leq u\} \geq P_\theta \{X \in S_u\} = u \quad for\ all\ 0 \leq u \leq 1$$

but from part (i) we have that $P_\theta \{\hat{p} \leq u\} \leq u \quad for\ all\ 0 \leq u \leq 1$ so this implies that

$$P_\theta \{\hat{p} \leq u\} = u \quad for\ all\ 0 \leq u \leq 1$$

which completes our proof. □

2 Theoretical Part

2.1 Definitions and Theorems

Definition 2.1. The normal distribution with mean μ and variance σ^2 is denoted with $N(\mu, \sigma^2)$ and has the probability density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

The standard normal distribution is the normal distribution with mean 0 and variance 1.

Definition 2.2. The Chi-square distribution (χ^2-distribution) with k degrees of freedom is the distribution of the sum of squares of k independent standard normal random variables. The Chi-square distribution with k degrees of freedom is denoted as χ^2_k and has the probability density function

$$f(x) = \begin{cases} \frac{x^{(k/2-1)}e^{-x/2}}{2^{k/2}\Gamma\left(\frac{k}{2}\right)} & x > 0; \\ 0 & otherwise, \end{cases}$$

where $\Gamma(k/2)$ denotes the Gamma function.

Definition 2.3. The Bernoulli distribution with parameter p is the distribution of a random variable which takes the value 1 with probability p and the value 0 with probability $1-p$. The probability mass function of this distribution is

$$P(X = k) = p^k(1-p)^{1-k} \qquad for\ k \in \{0, 1\}.$$

Definition 2.4. The Binomial distribution with parameters n and p is the discrete probability distribution of the number of successes in a sequence of n independent trials of which each trial has a Bernoulli distribution with parameter p. This distribution has the probability mass function:

$$P(X = k) = \binom{n}{k} p^k(1-p)^{1-k},$$

for $k = 1, 2, \ldots n$.

Definition 2.5. The Multinomial distribution is the probability distribution that models the probability of any particular combination of numbers of successes for various categories where n independent trials are conducted, each of which leads to a success for exactly one out of the k categories with each category having a given fixed success probability p_i. The multinomial distribution with n trials and $\mathbf{p} = (p_1, \ldots, p_k)$ defining the success probabilities has the probability mass function

$$P(X_1 = x_1 \, and \ldots and \, X_k = x_k) = \begin{cases} \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k} & when \ \sum_{i=1}^{k} x_i = n \\ 0 & otherwise, \end{cases}$$

Definition 2.6. The Hypergeometric distribution is a discrete probability distribution describing the probability of k successes in n draws from a finite population of size N containing exactly K successes, without replacement, where each draw is either a success or a failure.
The probability mass function of hypergeometric distribution is defined as

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}.$$

Definition 2.7. The Uniform continuous distribution on an interval $[a, b]$, denoted with $U(a, b)$, is the distribution with the probability mass function

$$f(x) = \begin{cases} \frac{1}{b-a} & for \ a \leq x \leq b \\ 0 & for \ x < a \ or \ x > b. \end{cases}$$

The cumulative distribution function of $U(0, 1)$ is

$$F(x) = \begin{cases} 0 & : \ x < 0 \\ x & : \ 0 \leq x < 1 \\ 1 & : \ x \geq 1. \end{cases}$$

Theorem 2.8. *Lebesgue's Dominated Convergence Theorem Suppose* $\{f_n\}$ *is a sequence of complex measurable functions on* X *such that*

$$f(x) = \lim_{n \longrightarrow \infty} f_n(x)$$

exists for every $x \in X$*. If there is a function* $g \in L^1(\mu)$ *such that*

$$|f_n(x)| \leq g(x)$$

for $n = 1, 2, 3...$ *then* $f \in L^1(\mu)$*,*

$$\lim_{n \longrightarrow \infty} \int_X |f_n - f| d\mu = 0$$

and

$$\lim_{n \longrightarrow \infty} \int_X f_n d\mu = \int_X f d\mu.$$

Proof. We omit the proof here. It is given in [15]. □

Theorem 2.9. *Central Limit Theorem Let $X_1, X_2...$ be a sequence of random variables, which are independent and identically distributed and let $S_n = X_1 + ... + X_n$. Assume $E(|X_I|) < \infty$ and $E(|X_I|^2) < \infty \ \forall i$, then for any $a < b$ we have*

$$\lim_{n \longrightarrow \infty} P\left\{ a \leq \frac{S_n - n\mu}{\delta\sqrt{n}} \leq b \right\} = \phi(b) - \phi(a),$$

where $\mu = E(X_1)$, $\delta = var(X_1)$ and ϕ is the distribution function of standard normal distribution.

Proof. We omit the proof here. It is given in [10]. □

Theorem 2.10. *Multidimensional Central Limit Theorem Let $\mathbf{X_1}, \ldots, \mathbf{X_n}$ be independent $\mathbb{R}^d$-valued vectors, each having mean zero. Write $\mathbf{S} = \sum_{i=1}^{n} \mathbf{X_i}$ and assume $\mathbf{\Sigma} = Cov\,[\mathbf{S}]$ is invertible. Let $\mathbf{Z} \sim \mathbf{N}(0, \mathbf{\Sigma})$ be a $d-$dimensional Multinormal distribution with the same mean and covariance matrix as $\mathbf{S}$. Then for all convex set $U \subseteq \mathbb{R}^d$ we have that*

$$|P[\mathbf{S} \in U] - P[\mathbf{Z} \in U]| \leq Cd^{1/4}\gamma,$$

where C is a universal constant, $\gamma = \sum_{i=1}^{n} E\left[||\mathbf{\Sigma}^{1/2}\mathbf{X_i}||^3\right]$.

Proof. We omit the proof here. It is given in [2]. □

In other words we can interpret this result as that $\mathbf{S}$ converges in distribution to $\mathbf{Z}$.

3 Pearson's Chi-Squared Test

The Pearson's Chi-Squared Test is one of the most common tests for testing the independence of two binomial proportions. The test is used to determine whether the two variables are independent of each other or whether there is a pattern of dependence between them. Under the null hypothesis it assumes that there is no relationship between the two variables, i.e. that the variables are independent, and under the alternative hypothesis that there is some relationship between the two variables, that is that there is a pattern of dependence between them. If we denote the variables as X and Y, the hypotheses are:

$$H_0 : \textit{ No relationship between } X \textit{ and } Y$$

$$H_a : \textit{ Some relationship between } X \textit{ and } Y$$

In terms of independence we can state them as follows:

$$H_0 : \textit{ } X \textit{ and } Y \textit{ are independent}$$

$$H_a : \textit{ } X \textit{ and } Y \textit{ are dependent}$$

The Chi-Squared Test uses 2×2 contingency tables to examine the nature of relationship between the two variables. The test will say whether the observed pattern between the two variables in the table is strong enough to conclude that the two variables are dependent on each other or not [17].

3.1 Chi-Squared Statistic

The value of the Pearson Chi-square test statistics is defined as follows:

$$\chi^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j} = N \sum_{j=1}^{k} \frac{\left(\frac{O_j}{N} - p_j\right)^2}{p_j},$$

where we use χ^2 to denote the cumulative test statistic, for which we will later prove that it approaches to a χ^2-distribution. The notation used is the following: O_i is the number of observations of type i, N the total number of observations, k the total

number of cells in the table and E_i is the expected number of outcomes of type i which is known to be (theoretically) Np_i since our sample comes from a Multinomial distribution.

If we want to test whether there is a statistical dependence between r observations, and the null hypothesis is chosen to be that the observations are statistically independent, the contingency table that arises will have r rows and c columns. The theoretical expectation of a cell in the contingency table, given the assumption of independence and due to the fact that the cells are chosen from a Multinomial distribution, is

$$E_{i,j} = Np_{i+}p_{+j},$$

where we define p_{i+} and p_{+j} as follows:

$$p_{i+} = \frac{O_{i+}}{N} = \sum_{j=1}^{c} \frac{O_{i,j}}{N} \quad and \quad p_{+j} = \frac{O_{+j}}{N} = \sum_{i=1}^{r} \frac{O_{i,j}}{N}.$$

Thus, plugging in gives us the test statistic to be

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} = N \sum_{i=1}^{r}\sum_{j=1}^{c} p_{i+}p_{+j} \left(\frac{\frac{O_{i,j}}{N} - p_{i+}p_{+j}}{p_{i+}p_{+j}} \right)^2.$$

Now, we will consider our case of interest, when we examine the case where our test consists of testing the statistical independence of two observations both having only two possible outcomes, where the null hypothesis is stated that the observations are statistically independent. Then our contingency table will have two rows and two columns, that is the contingency table will look as follows

	X	Y	Total
A	a	b	n_1
B	c	d	n_2
	m_1	m_2	N

where we denote $a = p_{11},\ b = p_{12},\ c = p_{21},\ d = p_{22}$ and $n_1 = Np_{1+},\ n_2 = Np_{2+},\ m_1 = Np_{+1},\ m_2 = Np_{+2}$ and thus our test statistics will look as follows

$$\chi^2 = \frac{1}{N}\left[\frac{(aN - n_1m_1)^2}{n_1m_1} + \frac{(bN - n_1m_2)^2}{n_1m_2} + \frac{(cN - n_2m_1)^2}{n_2m_1} + \frac{(dN - n_2m_2)^2}{n_2m_2} \right].$$

The above form of the test statistic simplifies to the following form

$$\chi^2 = \frac{N(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)},$$

which is the most common form of the Chi-Square test statistic for 2×2 contingency tables.

3.2 Pearson's Theorem

Let us consider the following situation: we have r labeled boxes $B_1, B_2, \ldots, B_r$ and we throw n balls $X_1, X_2, \ldots, X_n$ into the boxes randomly and independent of each other with probabilities of hitting a box with some ball given as

$$P(X_i \in B_1) = p_1, \ldots, P(X_i \in B_r) = p_r,$$

where we assume that every ball falls in some box, that is that the probabilities $p_1, \ldots, p_r$ add up to one. Let us now define ϑ_j as the number of balls that fall into box j that is

$$\vartheta_j = \sum_{l=1}^{n} I(X_l \in B_j).$$

Further we have that

$$E(\vartheta_j) = \sum_{l=1}^{n} 1 \cdot p_j + 0 \cdot p_j = np_j$$

Now we have the following theorem:

Theorem 3.1. *We have that the random variable*

$$\sum_{j=1}^{r} \frac{(\vartheta_j - np_j)^2}{np_j} \longrightarrow \chi^2_{r-1} \tag{3.1}$$

converges in distribution to the χ^2_{r-1} distribution with $r-1$ degrees of freedom.

Proof. Let us now observe a fixed box B_j. The random variables

$$I(X_1 \in B_j), \ldots, I(X_n \in B_j)$$

that indicate whether each observation X_i is in the fixed box B_j or not are independent and identically distributed with Bernoulli distribution $B(p_j)$ with expectation and variance

$$E\left(I(X_1 \in B_j)\right) = P\left(X_1 \in B_j\right) = p_j$$
$$Var\left(I(X_1 \in B_j)\right) = p_j(1-p_j).$$

Defining the vector $\boldsymbol{\Upsilon} = \left(\frac{\vartheta_1 - np_1}{\sqrt{np_1(1-p_1)}}, \ldots, \frac{\vartheta_r - np_r}{\sqrt{np_r(1-p_r)}}\right)$ and using the Multidimensional Central Limit Theorem we directly obtain that $\boldsymbol{\Upsilon}$ converges in distribution to $\mathbf{Z} \sim N(0, \Sigma)$ which has multinormal distribution. Further we know that, by the Central Limit Theorem, the random variable

$$\begin{aligned}
\frac{\vartheta_j - np_j}{\sqrt{np_j(1-p_j)}} &= \frac{\sum_{l=1}^{n} I(X_l \in B_j) - np_j)}{\sqrt{np_j(1-p_j)}} \\
&= \frac{\sum_{l=1}^{n} I(X_l \in B_j) - nE\left(I(X_l \in B_j)\right)}{\sqrt{nVar\left(I(X_l \in B_j)\right)}} \\
&\longrightarrow N(0,1)
\end{aligned}$$

converges to the standard normal distribution [14]. Therefore, we have that the random variable

$$\frac{\vartheta_j - np_j}{\sqrt{np_j}} \longrightarrow \sqrt{1-p_j} \cdot N(0,1) = N\left(0, 1-p_j\right),$$

converges to the normal distribution with mean zero and variance $\sqrt{1-pj}$. Let us now write that

$$\frac{\vartheta_j - np_j}{\sqrt{np_j}} \longrightarrow Z_j.$$

where Z_j is a random variable with distribution $N\left(0, \sqrt{1-p_j}\right)$. Here we need to note that Z_j is a marginal distribution of the multinormal $\mathbf{Z}$ since we have that $\boldsymbol{\Upsilon}$ converges in distribution to $\mathbf{Z}$. And we have that

$$\sum_{j=1}^{r} \frac{\left(\vartheta_j - np_j\right)^2}{np_j} \longrightarrow \sum_{j=1}^{r} Z_j^2. \tag{3.2}$$

Unfortunately we can not say much about the distribution of $\sum Z_j^2$ from this result since we do not know whether the random variables Z_j are independent or not. We can easily see that the random variables ϑ_j are not independent, because going back to the definition of the ϑ_j through balls and boxes we see that the total number of balls is n so we will have that $\sum \vartheta_j = n$ and thus if we know the value of $n-1$ variables we will automatically have the value for the $n-$th. This means that we will need the covariance between Z_i and Z_j, but first let us compute the covariance between $\frac{\vartheta_j - np_j}{\sqrt{np_j}}$ and $\frac{\vartheta_i - np_i}{\sqrt{np_i}}$. Since we have that

$$E\left(\frac{\vartheta_j - np_j}{\sqrt{np_j}}\right) = \frac{E\left(\vartheta_j\right) - np_j}{\sqrt{np_j}} = \frac{np_j - np_j}{\sqrt{np_j}} = 0.$$

Analogously we have also for $E\left(\frac{\vartheta_i - np_i}{\sqrt{np_i}}\right) = 0$, thus we have that the covariance of $\frac{\vartheta_j - np_j}{\sqrt{np_j}}$ and $\frac{\vartheta_i - np_i}{\sqrt{np_i}}$ is equal to

$$E\left(\frac{\vartheta_j - np_j}{\sqrt{np_j}} \frac{\vartheta_i - np_i}{\sqrt{np_i}}\right).$$

Now to compute this expression we have

$$\begin{aligned} E\left(\frac{\vartheta_j - np_j}{\sqrt{np_j}} \frac{\vartheta_i - np_i}{\sqrt{np_i}}\right) &= \frac{1}{n\sqrt{p_i p_j}} \left(E(\vartheta_i \vartheta_j) - E(\vartheta_i np_j) - E(\vartheta_j np_i + n^2 p_i p_j)\right) \\ &= \frac{1}{n\sqrt{p_i p_j}} \left(E(\vartheta_i \vartheta_j) - np_i np_j - np_j np_i + n^2 p_i p_j\right) \\ &= \frac{1}{n\sqrt{p_i p_j}} \left(E(\vartheta_i \vartheta_j) - n^2 p_i p_j\right). \end{aligned}$$

To compute $E(\vartheta_j \vartheta_i)$ we will again look at the definition of ϑ_j and ϑ_i with balls and boxes. We will use the fact that one ball can not simultaneously be in two boxes which means that

$$I\left(X_l \in B_i\right) I\left(X_l \in B_j\right) = 0.$$

Thus we will have

$$\begin{aligned}
E(\vartheta_j\vartheta_i) &= E\left(\left(\sum_{l=1}^{n} I\left(X_l \in B_i\right)\right)\left(\sum_{k=1}^{n} I\left(X_k \in B_j\right)\right)\right) \\
&= E\left(\sum_{l=k} I\left(X_l \in B_i\right) I\left(X_k \in B_j\right)\right) + E\left(\sum_{l\neq k} I\left(X_l \in B_i\right) I\left(X_k \in B_j\right)\right) \\
&= n(n-1)E\left(I\left(X_l \in B_i\right)\right) E\left(I\left(X_k \in B_j\right)\right) \\
&= n(n-1)p_ip_j.
\end{aligned}$$

So, the covariance of $\frac{\vartheta_j - np_j}{\sqrt{np_j}}$ and $\frac{\vartheta_i - np_i}{\sqrt{np_i}}$ is equal to

$$\frac{1}{n\sqrt{p_ip_j}}\left(n(n-1)p_ip_j - n^2p_ip_j\right) = -\sqrt{p_ip_j}.$$

Now by Lebesgue Dominant Convergence theorem it follows directly that

$$E(Z_iZ_j) = -\sqrt{p_jp_i} \quad \forall i, j.$$

And since $Z_j \sim N(0, 1-p_j)$ we have that $E(Z_j^2) = 1 - p_j$. To finish the proof we need to show that this covariance structure will imply that the sum of Z_i's converges to χ^2_{r-1}. Let define the random variables $G_1, \ldots, G_r$ be a sequence of independent identically distributed random variables with standard normal distribution. And let us define the vectors

$$\mathbf{G} = (G_1, \ldots, G_r) \quad and \quad \mathbf{p} = (\sqrt{p_1}, \ldots, \sqrt{p_r}).$$

Consider the vector $\mathbf{V} = \mathbf{G} - (\mathbf{G} \cdot \mathbf{p})\mathbf{p}$, where $\mathbf{G} \cdot \mathbf{p} = G_1\sqrt{p_1} + \ldots + G_r\sqrt{p_r}$ is the scalar product of $\mathbf{G}$ and $\mathbf{p}$. Now we will prove that $\mathbf{V}$ has the same joint distribution as $(Z_1, \ldots, Z_r) = \mathbf{Z}$. To show this let us consider two coordinates of the vector $\mathbf{V}$

$$V_i = G_i - \sum_{l=1}^{r} G_l\sqrt{p_l}\sqrt{p_i} \quad and \quad V_j = G_j - \sum_{l=1}^{r} G_l\sqrt{p_l}\sqrt{p_j}$$

and let us compute their covariance which is equal to

$$E\left(\left(G_i - \sum_{l=1}^{r} G_l\sqrt{p_l}\sqrt{p_i}\right)\left(G_j - \sum_{l=1}^{r} G_l\sqrt{p_l}\sqrt{p_j}\right)\right).$$

since $E\left(G_i - \sum_{l=1}^{r} G_l\sqrt{p_l}\sqrt{p_i}\right) = 0$, now we have that

$$\begin{aligned}
&E\left(\left(G_i - \sum_{l=1}^{r} G_l\sqrt{p_l}\sqrt{p_i}\right)\left(G_j - \sum_{l=1}^{r} G_l\sqrt{p_l}\sqrt{p_j}\right)\right) = \\
&= E\left(G_iG_j\right) - \sum_{l=1}^{r}\sqrt{p_j}\sqrt{p_l}E\left(G_lG_i\right) - \sum_{l=1}^{r}\sqrt{p_i}\sqrt{p_l}E\left(G_lG_j\right) + \\
&+ \sum_{l\neq k}\sqrt{p_l}\sqrt{p_k}\sqrt{p_j}\sqrt{p_i}E\left(G_lG_k\right) + \sum_{l=1}^{r} p_l\sqrt{p_i}\sqrt{p_j}E\left(G_i^2\right) \\
&= -\sqrt{p_j}\sqrt{p_i} - \sqrt{p_i}\sqrt{p_j} + \sqrt{p_i}\sqrt{p_j}\sum_{l=1}^{r} p_l \\
&= -\sqrt{p_i}\sqrt{p_j}.
\end{aligned}$$

Similarly we have also that

$$E\left(\left(G_i - \sum_{l=1}^{r} G_l\sqrt{p_l}\sqrt{p_i}\right)^2\right) = 1 - p_i.$$

This proves that we have the same joint distributions between $(V_1, \ldots, V_r)$ and $(Z_1, \ldots, Z_r)$ since they both have joint multinormal distributions and the same means and covariance structure, which gives us a way to formulate the convergence from 3.2 as

$$\sum_{j=1}^{r}\left(\frac{\vartheta_j - np_j}{\sqrt{np_j}}\right)^2 \to \sum_{i=1}^{r}\left(V_i\right)^2$$

Now looking at the vector $\mathbf{V}$ since we have that $|\mathbf{p}| = 1$ is a unit vector, it means that the vector $\mathbf{W} = (\mathbf{G}\cdot\mathbf{p})\,\mathbf{p}$ is a projection of the vector $\mathbf{G}$ on the line along $\mathbf{p}$, and therefore the vector $\mathbf{V}$ will be the projection of $\mathbf{G}$ onto the plane orthogonal to $\mathbf{p}$. Let us now consider a new orthonormal coordinate system with the last basis vector equal to $\mathbf{p}$, in this new coordinate system the vector $\mathbf{G}$ will have coordinates

$$\mathbf{G}' = (G_1', \ldots, G_r') = \mathbf{G}T$$

obtained from $\mathbf{G}$ by orthogonal transformation T that maps the canonical basis into the new basis. But this means that $G_1' \ldots, G_r'$ will also be independent and identically standard normally distributed. Also it can be seen that the vector $\mathbf{V}$ will have coordinates $\left(G_1', \ldots, G_{r-1}', 0\right)$ in the new coordinate system, and therefore we will have that

$$\sum_{i=1}^{r}\left(V_i\right)^2 = \left(G_1'\right)^2 + \ldots + \left(G_{r-1}'\right)^2$$

but by definition the right hand side of the above equation has Chi-square distribution with $r-1$ degrees of freedom since G_i' have standard normal distribution. And this

ends our proof since we have that

$$\sum_{j=1}^{r}\left(\frac{\vartheta_j - np_j}{\sqrt{np_j}}\right)^2 \to \sum_{i=1}^{r}(V_i)^2 \sim \chi^2_{r-1},$$

where χ^2_{r-1} is the random variable distributed with the Chi-square distribution with $r-1$ degrees of freedom. □

Now that we proved the theorem we go back to the introduction where we defined the ϑ_j. It is obvious from the definition of the random variables ϑ_j and the definition of the multinomial distribution that all random variables ϑ_j come from a multinomial distribution. Having established this we go on to the test statistic of the χ^2-test which as we remember is

$$\sum_{j=1}^{k}\frac{(O_j - E_j)^2}{E_j} = \sum_{j=1}^{k}\frac{(O_j - Np_j)^2}{Np_j},$$

where O_j are observations from a multinomial distribution and Np_j is their expectation. Having this form we can directly apply Pearson's theorem to the χ^2-test statistic from where follows that the χ^2-test statistic converges in distribution to the Chi-Square distribution with $k-1$ degrees of freedom.

3.3 z-test for two independent proportions

Let us consider the following problem, we are given two coins c_1 and c_2 and we toss both coins 20 times. For the first coin we get 12 heads and 8 tails, for the second coin we get 6 heads and 14 tails. Now does this result imply that the coins have different probabilities of obtaining head, or is this difference from our trial due to chance only? The $z-$test for two independent proportions can be used to address this and similar problems involving two levels of a discrete random variable. Now we go on to formally define the $z-$test.

The z-test for two independent proportions is a statistical test involving proportions from two levels of a discrete independent variable. This variable may take only two discrete possible outcomes, mutually exclusive and exhaustive. The null hypothesis states that the two proportions, lets call them P_1 and P_2, are equal.

$$H_0 \ : \ P_1 = P_2$$

$$H_a \ : \ P_1 \neq P_2$$

In order for one to be able to use the $z-$test we must have randomly selected samples from two independent variables and the samples must be large enough in order to be able to use a normal approximation.

3.3.1 $z-$statistic

The z-statistic is defined as the ratio of the difference between the proportions and the standard error of the difference of the proportions [**?**]. That is,

$$z = \frac{difference\ between\ the\ proportions}{standard\ error}$$

Let us assume that we have the same situation as we examined for the χ^2 test, that is, we have two samples of sizes n_1 and n_2, the number of successes in the first sample is a, in the second sample is c and the number of failures are b and d, respectively. The estimate of the difference between two proportion is straight forward, we simply use the means for the two samples that is

$$\hat{p}_1 = \frac{a}{n_1} \qquad \hat{p}_2 = \frac{c}{n_2}$$

For the standard error of the difference we use quantity that is the square root of the sum of squares of the standard errors for the first and second sample. This is supported by the assumption that the two samples come from independent variables where then the variance is simply the sum of variances. Now the $z-$test statistics will be derived as follows

$$z = \frac{observed\ difference\ -\ expected\ difference}{SE\ for\ difference}$$

Going on to find the components for the formal definition of the $z-$test statistic we first note that under the null the expected difference $p_1 - p_2 = 0$, since $p_1 = p_2$, so we can forget about this part, the observed difference is also easy to obtain, one has only to take the difference between the two already defined estimates $\hat{p}_1 - \hat{p}_2$. The standard error will be computed as the as follows:

$$SE\ for\ difference\ = \sqrt{SE_1^2 + SE_2^2}$$

where SE_1 is the standard error for the proportion of variable one and SE_2 the standard error for the proportion of variable two. From here we calculate SE_1 and SE_2 which are

$$SE_1 = \sqrt{\frac{\hat{p}(1-\hat{p})}{n_1}} \qquad SE_2 = \sqrt{\frac{\hat{p}(1-\hat{p})}{n_2}}$$

where we have that $\hat{p}$ is the estimate of overall proportion, i.e. $\hat{p} = \frac{a+c}{n_1+n_2}$. So finally we formally combine all this to define the $z-$test statistic as

$$z = \frac{\frac{a}{n_1} - \frac{c}{n_2}}{\sqrt{\frac{\frac{a+c}{n_1+n_2}\left(1-\frac{a+c}{n_1+n_2}\right)}{n_1} + \frac{\frac{a+c}{n_1+n_2}\left(1-\frac{a+c}{n_1+n_2}\right)}{n_2}}}$$

We still have to prove that this test statistic converges to the standard normal random variable. In order to prove this we note that the standard error will converge to the standard deviation of average of the two variables, and since our nominator is the estimated mean minus the expectation, which is zero we have directly by the Central Limit Theorem that

$$\lim_{N\to\infty} \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}} \sim N(0,1)$$

3.4 Equivalence of the Chi-square test and the $z-$test

As we showed already in the previous section when having a sample from the contingency table

	X	Y	Total
A	a	b	n_1
B	c	d	n_2
	m_1	m_2	N

the test statistic for the χ^2-test is

$$\chi^2 = \frac{1}{N}\left[\frac{(aN - n_1m_1)^2}{n_1m_1} + \frac{(bN - n_1m_2)^2}{n_1m_2} + \frac{(cN - n_2m_1)^2}{n_2m_1} + \frac{(dN - n_2m_2)^2}{n_2m_2}\right]$$

For the same sample the test statistic for the $z-$test is

$$z = \frac{\frac{a}{n_1} - \frac{c}{n_2}}{\sqrt{\frac{\frac{a+c}{n_1+n_2}\left(1-\frac{a+c}{n_1+n_2}\right)}{n_1} + \frac{\frac{a+c}{n_1+n_2}\left(1-\frac{a+c}{n_1+n_2}\right)}{n_2}}}$$

which is equivalent to

$$z = \frac{\frac{an_2 - cn_1}{n_1n_2}}{\sqrt{\frac{(a+c)(b+d)}{(n_1+n_2)n_1n_2}}}$$

Now we will prove that the χ^2-statistic is equivalent to the $z-$statistic squared.

To do this we will introduce some new notation. Let us define $x_1 = \frac{a}{n_1}$, $y_1 = \frac{b}{n_1}$, $x_2 = \frac{c}{n_2}$ and $y_2 = \frac{d}{n_2}$, further we define

$$p = \frac{n_1x_1 + n_2x_2}{n_1 + n_2} \qquad q = \frac{n_1y_1 + n_2y_2}{n_1 + n_2}$$

We note that from this it follows that

$$m_1 = a + c = n_1x_1 + n_2x_2 = pN \qquad m_2 = b + d = n_1y_1 + n_2y_2 = qN$$

and we have that $q = 1 - p$. Plugging this into the formula for the χ^2 test statistic we obtain

$$\begin{aligned}
\chi^2 &= \frac{1}{N}\left[\frac{(x_1n_1N - n_1Np)^2}{n_1Np} + \frac{(y_1n_1N - n_1Nq)^2}{n_1Nq} + \frac{(x_2n_2N - n_2Np)^2}{n_2Np} + \frac{(y_2n_2N - n_2Nq)^2}{n_2Nq}\right] \\
&= \frac{1}{N}\left[\frac{n_1N(x_1 - p)^2}{p} + \frac{n_1N(y_1 - q)^2}{q} + \frac{n_2N(x_2 - p)^2}{p} + \frac{n_2N(y_2 - q)^2}{q}\right] \\
&= n_1\left[\frac{(x_1 - p)^2}{p} + \frac{(y_1 - q)^2}{q}\right] + n_2\left[\frac{(x_2 - p)^2}{p} + \frac{(y_2 - q)^2}{q}\right] \\
&= \frac{n_1(x_1 - p)^2(1 - p) + n_1(1 - x_1 - 1 + p)^2p + n_2(x_2 - p)^2(1 - p) + n_2(1 - x_2 - 1 + p)^2p}{pq} \\
&= \frac{n_1(x_1 - p)^2(1 - p) + n_1(p - x_1)^2p + n_2(x_2 - p)^2(1 - p) + n_2(p - x_2)^2p}{pq} \\
&= \frac{\left[n_1(x_1 - p)^2\right][(1 - p) + p] + \left[n_2(x_2 - p)^2\right][(1 - p) + p]}{pq} \\
&= \frac{n_1(p - x_1)^2 + n_2(p - x_2)^2}{pq}
\end{aligned}$$

Now plugging $p = \frac{n_1x_1 + n_2x_2}{n_1 + n_2}$ into the transformed test statistics we obtain

$$\begin{aligned}
\chi^2 &= \frac{n_1\left(\frac{n_2x_1 - x_2x_2}{n_1 + n_2}\right)^2 + n_2\left(\frac{n_1x_2 - n_1x_1}{n_1 + n_2}\right)^2}{pq} \\
&= \frac{(x_1 - x_2)^2(n_1^2n_2 - n_1n_2^2)}{pqN^2} \\
&= \frac{(x_1 - x_2)^2}{pq\frac{(n_1 + n_2)^2}{(n_1^2n_2 - n_1n_2^2)}} = \frac{(x_1 - x_2)^2}{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}
\end{aligned}$$

Now substituting back to the standard notation we obtain

$$\chi^2 = \frac{\left(\frac{a}{n_1} - \frac{c}{n_2}\right)^2}{\left(\frac{a+c}{n_1+n_2}\right)\left(\frac{b+d}{n_1+n_2}\right)\left(\frac{n_1+n_2}{n_1n_2}\right)} = z^2$$

which is exactly what we needed.

4 Fisher's Exact Test

Fisher's Exact Test, named after its inventor Ronald Fisher, is a statistical significance test that is most commonly used in the analysis of contingency tables, that is in the analysis of 2×2 contingency tables. The test is in practice employed mostly with small sample sizes even though it is valid for all sample sizes. Fisher's exact test determines whether there is a statistically significant association between two categorical variables. While examining Fisher's test we will assume that we are given two discrete binomial random variables from which one tells us whether an individual out of the population is in category A or B and the other tells us whether one individual is in category 1 or category 2. The setup of Fisher's exact test when having a population of size n and taking a sample of total size N will look as follows

- n_1 of the subjects belong to category A and n_2 belong to category B, such that $n_1 + n_2 = N$
- m_1 of the subjects belong to category 1 and m_2 belong to category 2, such that $m_1 + m_2 = N$
- There are a subject of category A that belong to category 1 and b subject of category A belonging to category 2, such that $a + b = n_1$
- There are c subject of category B that belong to category 1 and d subject of category B belonging to category 2, such that $c + d = n_2$
- Similarly it holds that $a + c = m_1$ and $b + d = m_2$

The above stated setup for Fisher's test can be put in a contingency table, given below,

	Category 1	Category 2	Total
Category A	a	b	n_1
Category B	c	d	n_2
Total	m_1	m_2	N

Table 2: Contingency table

Once we are given the table with the values we fix the marginal values $m_1,\ m_2$ and $n_1,\ n_2$ and then we can calculate the exact probability of the table occurring among

all tables with the same marginal sums. Calculating this probability relies heavily on the hypergeometric distribution which we defined previously. In our case, we have that in the finite population of size n, n_1 subjects are of one type and n_2 of the other type, where each subject is either one type or the other. Also, we need to choose a particular number of subjects of each type, meaning that the probability of the particular table occurring has hypergeometric distribition. From here we can directly compute the probability of a given contingency table occurring which is defined as

Definition 4.1. Probability of a given contingency table

$$p_t = P(a = t) = \frac{\binom{n_1}{t}\binom{n_2}{m_1 - t}}{\binom{n}{m_1}}$$

for $\max\{0, n_1 + m_1 - n\} \leq t \leq \min\{n_1, m_1\}$

We note that a 2×2 contingency table with both margins fixed is completely determined by one of its elements, that is if we choose a value for any of the elements a, b, c, d of the table, all the other value will be determined due to the fixed margins. Thus it follows that the probability of a given table occurring, p_t, can be calculated as the probability of the element a being equal to a given value t.
The $p-$value of the test is calculated as the sum of the p_t values of the contingency tables that give stronger evidence in favour of H_a.

$$P - value = P(a \geq t) = \sum_{t_0 \geq t} p_{t_0} = \sum_{t_0 \geq t} \frac{\binom{n_1}{t_0}\binom{n_2}{m_1 - t_0}}{\binom{n}{m_1}}$$

4.1 Two-Sided P-Values

In the case when we test the null hypothesis against an alternative hypothesis which is two-sided, we have a few different approaches how to calculate the $p-$value [1].

- $P = P\left[p_a \leq p_{t_0}\right]$ for the observed value t_0.
- another possibility is

  $$P = P\left[|a - E(a)| \geq |t_0 - E(a)|\right]$$

  where the hypergeometric $E(a) = m_1 n_1 / n$.
- $P = min\left[P(a \geq t_0), P(a \leq t_0)\right]$ plus an attainable probability in the other tail that is close as possible to, but not greater than one-tailed probability.

We will omit to deal with two sided $p-$values in our work and concentrate only on one sided $p-$value. The only part in which we use two sided $p-$value is in our simulation study where we are analysing the data using the **R** language which gives two sided $p-$values as a result, but the computation is based on multiplying the one sided $p-$value with two and thus being equivalent to computing one sided $p-$values.

4.2 Tea-tasting Experiment

If we go back to the previously defined Lady Tasting Tea experiment we can analyse this experiment with the Fisher's Exact Test. Firstly, we observe that there are 4 cups with milk poured first and 4 with tea poured first, therefore one of the margins is fixed. Next, we also know, that this information was given to the lady, so she knew that she had to choose 4 cups of each category. Therefore, the other margin is fixed. So, we need to calculate the p_t values for all tables with all margins equal to 4.

	Guess poured first		
Poured first	Milk	Tea	Total
Milk	a	b	4
Tea	c	d	4
Total	4	4	8

Table 3: Contingency table for the tea-testing experiment

The variable a may take the values $0, 1, 2, 3, 4$, thus we will have 5 possible contigency tables. We will calculate the p_t for $t \in \{0, 1, 2, 3, 4\}$.

$$p_4 = \frac{\binom{4}{4}\binom{4}{0}}{\binom{8}{4}} = \frac{1}{70} = 0.014$$

$$p_3 = \frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}} = \frac{32}{70} = 0.229$$

$$p_2 = \frac{\binom{4}{2}\binom{4}{2}}{\binom{8}{4}} = \frac{36}{70} = 0.514$$

$$p_1 = \frac{\binom{4}{1}\binom{4}{3}}{\binom{8}{4}} = \frac{32}{70} = 0.229$$

$$p_0 = \frac{\binom{4}{0}\binom{4}{4}}{\binom{8}{4}} = \frac{1}{70} = 0.014$$

Also, it is easy to see that their sum is equal to 1, which is what we expect from a distribution.

$$\begin{aligned} \sum_{i=0}^{4} p_i &= p_0 + p_1 + p_2 + p_3 + p_4 \\ &= 0.014 + 0.229 + 0.514 + 0.229 + 0.014 \\ &= 1 \end{aligned}$$

Now, lets calculate the $p-$value for each of the possible contingency tables:

- the $p-$value for the table with $a = 4$ is 0.014,
- for $a = 3$ is 0.014+0.229=0.243,
- for $a = 2$ it is 0.014+0.229+0.514=0.757,
- for $a = 1$ it is 0.014+0.229+0.514+0.229=0.986
- for $a = 0$ it is 0.014+0.229+0.514+0.229+0.014 =1.

We already said that we have statistically significant results only when $p < \alpha$, and since our predefined value of α is 0.05, the only such $p-$value is 0.014, which is the only case when we can reject the H_0, when the lady successfully guesses all four cups she chooses.

The possible reasons behind the inexactness of the Fisher's Exact Test are the following:

- discrete null distribution, because the hypergeometric distribution is very discrete and the p_t values can take only a few values in the interval $(0, 1)$ such that the sum of the all such values is 1,
- conditioning on both margins, which is a fact we already mentioned. Actually, if we condition on one margin only, we get more possible cases making the statistic less discrete. In the most extreme case, if we do not condition on any of the margins we will not get enough information to run a test [3].

5 Simulations

Definition 5.1. Size of a test is the probability of incorrectly rejecting the null hypothesis. In other words, it is the effective significance level of a test, denoted by α_e.

Definition 5.2. A conservative test is a statistical test for which the true probability of incorrectly rejecting the null hypothesis is never greater than the predefined nominal level.

5.1 Inexactness of the Fisher's Exact Test

As we already mentioned we will use simulations to show that the Fisher's Exact test is conservative and that for small sample sizes, for which is mostly used, the effective significance level never reaches the nominal level. The nominal significance level in our simulations is chosen to be $\alpha = 0.05$. We simulate the test for different sample sizes, denoted by N and for different probabilities in the binomial samples that we use, denoted by p. The values for the sample sizes that we use are $(5, 10, 15, 20, 25, 30, 50, 75, 100)$ and each of them was tested for different p, probabilities in the binomial distribution from where we generate the sample. Each step was of the simulation was run 10000 times. **R** language for statistical computation was used to conduct the analysis [18].

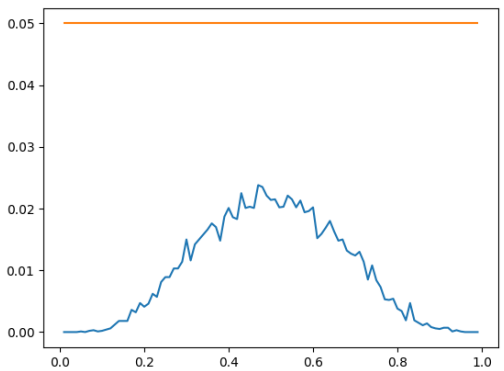

Figure 1: Sample size $N = 5$

The first case was for a small sample size of only 5. In this case we generated a curve showing how the effective rate behaves as we change the value of the p. We can easily see that the effective level is much smaller than the expected nominal value of 0.05. The highest effective rate we get is 0.0238 for the value $p = 0.47$ and in such a small sample this test is really conservative, it never reaches any value near the nominal 0.05.

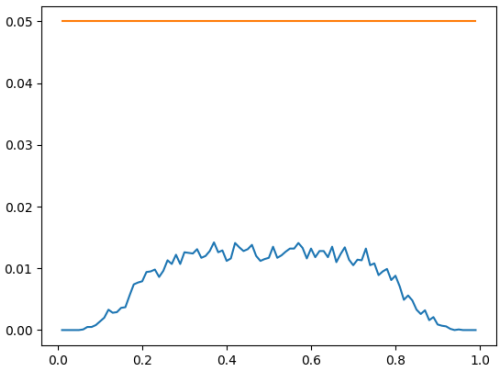

Figure 2: Sample size $N = 10$

The second case we considered was taking a sample of size 10. Similarly as before, we generated a curve showing how the effective rate behaves as we change the value of the p. We can easily see that the effective level is in a very small range around 0.012, which is much below the expected nominal value of 0.05. The highest effective rate we get is 0.0141 for the value $p = 0.58$ and again, in such a small sample this test is really conservative, it never reaches any value near the nominal 0.05.

p	0.1	0.25	0.4	0.5	0.8	0.95
α_e	0.0055	0.0171	0.0165	0.0157	0.0173	0.0009

Table 4: Sample size $N = 15$

Then we took slightly larger sample of 15, for which we only considered a few different values of p, namely the values 0.1,0.25,0.4,0.5,0.8 and 0.95. Analysing the table above we can again see that the effective rate again stays very small, and the highest value we produced, 0.0173, is much smaller then our given nominal level of 0.05.

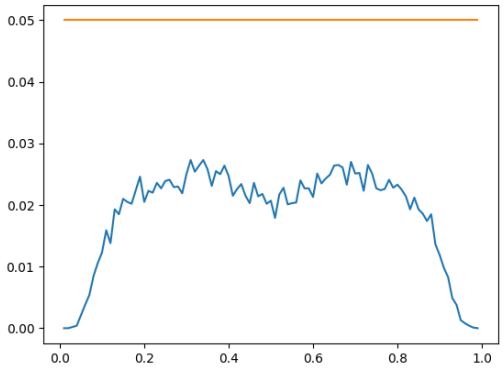

Figure 3: Sample size $N = 20$

Next, we considered a sample of size 20. Similarly as before, we generated a curve showing how the effective rate behaves as we change the value of the p. We can easily see that the effective level is in a very small range around 0.025, which is below the expected nominal value of 0.05. The highest effective rate we get is 0.0273 for the value $p = 0.31$ and again, in such a small sample this test is really conservative, it never reaches any value near the nominal 0.05.

p	0.1	0.25	0.4	0.5	0.8	0.95
α_e	0.0073	0.0221	0.0301	0.0332	0.0207	0.0011

Table 5: Sample size $N = 25$

Then we took slightly larger sample of 25, for which we only considered a few different values of p, namely the values 0.1,0.25,0.4,0.5,0.8 and 0.95. Analysing the table above we can again see that the effective rate again stays very small, and the highest value we produced, 0.0332, is greater then the values we had before, but again not close enough to our given nominal level of 0.05.

p	0.1	0.25	0.4	0.5	0.8	0.95
α_e	0.0086	0.0248	0.023	0.0253	0.0238	0.0014

Table 6: Sample size $N = 30$

As we had a trend of getting greater α_e we increased the sample a little bit, and considered a sample of size 30. Again, we used a few different values of p, the values 0.1,0.25,0.4,0.5,0.8 and 0.95. Analysing the table above we can again see that the effective rate again stays very small, and the highest value we produced, 0.0253, is less

then the values we had before for the sample of 25, so there is no trend of increasing the α_e as we increase sample size, when considering small samples.

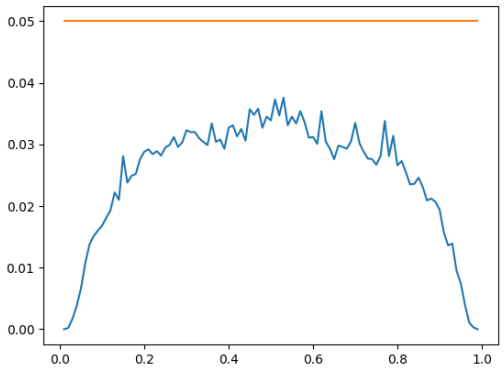

Figure 4: Sample size $N = 50$

Next, we considered a sample of size 50. Again we generated a curve showing how the effective rate behaves as we change the value of the p. We can easily see that the effective level is in a very small range around 0.03, which is below the expected nominal value of 0.05. The highest effective rate we get is 0.0358 for the value $p = 0.47$ and again, in such a small sample this test is really conservative, it never reaches any value near the nominal 0.05.

p	0.1	0.25	0.4	0.5	0.8	0.95
α_e	0.0282	0.0321	0.0378	0.0457	0.0286	0.0152

Table 7: Sample size $N = 75$

Then we consider a sample of size of 75. We did the simulations for a few values of 5, 0.1,0.25,0.4,0.5,0.8 and 0.95, and we got better results then before. In this case the α_e for $p = 0.1$, which is quite an extreme case, is 0.0282 which is greater then all such values for smaller samples. For $p = 0.5$ we have $\alpha_e = 0.0457$ which is very close to 0.05, but it will never reach it, because as p will increase the α_e will decrease.

p	0.1	0.25	0.4	0.5	0.8	0.95
α_e	0.0276	0.0359	0.035	0.0422	0.0297	0.0219

Table 8: Sample size $N = 100$

Next, we increase the sample size to 100. In this case we have greater α_e for $p = 0.25$, $p = 0.8$ and $p = 0.95$, so all extreme cases that are very close, then in the previous case,

but in the other cases this does not happen. For $p = 0.05$ we are getting smaller α_e, of 0.0422. So, again the test is conservative and does not reach that close the nominal level.

p	0.1	0.25	0.4	0.5	0.8	0.95
α_e	0.02986	0.0395	0.0408	0.0426	0.0343	0.0325

Table 9: Sample size $N = 200$

Now, we consider a sample of size 200. From the table, we can easily see that the results are getting closer to the nominal level. Even thought we twice increased the sample size, from 100 to 200, α_e for $p = 0.05$ changes just a little bit, from 0.0422 to 0.0426. However, all the α_e got higher and closer to 0.05, but never 0.05.

p	0.1	0.25	0.4	0.5	0.8	0.95
α_e	0.0387	0.045	0.0452	0.04832	0.04267	0.0347

Table 10: Sample size $N = 500$

Finally, we consider a sample of size 500. Looking at the table with our α_e it is easy to see that almost all α_e are quite close to 0.05 then in the other case for smaller sample size. α_e even reaches 0.04832 which is the closest we get to 0.05.

Comparing all these results, shows us that the Fisher's Exact Test is indeed conservative. Analysing the simulations, we can see that as we increase the sample sizes the effective rate gets bigger, but not big enough to reach the nominal level of 0.05. However, this trend is not consistent for small samples, but is what we expect from a conservative test. Also, it is easy to see that if the binomial samples are having value p close to 0 or 1, the test works much worse then it works for values of p near 0.5. In all our cases, we got the highest α_e in a small range around the $p = 0.5$. We can expect that as the sample size increases the Fisher's test will work better, but only when we have infinite sample we may reach to have equal nominal and effective rate.

All in all, even that the test is said to be exact, it is conservative for small sample sizes, for which it is mostly used and is believed to be very exact. It incorrectly rejects the null hypothesis much less that the nominal vale which is what we expected to see.

5.2 Comparison between the Fisher's Exact Test and the χ^2 Test

When we want to distinguish which test to use for a specific case, usually it is said to use the Fisher' Exact Test for small samples and use the χ^2 Test for large samples. However, we already saw that the Fisher's Test is very conservative for small samples, and that even for larger samples is stays quite conservative. For the usage of the χ^2 test it is usually said that the only restriction we need to consider is that the expected number of cases should exceed 5 in most cells of the contingency table, because the test statistic only asymptotically converges to the χ^2 distribution. Therefore it is advised to use it for large sample sizes, while for small sample sizes to use the Fisher's test.

That is why we ran simulations for different sample sizes, and compared the effective rates for the Fisher's and the χ^2 test. We took sample sizes of 30, 50, 100 and 200.

In the first case, the sample size was 30, and we got very interesting results. Almost everywhere the α_e for both tests was the same. This is quite opposite of what we expected, because the Fisher's test was expected to perform better, but it performed exactly as the χ^2 test. In the table some of the values of α_e for the tests is given and we can easily see that those values are indeed the same for both tests. Also, in the appendix all the values can be seen that we got from the simulations.

p	0.1	0.25	0.4	0.5	0.8	0.95
Fisher's effective rate	0.0103	0.0269	0.0280	0.0287	0.0255	0.0012
Chi-squared effective rate	0.0103	0.0269	0.0280	0.0287	0.0255	0.0012

Table 11: Sample size $N = 30$

In the second case, the sample size was 50, and we got similar results as before. Almost everywhere the α_e for both tests was the same. Again, the Fisher's test performed exactly as the χ^2 test. In the table some of the values of α_e for the tests is given and we can easily see that those values are indeed the same for both tests. Also, in the appendix all the values can be seen that we got from the simulations.

p	0.1	0.25	0.4	0.5	0.8	0.95
Fisher's effective rate	0.0163	0.0307	0.0297	0.0391	0.0289	0.0081
Chi-squared effective rate	0.0163	0.0307	0.0297	0.0391	0.0289	0.0081

Table 12: Sample size $N = 50$

Then we performed simulations for bigger samples, of 100 and 200. In both cases we got similar results, the α_e for both cases are again very close. This is an expected result, because we expect that for large sample sizes he both test to work very similar. In the case of sample of size 100, there is just a small difference for values of p on the boundaries, when it is very close to 0 or to 1. And, in the other case, when the sample size is 200, the α_e matches for both of the test almost everywhere, even at the boundaries of the value of p.

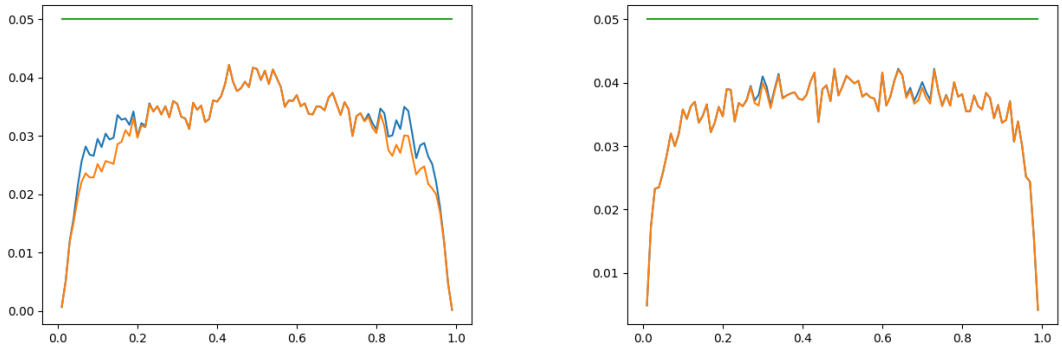

Figure 5: Samples of sizes $N = 100$ and $N = 200$

All in all, we can conclude that both tests, Fisher's Exact Test and the χ^2 test perform similarly, both for small and large sample sizes. Also, they are both very conservative, and produce values of α_e much smaller then the nominal level, $\alpha = 0.05$.

6 Conclusion

Our goal was to see the inexact behaviour of the Fisher's exact test and see its behaviour compared to the other tests used instead of Fisher's test.

Firstly, we elaborately explained the concept of hypothesis testing, what categorical variable is, the 2×2 contingency table we used and included some important results about hypothesis testing. One of them is the proof that under the null hypothesis the distribution of the $p-$values is uniform on the interval $(0, 1)$ and the Neyman-Pearson Fundamental Lemma about uniformly most powerful tests.

For testing the association between two binary categorical variables, a few well known tests are used. One of them is the χ^2-test, for which we showed how to derive the test statistic in the general case for $r \times c$ contingency tables. Then using the Person's Theorem we showed that it asymptotically converges to the χ^2 distribution. Afterwards, we focused on the special case of 2×2 contingency tables, used for variables having only two levels. In this case we proved that the χ^2 test statistic is equal to the square of the z-test statistic for two proportions, from which we concluded that the χ^2 test for 2×2 contingency tables is equivalent to the $z-$test for testing the equality of two independent proportions. The $z-$test is another statistical test involving proportions from two levels of a discrete independent variable. This variable may take only two discrete possible outcomes, mutually exclusive and exhaustive.

Next, we had Fisher's exact test, as one of the most commonly used methods for testing the association between two categorical variables. The historical background of this test is very interesting, as Fisher came up with this idea because he doubted that one of his friends can distinguish whether in a cup of tea the tea or the milk was poured first. Therefore, we also included the historical background of the test and the theory behind the well known Lady Tasting Tea experiment. Further, we explained how the Fisher's exact test is based on the hypergeometric discrete distribution of the predefined test statistic. We included an example how we can perform the test, and how different test statistics are defined, based whether we have one-sided or two-sided test.

At the end, after defining all the theoretical background, we did simulations and showed the inexactness of Fisher's test. While the χ^2-test is known to be only approximate test, i.e. its size is equal to the nominal level when n goes to infinity, we show that also the Fisher's exact test is inexact with small samples using 2×2 contingency tables. The Fisher's exact test is said to be a test for which the nominal level is equal to the effective type I error, but that is not the case. It is actually a conservative test, meaning that the true probability of incorrectly rejecting the null hypothesis is never greater or equal to the nominal level. There are a few reasons for its conservatism, such as the discrete null distribution, the usage of a fixed nominal level and the conditioning on two margins. We showed this in the analysis of our simulation study. We also wanted to see how Fisher's test behaves with comparison to the χ^2 test, and even though we expected better results from Fisher's test we actually got the same effective rates for both tests, even thought we expected that Fisher's test will work better for small samples, as it is usually advised to be used in such cases. To conclude with, both Fisher's exact test and the asymptotic χ^2-test are both conservative statistical test, giving very similar results both for small and large samples.

7 Povzetek naloge v slovenskem jeziku

Testiranje hipotez je zelo pomemben koncept v statistiki. Statistična hipoteza je narejena na podlagi opazovanj, ki jih imamo in nato odvisno od tega kakšne podatke imamo, uporabljamo poseben test. Ničelna hipoteza je postavljena a na takšen način, da jo želimo zavrnuti, ne pa sprejeti. Testiranje hipotez ima veliko različnih aplikacij, eden od njih pa preverja, ali obstaja povezava med kategoričnimi spremenljivkami. Kategorična spremenljivka je spremenljivka, ki je določena z nivoji ali kategorijami. V našem primeru nas je zanimalo testiranje povezave med dvema kategoričnima spremenljivkama. Na primer, želimo vedeti, ali obstaja povezava med deležem ljudi, ki kadijo v ženski in moški populaciji. Torej, oseba je bodisi kadilec ali ne, bodisi moški ali ženska, kar nam daje dve kategorični spremenljivki. Pojasnili smo koncept testiranja hipotez, kaj je kategorični datum, kaj je tabela kontigentnosti 2×2, ki smo jo uporabili, in vključili smo nekaj pomembnih rezultatov testiranja hipotez. Eden od njih je dokaz, da je pod ničelno hipotezo porazdelitev $p-$vrednosti enakomerna na intervalu $(0, 1)$ in Neyman-Pearsonova Fundamentala Lema o enakomerno najmočnejših testih. Za testiranje povezave dveh binarnih kategoričnih spremenljivk se uporablja nekaj dobro znanih testov. Točen test Fisherja je ena od najpogostejše uporabljenih metod za testiranje povezave med dvema kategoričnima spremenljivkama. Zgodovinsko ozadje tega testa je zelo zanimivo, saj je Fisher prišel do te ideje, ker je dvomil, da lahko eden od njegovih prijateljev razlikuje, ali je v skodelici čaja najprej nalit čaj ali mleko. Zato smo vključili tudi zgodovinsko ozadje testa in teorijo v ozadju zelo znanega Lady Tasting Tea testa. Temelji na hipergeometrični diskretni porazdelitvi vnaprej določene testne statistike. Drugi test je χ^2 test, za katerega smo pokazali, kako izpeljati testno statistiko v splošnem primeru za $r \times c$ kontingenčno tabelo. Potem smo z uporabo Personeva izreka pokazali, da asimptotsko konvergira proti χ^2 porazdelitvi. Nato smo se osredotočili na poseben primer 2×2 kontingenčne tabele, ki so uporablja za spremenljivke, ki imajo le dva nivoja. V tem primeru smo dokazali, da je χ^2 statistika enaka kvadratu statistike $z-$testa za dva deleža, iz katerega smo ugotovili, da je χ^2 test za 2×2 kontingenčne tabele enakovreden z-testu za testiranje enakosti dva neodvisna deleža. Z-test je še en statistični test, ki vključuje deleže dva nivoja diskretne

neodvisne spremenljivke. Ta spremenljivka lahko vključuje le dva ločena možna izida, medsebojno izključujoči in izčrpni. Medtem, ko sta zadnja dva testa znana le kot približna testa, oziroma njihova velikost je enaka nominalnem nivoju, ko n gre proti neskončno, smo mi pokazali, da je tudi Fisherjev natančen test nenatančen z majhnimi vzorci z uporabo 2×2 kontingenčnih tabel. Točen test Fisherja naj bi bil točen test, kar pomeni, da je statistični test, pri katerem je nominalni nivo enak napaki tipa I, vendar temu ni tako. Pravzaprav je konzervativni test, kar pomeni, da prava verjetnost nepravilne zavrnitve ničelne hipoteze nikoli ni večja ali enaka nominalnem nivoju. Obstaja nekaj razlogov za njegov konservatizem, kot je diskretna ničelna porazdelitev, uporaba točneg nominalnega nivoja pogojevano na dveh robovih. Prav tako smo želeli videti, kako se Fisherjev test obnaša v primerjavi s testom χ^2, in čeprav smo pričakovali boljše rezultate iz Fisherjevega testa, smo dejansko dobili enake efektivne stopnje za oba testa, čeprav smo pričakovali, da bo Fisherjev test boljše deloval pri majhnih vzorcih, saj je v takšnih primerih priporočljivo, da se uporablja. Da zaključimo, Fisherjev natančen test in asimptotski χ^2 test sta oba konzervativna statistična testa, ki dajeta zelo podobne rezultate tako za majhne kot velike vzorce.

8 Bibliography

[1] A. Agresti, *Categorical Data Analysis*, Wiley-Interscience, Second Edition, 2003. *(Cited on pages 4 and 22.)*

[2] V. Bentkus, A Lyapunov type bound in $\mathbb{R}^d$. *Rossiiskaya Akademiya Nauk. Teoriya Veroyatnostei i ee Primeneniya* 49 (2004) 400–410. *(Cited on page 10.)*

[3] J. Berkson, In Dispraise of the Exact Test. *Journal of Statistical Planning and Inference* 2 (1978) 27–42. *(Cited on page 24.)*

[4] R. D'Agostino, W. Chase, and A. Belanger, The Appropriateness of Some Common Procedures for Testing the Equality of Two Independent Binomial Populations. *The American Statistician* 42 (1988) 198–202. *(Not cited.)*

[5] J. De Muth, *Basic Statistics and Pharmaceutical Statistical Applications*, Chapman and Hall/CRC, Third Edition, 2014. *(Not cited.)*

[6] R. Fisher, Mathematics of a Lady Tasting Tea. *The World of Mathematics* 3 (1956) 1514–1521. *(Cited on page 2.)*

[7] R. Fisher, *The Design of Experiments*, Oliver and Boyd, First Edition, 1935. *(Cited on page 1.)*

[8] P. Good, *Permutation, Parametric and Bootstrap Tests of Hypotheses*, Springer-Verlag, New York, Third Edition, 2005. *(Not cited.)*

[9] P. Good, *Permutation Tests, A Practical Guide to Resampling Methods for Testing Hypotheses*, Springer-Verlag, New York, 1994. *(Not cited.)*

[10] B. Klartag, A central limit theorem for convex sets. *Inventiones mathematicae* 168 (2007) 91–131. *(Cited on page 10.)*

[11] E. Lehmann and J. Romano, *Testing Statistical Hypotheses*. Springer-Verlag, New York, Third Edition, 2005. *(Cited on page 3.)*

[12] R. Little, Testing the Equality of Two Independent Binomial Proportions. *The American Statistician* 43 (1989) 283–288. *(Not cited.)*

[13] S. Lydersen, M. Fagerland, and P. Laake, Tutorial in Biostatistic, Recommended tests for association in 2×2 tables. *Statistic in Medicine* 28 (2009) 1159–1175. *(Not cited.)*

[14] K. Mardia, J. Kent, and J. Bibby, *Multivariate Analysis*. Academic Press, London, 1979. *(Cited on page 14.)*

[15] W. Rudin, *Real and Complex Analysis*, McGraw-Hill International, Third Edition 1987. *(Cited on page 10.)*

[16] G. Upton, Fisher's Exact Test. *Journal of the Royal Statistical Society* 155 (1992) 395–402. *(Not cited.)*

[17] F. Yates, Contingency Tables Involving Small Numbers and the χ^2 Test. *Journal of the Royal Statistical Society* 1 (1934) 217–235. *(Cited on page 11.)*

[18] *The R Project for Statistical Computing*, `https://www.r-project.org/`. (Viewed on: 06/08/2017.) *(Cited on page 25.)*

Appendices

A Tables for comparison of α_e for Fisher's and χ^2 test

Sample size $N = 30$

p	α_e for Fisher's test	α_e for χ^2 test	p	α_e for Fisher's test	α_e for χ^2 test
0.01	0.0000	0.0000	0.51	0.0238	0.0238
0.02	0.0000	0.0000	0.52	0.0284	0.0284
0.03	0.0002	0.0002	0.53	0.0264	0.0264
0.04	0.0005	0.0005	0.54	0.0242	0.0242
0.05	0.0020	0.0020	0.55	0.0270	0.0270
0.06	0.0024	0.0024	0.56	0.0251	0.0251
0.07	0.0049	0.0049	0.57	0.0276	0.0276
0.08	0.0064	0.0064	0.58	0.0229	0.0229
0.09	0.0080	0.0080	0.59	0.0242	0.0242
0.10	0.0103	0.0103	0.60	0.0257	0.0257
0.11	0.0112	0.0112	0.61	0.0240	0.0240
0.12	0.0126	0.0126	0.62	0.0257	0.0257
0.13	0.0130	0.0130	0.63	0.0273	0.0273
0.14	0.0153	0.0153	0.64	0.0266	0.0266
0.15	0.0178	0.0178	0.65	0.0253	0.0253
0.16	0.0203	0.0203	0.66	0.0299	0.0299
0.17	0.0190	0.0190	0.67	0.0263	0.0263
0.18	0.0191	0.0191	0.68	0.0281	0.0281
0.19	0.0228	0.0228	0.69	0.0266	0.0266
0.20	0.0212	0.0212	0.70	0.0284	0.0284
0.21	0.0204	0.0204	0.71	0.0242	0.0242
0.22	0.0239	0.0239	0.72	0.0256	0.0256
0.23	0.0266	0.0266	0.73	0.0245	0.0245
0.24	0.0257	0.0257	0.74	0.0279	0.0279
0.25	0.0269	0.0269	0.75	0.0252	0.0252
0.26	0.0257	0.0257	0.76	0.0268	0.0268
0.27	0.0263	0.0263	0.77	0.0251	0.0251
0.28	0.0220	0.0220	0.78	0.0255	0.0255
0.29	0.0233	0.0233	0.79	0.0219	0.0219
0.30	0.0287	0.0287	0.80	0.0255	0.0255
0.31	0.0257	0.0257	0.81	0.0214	0.0214
0.32	0.0267	0.0267	0.82	0.0206	0.0206
0.33	0.0231	0.0231	0.83	0.0183	0.0183
0.34	0.0281	0.0281	0.84	0.0179	0.0179
0.35	0.0270	0.0270	0.85	0.0162	0.0162
0.36	0.0256	0.0256	0.86	0.0139	0.0139
0.37	0.0242	0.0242	0.87	0.0160	0.0160
0.38	0.0269	0.0269	0.88	0.0146	0.0146
0.39	0.0245	0.0245	0.89	0.0138	0.0138
0.40	0.0280	0.0280	0.90	0.0108	0.0108
0.41	0.0251	0.0251	0.91	0.0068	0.0068
0.42	0.0289	0.0289	0.92	0.0069	0.0069
0.43	0.0271	0.0271	0.93	0.0051	0.0051
0.44	0.0245	0.0245	0.94	0.0025	0.0025
0.45	0.0242	0.0242	0.95	0.0012	0.0012
0.46	0.0262	0.0262	0.96	0.0001	0.0001
0.47	0.0301	0.0301	0.97	0.0002	0.0002
0.48	0.0312	0.0312	0.98	0.0000	0.0000
0.49	0.0288	0.0288	0.99	0.0000	0.0000
0.50	0.0287	0.0287	-	-	-

Sample size $N = 50$

p	α_e for Fisher's test	α_e for χ^2 test	p	α_e for Fisher's test	α_e for χ^2 test
0.01	0.0000	0.0000	0.51	0.0336	0.0336
0.02	0.0003	0.0003	0.52	0.0340	0.0340
0.03	0.0019	0.0019	0.53	0.0315	0.0315
0.04	0.0043	0.0043	0.54	0.0344	0.0344
0.05	0.0075	0.0075	0.55	0.0368	0.0368
0.06	0.0117	0.0117	0.56	0.0351	0.0351
0.07	0.0124	0.0124	0.57	0.0337	0.0337
0.08	0.0146	0.0146	0.58	0.0304	0.0304
0.09	0.0161	0.0161	0.59	0.0349	0.0349
0.10	0.0163	0.0163	0.60	0.0319	0.0319
0.11	0.0182	0.0182	0.61	0.0301	0.0301
0.12	0.0201	0.0201	0.62	0.0356	0.0356
0.13	0.0209	0.0209	0.63	0.0358	0.0358
0.14	0.0236	0.0236	0.64	0.0319	0.0319
0.15	0.0234	0.0234	0.65	0.0298	0.0298
0.16	0.0236	0.0236	0.66	0.0285	0.0285
0.17	0.0262	0.0262	0.67	0.0321	0.0321
0.18	0.0245	0.0245	0.68	0.0321	0.0321
0.19	0.0251	0.0251	0.69	0.0264	0.0264
0.20	0.0266	0.0266	0.70	0.0320	0.0320
0.21	0.0282	0.0282	0.71	0.0338	0.0338
0.22	0.0328	0.0328	0.72	0.0329	0.0329
0.23	0.0268	0.0268	0.73	0.0302	0.0302
0.24	0.0308	0.0308	0.74	0.0306	0.0306
0.25	0.0307	0.0307	0.75	0.0312	0.0312
0.26	0.0279	0.0279	0.76	0.0302	0.0302
0.27	0.0292	0.0292	0.77	0.0290	0.0290
0.28	0.0288	0.0288	0.78	0.0319	0.0319
0.29	0.0275	0.0275	0.79	0.0291	0.0291
0.30	0.0307	0.0307	0.80	0.0289	0.0289
0.31	0.0315	0.0315	0.81	0.0289	0.0289
0.32	0.0287	0.0287	0.82	0.0264	0.0264
0.33	0.0311	0.0311	0.83	0.0251	0.0251
0.34	0.0307	0.0307	0.84	0.0248	0.0248
0.35	0.0296	0.0296	0.85	0.0236	0.0236
0.36	0.0282	0.0282	0.86	0.0225	0.0225
0.37	0.0298	0.0298	0.87	0.0206	0.0206
0.38	0.0268	0.0268	0.88	0.0193	0.0193
0.39	0.0320	0.0320	0.89	0.0191	0.0191
0.40	0.0297	0.0297	0.90	0.0167	0.0167
0.41	0.0356	0.0356	0.91	0.0169	0.0169
0.42	0.0336	0.0336	0.92	0.0153	0.0153
0.43	0.0335	0.0335	0.93	0.0118	0.0118
0.44	0.0335	0.0335	0.94	0.0103	0.0103
0.45	0.0318	0.0318	0.95	0.0081	0.0081
0.46	0.0365	0.0365	0.96	0.0040	0.0040
0.47	0.0327	0.0327	0.97	0.0026	0.0026
0.48	0.0377	0.0377	0.98	0.0001	0.0001
0.49	0.0343	0.0343	0.99	0.0000	0.0000
0.50	0.0391	0.0391	-	-	-