Data Analytics in R
Session 11

Maria Kunevich

Assignments

Assignment	Date of assignment	Deadline (midnight 23:59)
HW1	22 Sept 2022	28 Sept 2022
HW2	29 Sept 2022	5 Oct 2022
HW3	6 Oct 2022	12 Oct 2022
HW4	13 Oct 2022	19 Oct 2022
HW5	20 Oct 2022	2 Nov 2022
Paper summary	20 Oct 2022	20 Nov 2022
HW6	3 Nov 2022	9 Nov 2022
HW7	17 Nov 2022	27 Nov 2022
HW8	1 Dec 2022	11 Dec 2022
Project	20 Oct 2022	
Interim pitch		17 Nov 2022
Project Presentations		15 Dec 2022
Submission of Final report		29 Dec 2022

Revision - hypothesis testing

- What is hypothesis testing?
- What terms (sample statistic or population parameter) are used for hypotheses in a significance test?
- What are the three types of statistical tests?
- What are the steps in hypothesis testing? Comment on each of them
- What hypothesis states equality or no difference, or no relationship/effect?
- What hypothesis states no equality or existence of differences, relationship, or effect?
- What does a level of significance 5% mean?
- What does a p-value indicate?

Correlation tests

Correlation tests determine the extent to which two variables are associated:

Correlation test	Parametric?	Variables
Pearson's *r*	Yes	Interval/ratio variables
Spearman's *r*	No	Ordinal/interval/ratio variables
Chi square test of independence	No	Nominal/ordinal variables

The Chi-Square statistic

A chi-square ($\chi 2$) statistic is a single number that measures how much difference exists between **a model** (counts we would expect) and **actual observed data** (observed counts) if there were no relationship at all in the population.

The data used in calculating a chi-square statistic must be:

- ❏ random
- ❏ mutually exclusive
- ❏ drawn from independent variables
- ❏ drawn from a large enough sample

Example: tossing a coin

The Pearson’s Chi-Square test

Two main kinds of chi-square tests:

- **the test of independence** (exploring the relationship between two categorical variables), i.e. the RQ can be formulated like: “Is there a relationship between student gender and course choice?”
- **the goodness-of-fit test** (exploring how the sample data matches characteristics of the larger population), i.e. the RQ can be formulated like: “How well does the coin in my hand match a theoretically fair coin?”

The Chi-Square goodness-of-fit test

- Evaluate how ‘close’ the observed values are to those which would be expected under the predicted model
- **Example: tossing a coin**
- “How well does the coin in my hand match a theoretically fair coin?”
- Toss the coin 100 times: in the predicted model we can expect heads 100x0.5=50 times -> the probability of getting a head is 0.5
- But we might get more or less of heads in our actual data
- How much variation in the number of heads will we allow before we are confident in rejecting the hypothesis that p=0.5?
- What is our null hypothesis (H0) and alternative hypothesis (H1)?

  H0: the coin is equally likely to land head-up or tails-up every time (there is no relationship between landing on head or tail), i.e. any deviations from the 50% rate are due to chance.

  H1: there is a statistically significant difference between the observed and expected values, i.e the difference are not due to chance

The Chi-Square goodness-of-fit test

- To test the hypothesis we create a contingency table:

	Observed	Expected
Heads	38	50
Tails	62	50

- We can compute chi-square statistic and use it to determine the fairness of the coin: x^2 =5.76
- We compare the statistic with chi-square distribution: if it is in the tail of the distribution, then the probability of getting 38 heads using a fair coin would be rare
- if it is in the middle of the distribution, then it might be quite common to obtain 38 heads in 100 tosses from a fair coin

The Chi-Square goodness-of-fit test

Goodness of fit is a measure of how well a statistical model fits a set of observations.

- When goodness of fit is **high**, the values expected based on the model are close to the observed values, the **chi-square statistic is low**
- When goodness of fit is **low**, the values expected based on the model are far from the observed values, the **chi-square statistic is high**
- In our example is x^2 =5.76 a low or high goodness of fit?

To interpret the chi-square goodness of fit, we compare it to the appropriate chi-square distribution to decide whether **to reject the null hypothesis**

Source: https://www.scribbr.com/statistics/chi-square-goodness-of-fit/

The Chi-Square goodness-of-fit test

- From a **chi-square distribution** calculate the critical value:

  https://cdn.scribbr.com/wp-content/uploads/2022/05/Chi-square-table.pdf

To find the critical chi-square value, you'll need to know two things:

- the degrees of freedom (df): For chi-square **goodness of fit** tests, the df= the number of groups - 1
- Significance level (α): By convention, the significance level is usually .05.
- For a test of significance at α = .05 and df = 1, the x^2 critical value is 3.841.
- Compare the chi-square value to the critical value to determine which is larger.
- x^2 =5.76 is **greater** than the critical value

Source: https://www.scribbr.com/statistics/chi-square-goodness-of-fit/

The Chi-Square goodness-of-fit test

- If the x^2 value is **greater than the critical value**, then the difference between the observed and expected distributions is statistically significant ($p < \alpha$)
  The data allows you **to reject the null hypothesis** and provides support for the alternative hypothesis
- If the x^2 value is **less than the critical value**, then the difference between the observed and expected distributions is not statistically significant ($p > \alpha$).
  The data **doesn't allow you to reject** the null hypothesis and doesn't provide support for the alternative hypothesis
- **In our example**: the probability that the coin is fair is incorrect. Since the x^2 value is higher than the critical value, we can reject the null hypothesis that there is no difference between landing on heads or tails, it looks like our coin favours tails.

Source: https://www.scribbr.com/statistics/chi-square-goodness-of-fit/

The Chi-Square goodness-of-fit test: hypothesis testing

Using this test we can draw conclusions about the distribution of a population based on a sample and test whether the goodness of fit is “good enough” to conclude that the population follows the distribution:

The questions we can ask: was this sample drawn from a population that has...

- Equal proportions of male and female turtles?
- Equal proportions of red, blue, yellow, green, and purple jelly beans?
- 90% right-handed and 10% left-handed people?

Source: https://www.scribbr.com/statistics/chi-square-goodness-of-fit/

The Chi-Square goodness-of-fit test: hypothesis testing

The following conditions (pre-requisites) are necessary for perform a chi-square goodness of fit test:

- ❏ You want to test a hypothesis about the distribution of **one categorical** variable (if your variable is continuous, convert it to a categorical variable by separating the observations into intervals)
- ❏ The sample is **randomly** selected from the population
- ❏ There are a minimum of **five observations** expected in each group

Source: https://www.scribbr.com/statistics/chi-square-goodness-of-fit/

The Chi-Square test of independence

The **chi-square** (χ^2) test - evaluates a relationship between two **categorical** variables (especially nominal where order doesn't matter)

Synonyms: we are testing whether the variables are **related, associated, contingent, or dependent**

What kind of test is this? (parametric or **nonparametri**c)

The null and alternative hypotheses are:

- H0: the variables are independent, there is **no relationship** between the two categorical variables.
- H1: the variables are dependent, **there is a relationship** between the two categorical variables.

Source: https://www.scribbr.com/statistics/chi-square-test-of-independence/

The Chi-Square test of independence

The **chi-square** (χ 2) test is a **test of independence**, we compare the **observed** frequencies to the **expected** frequencies if there was **no relationship** between the two categorical variables

Small difference between the observed frequencies and the expected frequencies -> we cannot reject the null hypothesis of independence and thus we cannot reject the fact that the **two variables are not related**

Large difference between the observed frequencies and the expected frequencies -> we can reject the null hypothesis of independence and thus we can conclude that the **two variables are related**

Source: https://www.scribbr.com/statistics/chi-square-test-of-independence/

The Chi-Square test of independence

- The Chi-square distribution value (i.e. **critical value**) is the threshold between a small and large difference (critical region vs non-critical region)
- This critical value is calculated in the statistical table of the Chi-square distribution
- It depends 1) **on the significance level** α (usually set equal to 5%)
  2) **on the degrees of freedom**

  For a chi-square test of independence, the df is (number of variable 1 groups – 1) * (number of variable 2 groups – 1)

  For example: 1 variable: education level: 5 levels of education
  2 variable: income levels: 7 levels of income

  What is the df? (5-1)*(7-1)=24

The Chi-Square test of independence

To perform a chi-square test of independence, the best way to organize your data is type of frequency distribution table called a **contingency table** (cross tabulation) and a bar graph for visualisation

Steps to perform the chi-square test of independence:

1. **Calculate the expected frequencies** (using a contingency table)
2. **Calculate the chi-square value**
3. **Find the critical value**
4. **Compare the chi-square value to the critical value**
5. **Decide whether to reject the null hypothesis**

Optional post hoc tests: a significant difference doesn't tell you which groups' proportions are significantly different from each other

Source: https://www.scribbr.com/statistics/chi-square-test-of-independence/

The Chi-Square test for independence: hypothesis testing

The following conditions (pre-requisites) are necessary for perform a chi-square test for independence:

- You want to test a hypothesis about the relationship between two categorical variables (binary or nominal) (Chi-square tests of independence can be performed on ordinal variables with fewer than five groups)
- The sample is **randomly** selected from the population
- There are a minimum of **five observations** expected in each group

Source: https://www.scribbr.com/statistics/chi-square-test-of-independence/

Chi-square in R

- ***chisq.test()*** is a function used to perform test
- ***chisq.test(data)*** the output:

```
        Pearson's Chi-squared test

data:  c.table.1
X-squared = 486.7, df = 99, p-value < 2.2e-16
```

- Let's practise in R (the R script in is Github)
- Next session: t-tests and ANOVA (if we have time)

Thank you!