Exercise 3 Report

Ze Hong Zhou

13/10/2021

Contents

1 Chi-Squared Tests

A chi-squared test is a statistical test wherein the test statistic is distributed as the chi-squared distribution under the null hypothesis [5]. This report will cover the most common type, Pearson's chi-squared test.

1.1 The Chi-Squared Distribution Family

The chi-squared distribution family is a family of continuous distributions. This family includes all distributions of the sum of a positive integer number of independent standard normal random variables. The number of random variables summed of a member in this family is its number of degrees of freedom, k, the only parameter of chi-squared distributions. This is illustrated in relation (1).

Suppose that for all $i \in [1, k]$, $Z_i \sim N(0, 1)$. Then,

$$\sum_{i=1}^{k} (Z_i^2) \sim \chi_k^2 \tag{1}$$

Where χ_k^2 is the chi-squared distribution with k degrees of freedom.

The probability density function for chi-squared distributions can be seen in figure 1.
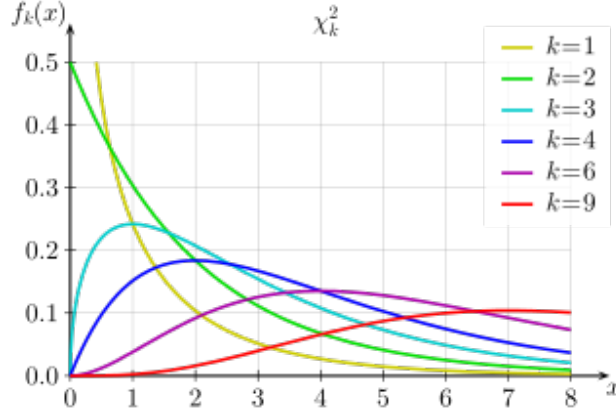
Figure 1: probability density functions for various k. Source: [4].

1.2 Pearson's Chi-Squared Test

Suppose that there are k mutually exclusive events with individual probabilities p_i, which sum to 1.

$$\sum_{i=1}^{k}(p_i) = 1$$

Suppose that n independent trials were taken and the number of times each event occurred (n_i) was recorded. In 1900, Pearson proposed the following test statistic to test a set of hypothesised values for p_i [3].

$$X^2 = \sum_{i=1}^{k}\left(\frac{(n_i - np_i)^2}{np_i}\right) \tag{2}$$

Pearson proved that as the number of trials approached infinity, if all $p_i > 0$, the distribution of his test statistic approached the chi-squared distribution with $k - l$ degrees of freedom, where l is the number of independent linear restrictions placed on the event probabilities [3]. That is:

$$\lim_{n\to\infty}(X^2) \sim \chi_{k-1}^2 \tag{3}$$

Seven proofs have been given in [1]. I wish that I can understand the maths.

A widely accepted rule of thumb of when the chi-squared distribution approximation is appropriate is that the expected value of the number of times that each event occurs is at least 5.

1.2.1 Example Test

Suppose that we want to test whether the proportion of studying teenagers is the same for both genders or not with the following random sample of teenagers. We sample the same number for each gender. Our null hypothesis is that men and women are equally likely to study and our alternative hypothesis is that gender affects studying probability. We use a 5% significance level.

	Men (M) $P(M) = 0.5$	Women (F) $P(F) = 0.5$
Studying (Y) $P(Y) = 5/12$	$n_{MY} = 3$ $P(MY) = 5/24$ $E(n_{MY}) = 15$	$n_{FY} = 27$ $P(FY) = 5/24$ $E(n_{FY}) = 15$
Not-studying (N) $P(N) = 7/12$	$n_{MN} = 33$ $P(MN) = 7/24$ $E(n_{MN}) = 21$	$n_{FN} = 9$ $P(FN) = 7/24$ $E(n_{FN}) = 21$

Table 1: an example for Pearson's chi-squared test. Probabilities and expected values are calculated assuming that the null hypothesis is true.

In table 1, each expected value is at least 5 so Pearson's chi-squared test is appropriate here.

We have three independent linear restrictions placed on the event probabilities:

$$P(M) = 0.5 = P(MY) + P(MN)$$
$$P(F) = 0.5 = P(FY) + P(FN)$$
$$P(Y) = 5/12 = P(MY) + P(FY)$$

The remaining two restrictions can be derived from the above three restrictions, so they aren't independent.

$$P(N) = 7/12 = P(MN) + P(FN)$$
$$1 = P(MY) + P(FY) + P(MN) + P(FN)$$

Thus, the test statistic is approximately distributed as the chi-squared distribution with $4 - 3 = 1$ degree of freedom.

Computing the test statistic gives:

$$X^2 = \frac{(3 - 15)^2}{15} + \frac{(27 - 15)^2}{15} + \frac{(33 - 21)^2}{21} + \frac{(9 - 21)^2}{21} = 32.91$$

Hence, the corresponding two sided p-value is $P(\chi_1^2 \geq 32.91) = 9.6312705 \times 10^{-9}$. We reject the null hypothesis.

2 Fisher's Exact Test

After Pearson built his test, Fisher proposed a statistical test that is accurate and appropriate for low expected values [6].

Fisher used the same assumptions as Pearson did in his test but also assumed that the row and column totals do not change accross samples. He then identified cases that are equally as extreme as or more extreme than what was observed in the sample, and added them up to obtain the p-value.

In this approach, defining extremity can be problematic when two sided tests are introduced or multivariate contingency tables are introduced [6]. An approach used by the `fisher.test` function in R is to sum the probabilities of the relevant contingency tables that are less than or equal to the probability of the observed contingency table [2].

2.1 Example

We use the same example as the example in the previous chapter. We test the same hypothesis at the same significance level, but using Fisher's exact test method.

	Men (M) $n_M = 36$	Women (F) $n_F = 36$
Studying (Y) $n_Y = 30$	$n_{MY} = 3$	$n_{FY} = 27$
Not-studying (N) $n_N = 42$	$n_{MN} = 33$	$n_{FN} = 9$

Table 2: a duplicated example for Fisher's exact test. Probabilities and expected values are calculated assuming that the null hypothesis is true.

The probability that $n_{MY} = j$, where j is a positive integer between 0 and 30, follows the distribution Hypergeometric(72, 30, 36) with probability mass function:

$$P(n_{MY} = j) = \frac{\binom{30}{j} \cdot \binom{42}{36-j}}{\binom{72}{36}}$$

The observed value of n_{MY} is 3. For a two sided test, values of n_{MY} that are equally as extreme or more extreme than this are 0, 1, 2, 3, 27, 28, 29, 30. Inputting these values into the probability mass function and summing them together yields 8.4177433×10^{-9}. We reject the null hypothesis.

2.2 Fisher's Exact Test is Conservative

Some authors have argued that Fisher's exact test is conservative [6]. This is because the test statistic used is discrete. For example, the cumulative distribution function may be 3% for a certain level of extremity, but it jumps up to 7% going to the next lower level of extremity. This means that when conducting a hypothesis test with a significance level of 5%, in reality we really are testing with a significance level of 3%.

We test this by producing simulations of the studying experiment. Suppose that we sample n males and n females. The number of studying males and studying females both follow the binomial distribution with success rate p and number of Bernoulli trials n: B(n, p). This is used to construct 2x2 contingency tables, which are then used to compute p-values.

We produce 10000 simulations for each of the 4 different combinations of $n \in \{20, 100\}$ and $p \in \{0.2, 0.5\}$ for a total of 40000 simulations. The proportion of simulation tests rejected using significance levels of 1%, 5%, and 10% are shown in the summary below under the eponymous columns for both Fisher's and Pearson's test.

n	p	Fisher 1%SL	Fisher 5%SL	Fisher 10%SL	Pearson 1%SL	Pearson 5%SL	Pearson 10%SL
20	0.2	0.0031	0.0249	0.0485	0.0104	0.0528	0.1087
20	0.5	0.0042	0.0214	0.0453	0.0102	0.0453	0.0859
100	0.2	0.0075	0.0345	0.0700	0.0126	0.0520	0.1002
100	0.5	0.0063	0.0391	0.0751	0.0090	0.0538	0.1032

Table 3: a simulation summary table.

This experiment provides evidence that Fisher's exact test is indeed conservative since the actual rejection rate is likely to be less than the significance level set. In contrast, Pearson's chi-squared test is quite accurate

even for low n and p.

3 References

[1]   Benhamou, E. and Melot, V. 2018. Seven proofs of the pearson chi-squared independence test and its graphical interpretation.

[2]   R Core Team 2021. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing.

[3]   Wackerly, D. et al. 2008. *Mathematical statistics with applications.* Thomson Brooks/Cole.

[4]   Wikipedia contributors 2021. Chi-squared distribution. https://en.wikipedia.org/wiki/Chi-squared_distribution. [Online; accessed 13-October-2021].

[5]   Wikipedia contributors 2021. Chi-squared test. https://en.wikipedia.org/wiki/Chi-squared_test. [Online; accessed 13-October-2021].

[6]   Wikipedia contributors 2021. Fisher's exact test. https://en.wikipedia.org/wiki/Fisher%27s_exact_test. [Online; accessed 13-October-2021].

4 This assignment was way too time consuming :(