Biostatistics I: Hypothesis testing

Categorical data: Chi-square and Fisher's exact tests

Eleni-Rosalina Andrinopoulou

Department of Biostatistics, Erasmus Medical Center

e.andrinopoulou@erasmusmc.nl

@erandrinopoulou

Erasmus MC
University Medical Center Rotterdam

In this Section

- Chi-square test
- Fisher's exact test
- Examples

Chi-square test: Theory

Assumptions

- The study groups must be independent
- There are 2 variables, and both are measured as categories, usually at the nominal level
- The levels (or categories) of the variables are mutually exclusive

Chi-square test: Theory

The chi-square test tests the statistical significance of the observed relationship with respect to the expected relationship

- Two variables are related or independent
- Goodness-of-fit between observed distribution and theoretical distribution of frequencies

Chi-square test: Theory

Scenario

Is there a relationship between gender and whether or not someone followed an online course?

Hypothesis

H_0: there is not association between gender and whether someone followed an online course
H_1: there is an association between gender and whether someone followed an online course

If a chi-square goodness of fit test is performed then: The null and alternative hypotheses for our goodness of fit test reflect the assumption that we are making about the population

Chi-square test: Theory

Connection with linear regression

Let's assume a 2x2 table with variable *A* (*i*-th categories) and variable *B* (*j*-th categories). A multiplicative model that reproduces the cell frequencies exactly is:

$n_{ij} = N * \alpha_i * \beta_j * \alpha\beta_{ij}$ where

- α_i : the main effect of variable A at category *i*
- β_j : the main effect of variable B at category *j*
- $\alpha\beta_{ij}$: interaction between the two variables
- N : total number of subjects

If we take the logarithm of both sides, we can rewrite it as:

$log(n_{ij}) = log(N) + log(\alpha_i) + log(\beta_j) + \log(\alpha\beta_{ij})$,
which is a log-linear model

Chi-square test: Theory

Test statistic

- We must know the observed and expected values
- The test statistic is: $X^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i}$,
  where K are the contingency table cells, O is the observed value and E the expected value.

When the values in the contingency table are fairly small a “correction for continuity” known as the “Yates’ correction” may be applied to the test statistic: $X^2 = \sum_{i=1}^{K} \frac{(|O_i - E_i| - 1/2)^2}{E_i}$

Chi-square test: Theory

Sampling distribution

- χ^2-distribution with
  $df = (\textit{number of rows} - 1) * (\textit{number of columns} - 1)$
- Critical value and p-value

If a chi-square goodness of fit test is performed then:
$df = \textit{number of categories} - 1$

Chi-square test: Theory

Type I error

- Normally $\alpha = 0.05$

Draw conclusions

- Compare test statistic (X^2) with the critical value or the p-value with α

Chi-square test: Application

Scenario

Is there a relationship between gender and whether or not someone followed an online course?

Hypothesis

H_0 : there is not association between gender and whether someone followed an online course
H_1 : there is an association between gender and whether someone followed an online course

Chi-square test: Application

Collect and visualize data

Observed:

	Yes: online course	No: online course	Sum
Male	33	14	47
Female	29	24	53
Sum	62	38	100

Expected:

For each cell we calculate =

(total number of obs for the row) * (total number of obs for the column) / (total number of obs)

	Yes: online course	No: online course
Male	29.1	17.9
Female	32.9	20.1

Chi-square test: Application

Test statistic

$$X^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} = \frac{(33-29.1)^2}{29.1} + \frac{(14-17.9)^2}{17.9} + \frac{(29-32.9)^2}{32.9} + \frac{(24-20.1)^2}{20.1} = 2.59$$

Degrees of freedom

$$df = (number\ of\ rows - 1) * (number\ of\ columns - 1) = (2-1) * (2-1) = 1$$

Type I error

$$\alpha = 0.05$$

Chi-square test: Application

Critical values

Using `R` we get the critical value from the χ^2-distribution:
critical value$_\alpha$ = critical value$_{0.05}$

```
qchisq(p = 0.05, df = 1, lower.tail = FALSE)
```

```
[1] 3.841459
```

Chi-square test: Application

Draw conclusions

We reject the H_0 if:

- X^2 > critical value$_\alpha$

We have 2.59 < 3.84 $\Rightarrow$ we do not reject the H_0

Using R we obtain the p-value from the χ^2-distribution:

```r
pchisq(q = 2.59, df = 1, lower.tail = FALSE)
```

```
[1] 0.1075403
```

Fisher's Exact Test: Theory

- Fisher's exact test is an exact test - but has type I error rates less than the specified value (because it is based on a discrete test statistic)
- Fisher's exact test is a special case of **permutation** tests

  - Calculate the original test statistic
  - Shuffle (permute) the data and calculate the test statistic
  - Repeat the above step for every possible permutation of the sample
  - Calculate the fraction of the values of the test statistic that are as extreme or more to the original test statistic

Fisher's Exact Test: Theory

Assumptions

- The study groups must be independent
- The variables should be dichotomous
- Both row and column marginal totals are fixed in advance

Fisher’s Exact Test: Theory

Scenario

Is there a relationship between gender and whether or not someone followed an online course?

Fisher's Exact Test: Theory

	Yes: online course	No: online course	Total
Male	O11	O12	TotalR1
Female	O21	O22	TotalR2
Total	TotalC1	TotalC2	Total

- The test assumes that both the row and column totals (TotalR1, TotalR2, TotalC1 and TotalC2) are known
- It calculates the probability that we would have obtained the observed frequencies that we did (O11, O12, O21 and O22) given those totals

Fisher's Exact Test: Theory

If we assume the marginal totals as given, the value of $O11$ determines the other cells. Assuming fixed marginals, the distribution of the four cell counts follows the hypergeometric distribution, e.g for $O11$:

$$Pr(O11) = \frac{\binom{TotalR1}{O11}\binom{TotalR2}{O21}}{\binom{Total}{TotalC1}} = \frac{\frac{TotalR1!}{O11!O12!}\frac{TotalR2!}{O21!O22!}}{\frac{N!}{TotalC1!TotalC2!}} = \frac{TotalR1!TotalR2!TotalC1!TotalC2!}{Total!O11!O12!O21!O22!},$$

- $\binom{TotalR1}{O11} = \frac{TotalR1!}{O11!(TotalR1-O11)!}$
- ! denotes the factorial, e.g: $N! = N(N-1)(N-2)(N-3)\ldots1$

Fisher's Exact Test: Theory

Steps

- For all possible tables (given that TotalR1, TotalR2, TotalC1 and TotalC2 are fixed), calculate the relevant hypergeometric probability
- The p-value is the sum of hypergeometric probabilities for outcomes at least as favorable to the alternative hypothesis as the observed outcome

Type I error

- Normally α = 0.05

Fisher's Exact Test: Application

Scenario

Is there a relationship between gender and whether or not someone followed an online course?

Collect and visualize data

	Yes: online course	No: online course	Sum
Male	1	3	4
Female	3	1	4
Sum	4	4	8

For this table:

$$p = \frac{TotalR1!TotalR2!TotalC1!TotalC2!}{Total!O11!O12!O21!O22!} = \frac{4!4!4!4!}{8!1!3!1!3!} = 0.2285714$$

Fisher's Exact Test: Application

Other alternatives:

	Yes: online course	No: online course	Sum
Male	0	4	4
Female	4	0	4
Sum	4	4	8

$$p = \frac{TotalR1!TotalR2!TotalC1!TotalC2!}{Total!O11!O12!O21!O22!} = \frac{4!4!4!4!}{8!0!4!0!4!} = 0.01428571$$

	Yes: online course	No: online course	Sum
Male	3	1	4
Female	1	3	4
Sum	4	4	8

$$p = \frac{TotalR1!TotalR2!TotalC1!TotalC2!}{Total!O11!O12!O21!O22!} = \frac{4!4!4!4!}{8!3!1!3!1!} = 0.2285714$$

	Yes: online course	No: online course	Sum
Male	2	2	4
Female	2	2	4
Sum	4	4	8

$$p = \frac{TotalR1!TotalR2!TotalC1!TotalC2!}{Total!O11!O12!O21!O22!} = \frac{4!4!4!4!}{8!2!2!2!2!} = 0.5142857$$

	Yes: online course	No: online course	Sum
Male	4	0	4
Female	0	4	4
Sum	4	4	8

$$p = \frac{TotalR1!TotalR2!TotalC1!TotalC2!}{Total!O11!O12!O21!O22!} = \frac{4!4!4!4!}{8!4!0!4!0!} = 0.01428571$$

Fisher's Exact Test: Application

For **one-tailed**: find extreme cases from the same direction as our data: 0.2285714 + 0.01428571 = 0.243

	Yes: online course	No: online course	Sum
Male	1	3	4
Female	3	1	4
Sum	4	4	8

Original data

	Yes: online course	No: online course	Sum
Male	0	4	4
Female	4	0	4
Sum	4	4	8

Shuffle

$$p = \frac{TotalR1!TotalR2!TotalC1!TotalC2!}{Total!O11!O12!O21!O22!} = \frac{4!4!4!4!}{8!1!3!1!3!} = 0.2285714$$

$$p = \frac{TotalR1!TotalR2!TotalC1!TotalC2!}{Total!O11!O12!O21!O22!} = \frac{4!4!4!4!}{8!0!4!0!4!} = 0.01428571$$

Fisher's Exact Test: Application

For **one-tailed**: find extreme cases from the other direction as our data: 0.2285714 + 0.5142857 + 0.2285714 + 0.01428571 = 0.986

	Yes: online course	No: online course	Sum
Male	1	3	4
Female	3	1	4
Sum	4	4	8

Original data

$$p = \frac{TotalR1!TotalR2!TotalC1!TotalC2!}{Total!O11!O12!O21!O22!} = \frac{4!4!4!4!}{8!1!3!1!3!} = 0.2285714$$

	Yes: online course	No: online course	Sum
Male	3	1	4
Female	1	3	4
Sum	4	4	8

Shuffle

$$p = \frac{TotalR1!TotalR2!TotalC1!TotalC2!}{Total!O11!O12!O21!O22!} = \frac{4!4!4!4!}{8!3!1!3!1!} = 0.2285714$$

	Yes: online course	No: online course	Sum
Male	2	2	4
Female	2	2	4
Sum	4	4	8

Shuffle

$$p = \frac{TotalR1!TotalR2!TotalC1!TotalC2!}{Total!O11!O12!O21!O22!} = \frac{4!4!4!4!}{8!2!2!2!2!} = 0.5142857$$

	Yes: online course	No: online course	Sum
Male	4	0	4
Female	0	4	4
Sum	4	4	8

Shuffle

$$p = \frac{TotalR1!TotalR2!TotalC1!TotalC2!}{Total!O11!O12!O21!O22!} = \frac{4!4!4!4!}{8!4!0!4!0!} = 0.01428571$$

Fisher's Exact Test: Application

- For a **two-tailed** test we must also consider tables that are equally extreme in both direction
- This is challenging, therefore we sum the probabilities that are equal or less than that from the observed data:
  $0.2285714 + 0.01428571 + 0.2285714 + 0.01428571 = 0.486$

Draw conclusions

If $\alpha = 0.05 \Rightarrow$ we do not reject the H_0 since p-value is > 0.05

Further reading

- Campbell I. Chi-squared and Fisher–Irwin tests of two-by-two tables with small sample recommendations. Statistics in medicine. 2007 Aug 30;26(19):3661-75.
- Crans GG, Shuster JJ. How conservative is Fisher's exact test? A quantitative evaluation of the two-sample comparative binomial trial. Statistics in medicine. 2008 Aug 15;27(18):3598-611.
- Lydersen S, Fagerland MW, Laake P. Recommended tests for association in 2× 2 tables. Statistics in medicine. 2009 Mar 30;28(7):1159-75.