Data Analytics in R
Session 12

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

Feedback for the assignments

The link to your reports with homework assignment 7:

- https://maria-13.github.io/DataReporting/
- Feedback: **things to consider**
  - Report structure (headings/subheading, etc)
  - Output from R (consider what is required)
  - Explain in more detail what you doing/analysing
  - Dataset should be related to your RQ
  - First formulate your RQ, then hypothesis testing
  - Consider sample size
  - Interpretation of results - independent / dependent variable
  - Nice plots and additional packages: `ggpubr` and `ggstatsplot`

- How are your projects going? Any questions?

Plan for today

1. Revision: chi-square test
2. Introduction to t-tests
3. One-sample t-test: practice session in R
4. Two-sample t-test: practice session in R
5. Paired t-test: example

Revision - Chi-square test

- Is a Chi-square test the same as a χ^2 test?
- What Chi-square test do we use when we have a single measurement variable?
- What Chi-square test do we use when we have two measurement variables?
- What are the observed and the expected values?
- What are the analysis steps for a Chi-square test?
- What are the assumptions for a Chi-square test?
- How do we find the degrees of freedom for a Chi-square test?
- The bigger the chi-square statistic, the ... the p-value.
- What is the difference between a Chi-square test and a correlation?

Revision

- You want to know if people's opinion on videogames and violence (video games promote violence: yes / no) is related to their level of education (5 levels). Which test should you use?
- You want to know if people's water consumption (daily intake in glasses) is related to the demographic cohort they belong to (Boomer/Gen X/Millennial/Gen Z). Which test should you use?
- You want to know if there are equal numbers of male and female students in universities in Tallinn. Which test should you use? (you want to know if there are 90% right-handed and 10% left-handed people in universities in Tallinn. Which test should you use?)
- You want to know if there is a difference in the mean test score between two groups of students: one group was given a lecture and then home exercises once a week for 4 weeks, after 4 weeks the students had a test. The other group was given resources to explore on their own, then they had a discussion session once a week for 4 week. Which test should you use?

t-tests

A *t*-test is a statistical test that is used to compare **the means** of two groups (when we want to find out if there is **an effect** of treatment or some process).

- Independent variable: categorical variable (two levels/two groups)
- **Dependent variable: continuous variable** (we are comparing means)

  ? Can a dependent variable be discrete?

ANOVA is used if you want to compare **more than two groups**

Assumptions for *t*-tests

What type of statistical test is the *t*-test? (**parametric**, nonparametric)

Assumptions for the *t*-test:

1. Your data are independent
2. Your data are collected from a representative, randomly selected portion of the total population
3. Your data for dependent variable are (approximately) normally distributed for each group
4. Your sample size is large enough
5. Your data have a similar amount of variance within each group being compared (a.k.a. homogeneity of variance)

Source: https://www.scribbr.com/statistics/t-test/

Types of *t*-tests

Consider: do you have data from a single population or do you have data from two different populations?

Do you have predictions about a specific direction of the difference (greater or less)?

One-sample, two-sample, or paired *t*-test?

- You have groups that come from *a single population* (e.g. measuring before and after an experimental test or treatment), perform **a paired *t*-test**.
- If the groups come from *two different populations* (e.g. two different species, or people from two separate cities), perform **a two-sample *t*-test** (independent *t*-test).
- If there is one group being compared *against a standard value* (e.g. population mean), perform **a one-sample *t*-test**.

One-tailed or two-tailed t-test?

- If you only care whether the two populations are different from one another, perform a **two-tailed** *t*-test.
- If you want to know whether one population mean is *greater than or less* than the other, perform a **one-tailed t-test** (provides more power to detect an effect)

Source: https://www.scribbr.com/statistics/t-test/

One-sample *t*-test: example

Imagine we want to explore monthly expenses of students in Tallinn.

Hypothesis testing steps:

0. Collect data and perform descriptive statistics

1. State your research hypothesis as a null hypothesis and alternate hypothesis (Ho) and (Ha or H1)
2. Prepare data to test the hypothesis
3. Select an appropriate statistical test and check the assumptions (pre-requisites). Perform the statistical test
4. Calculate the p-value, select significance level (1%, 5%). Decide whether to reject or fail to reject your null hypothesis
5. Present your results

One-sample *t*-test: example

Imagine we want to explore monthly expenses of students in Tallinn.

Hypothesis testing steps:

0. Collect data and perform descriptive statistics

1. We created a google form and collected anonymous data from 100 randomly selected students in Tallinn

   Age | Gender | Education | Monthly expenses

2. We perform descriptive statistics
3. What is our sample and what is our population?

Step 1: It is reported that students on average spend 850 euros per month. Based on that we formulate our Research question and null and alternative hypothesis

One-sample *t*-test: example

Step 1: RQ: is it possible that on average a university student spends 850 euros per month as expenses?

Null hypothesis: the population mean is equal to 850 **(no difference)**

Alternative hypothesis: the population mean is not equal to 850

Step 2: Prepare data to test the hypothesis

Monthly expenses

Step 3: an appropriate statistical test - one-sample *t*-test
Check the assumptions: 1) single variable: expenses -> normality (check the histogram); 2) data independence (randomly selected participants)

One-sample *t*-test: example

Step 4: select significance level (1%, 5%), calculate the p-value, decide whether **to reject or fail to reject** your null hypothesis

Let's do all these steps in R and look at the output:

- **The t value: -7.66** (it is negative that tells us something about the direction, but we usually care about the absolute value of the difference, i.e. the distance from 0)
- **The degrees of freedom:** df = 99. Degrees of freedom is calculated based on your sample size, and shows how many 'free' data points are available in your test for making comparisons. The greater the degrees of freedom, the better your statistical test will work.
- **The p value:** 1.237e-11 (i.e. 1.2 with 11 zeros in front). This describes the probability that you would see a *t* value as large as this one by chance.
- **A statement of the alternative hypothesis** (Ha). In this test, the Ha is that the population mean is not equal to 850
- **The 95% confidence interval**. This is the range of numbers within which the true difference in means will be 95% of the time.
- **The mean** for the group: x = 735

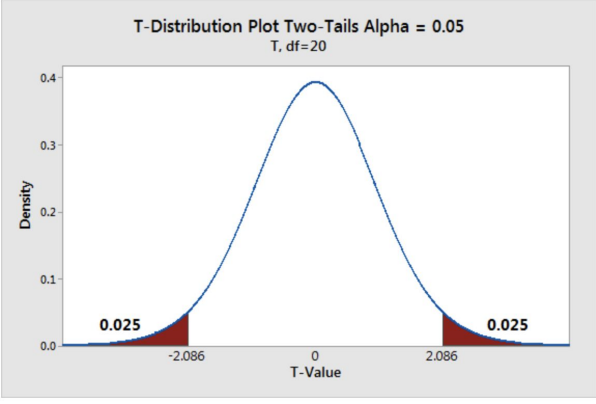
One-sample *t*-test: example

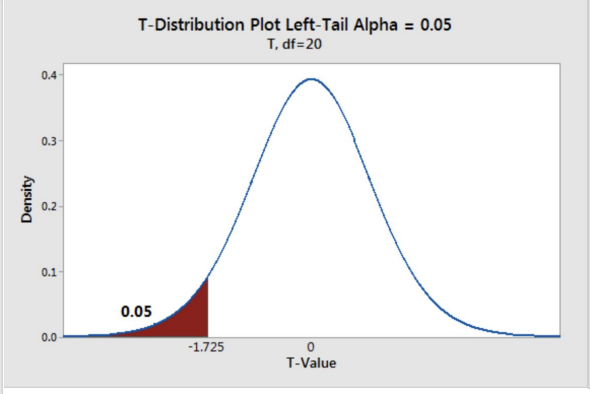
Step 5: Report the results:

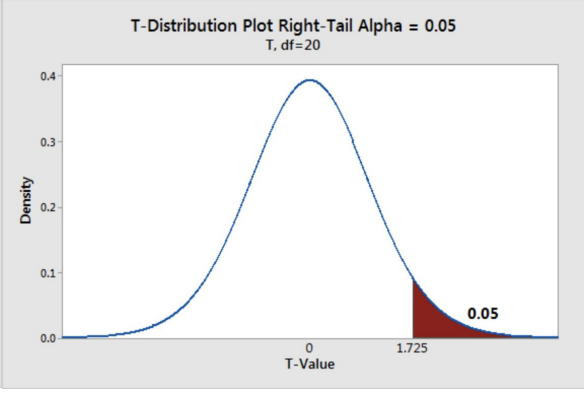
Monthly expenses

Students in Tallinn reported lower average monthly expenditure than found in the population as a whole, $t(99) = -7.66$, $P < .001$)

➔ If we know that the mean for the whole population should **be less or greater than** our mean, we can perform **one-tailed test** to make a more powerful conclusion







Two sample *t*-test

Let's look at our dataset with a different research question:

RQ: is there a difference between male and female students in terms of how much money they spend per month?

Describe all the steps for hypothesis testing:

0. Descriptive statistics

1. **Step 1:** *Null hypothesis:* the expenses of male and female students are the same/equal (there is NO difference)

   Alternative hypothesis: the expenses of male students are not equal to the expenses of female students

2. **Step 2:** Prepare the data

Gender

Monthly expenses

Two sample *t*-test

3 **Step 3:** an appropriate statistical test - two-sample *t*-test
Check the assumptions:

1) first variable: expenses for men -> normality (check the histogram); second variable: expenses for women -> normality (check the histogram)

2) data independence (randomly selected participants)

4 **Step 4:** select significance level (1%, 5%), calculate the p-value, decide whether **to reject or fail to reject** your null hypothesis

5 **Step 5:** report the results

Let's do it in RStudio!

Two sample *t*-test

5 **Step 5:**

report the results

```
        Two Sample t-test

data:  male$expenses and female$expenses
t = 0.3584, df = 98, p-value = 0.7208
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -49.15385  70.82174
sample estimates:
mean of x mean of y
 740.4386  729.6047
```

There was no significant effect for gender, *t*(98) = 0.35, p = .072, despite men (M = 740, SD = 150.6) spending on average more than women (M = 729, SD = 148.4).

Average text: There [was or was not] a significant difference in [response variable of interest] between [group1] (M = [Mean], SD = [standard deviation]) and [group2] (M = [Mean], SD = [standard deviation]); t(df) = [t-value], p = [p-value].

Paired *t*-test

For a paired t-test the research design will be different: e.g. we can collect data from **the same sample of students**, but asking them about their expenses per month during COVID (or in 2019) and current expenses (in 2022)

RQ: is there a difference between the amount of money students in Tallinn spent per month in 2019 and the amount of money they spend in 2022?

The same five steps

Reporting the results - *Average text:*

There [was or was not] a significant difference in [response variable of interest] between [group1] (M = [Mean], SD = [standard deviation]) and [group2] (M = [Mean], SD = [standard deviation]); t(df) = [t-value], p = [p-value].

Next session

- Revision of the course materials
- Introduction to linear regression

Thank you!