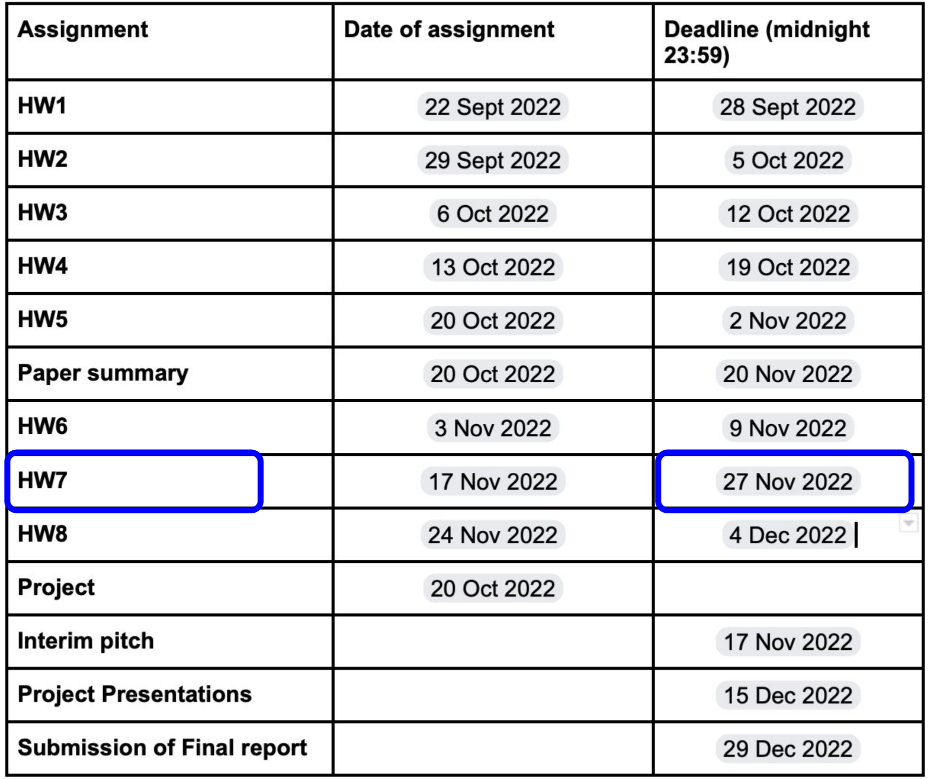
Data Analytics in R
Session 10

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	24 Nov 2022	4 Dec 2022
Project	20 Oct 2022	
Interim pitch		17 Nov 2022
Project Presentations		15 Dec 2022
Submission of Final report		29 Dec 2022

Project presentation

- Please add information about your project on the Miro board if you haven’t done so yet: https://miro.com/welcomeonboard/SUw3RHpXWjBpUDF2V0dwTWhkOFVVUnlUTnZ4Qm1oVGtSTVN0SmNCUmJyODhydU9hUzA3VUpNZVZHRnNBenhqVHwzMDc0NDU3MzYzNjI1MDIzMjY2fDI=?share_link_id=164034942325

Hypothesis testing

- Hypothesis testing is a formal process of statistical analysis using **inferential statistics**
- **Goal:** compare *populations* or relationships between variables using *samples*, i.e. make a decision about the value of a population parameter based on sample data

  Statistical tests are used for hypothesis testing, they also estimate sampling errors to make valid inferences

  Sample error: calculated by dividing the standard deviation of the population by the square root of the size of the sample, and then multiplying the resultant with the Z-score value, which is based on the confidence interval

- Online calculation:

  http://www.rogerwimmer.com/mmr/mmrsampling_error.htm

Source: https://www.scribbr.com/statistics/hypothesis-testing/

Hypothesis testing

Three forms of statistical tests: **tests of comparison**, **correlation** and **regression**

Two main types of statistical tests: parametric and non-parametric tests

- What is the difference between parametric and non-parametric tests?
- Which tests are more powerful? Why?
- What are the main assumptions for parametric tests?

1. Normality (data follows a normal distribution of scores)
2. Equal variance (a measure of spread for each group is similar)
3. Independence (data is sampled randomly and independently)
4. No outliers (no extreme data points)

Source: https://www.scribbr.com/statistics/hypothesis-testing/

Steps in hypothesis testing

There are 5 main steps in hypothesis testing:

0. Descriptive statistics

1. State your research hypothesis as a null hypothesis and alternate hypothesis (H_0) and (H_A or H_1)
2. Collect 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

Source: https://www.scribbr.com/statistics/hypothesis-testing/

Step 1: Null hypothesis and alternative hypothesis

Your initial research hypothesis is usually the alternative hypothesis, the null hypothesis predicts that there is no relationship between the variables we are interested in. For a statistical test we need to restate the initial hypothesis:

Null hypothesis (H0): There's **no effect** in the population.

Alternative hypothesis (Ha or H1): There's **an effect** in the population.

The effect is usually the effect of the **independent variable** on the **dependent variable**

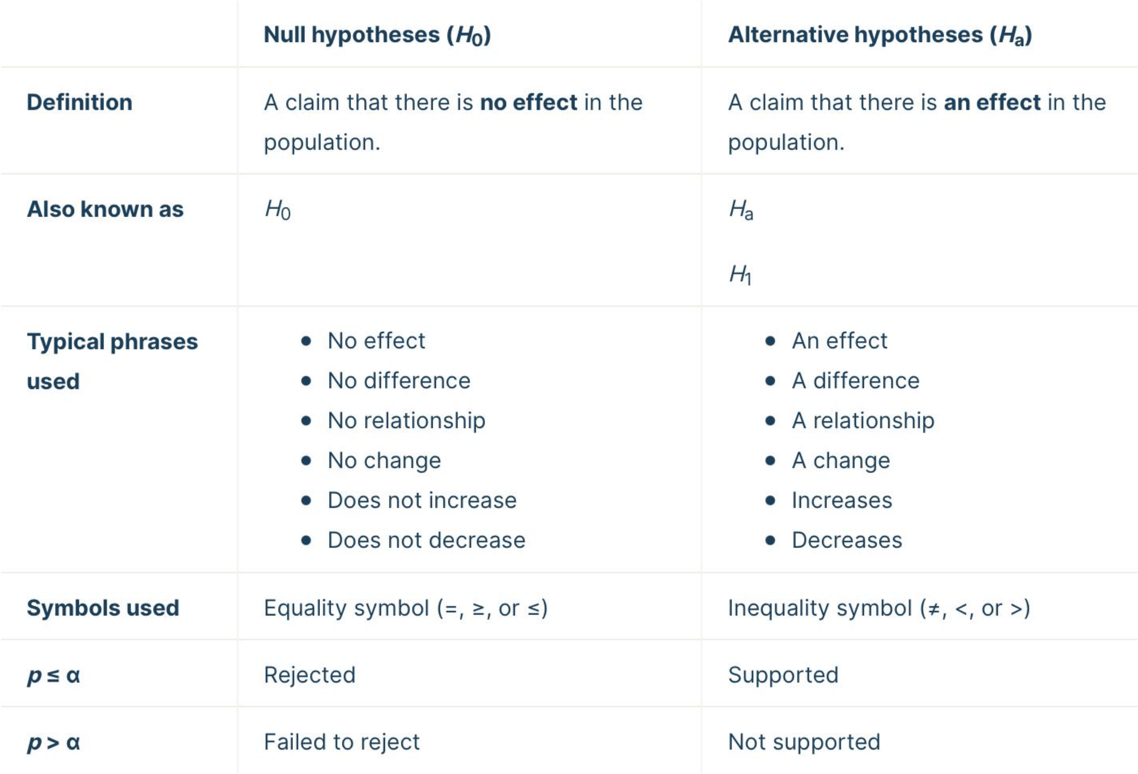
Salary and expenses

H0: There is NO relationship between salary and expenses

HA: There is a relationship between salary and expenses

Source: https://www.scribbr.com/statistics/hypothesis-testing/

Null hypothesis and alternative hypothesis



	Null hypotheses (H_0)	Alternative hypotheses (H_a)
Definition	A claim that there is **no effect** in the population.	A claim that there is **an effect** in the population.
Also known as	H_0	H_a H_1
Typical phrases used	• No effect • No difference • No relationship • No change • Does not increase • Does not decrease	• An effect • A difference • A relationship • A change • Increases • Decreases
Symbols used	Equality symbol (=, ≥, or ≤)	Inequality symbol (≠, <, or >)
$p \leq \alpha$	Rejected	Supported
$p > \alpha$	Failed to reject	Not supported

Source: https://www.scribbr.com/statistics/null-and-alternative-hypotheses/

Step 2: collect data

What steps are important when collecting data?

- defining the ‘right’ data for the research
- analysing the sampling techniques
- checking the quality of data
- considering research design
- considering how the independent variable is manipulated
- analysing how the data is coded

Define your variables: **dependent** variable/**independent** variable/**controlled** variable

❑ The independent variable (the cause) - does not change based on other variables
❑ The dependent variable (the effect) - depends on the independent variable
❑ The controlled variable - does not change during the experiment

Step 3: Statistical tests

❑ Statistical tests are used in hypothesis testing

❑ They determine whether a predictor variable (independent variable) has a **statistically significant relationship** with an outcome (dependent) variable

OR

❑ They are used to **estimate the difference** between two or more groups

❑ Help to determine whether the observed data fall outside of the range of values predicted by the null hypothesis

❑ Depend on the **types of variables** in your data and on **the level of measurement** (nominal, ordinal, interval, ratio)

Source: https://www.scribbr.com/statistics/statistical-tests/

Step 4: p-value

- the **p-value** is the *probability* of obtaining test results at least as extreme as the result actually observed, under the assumption that the null hypothesis is correct *(Wikipedia)*
- **Low p-value** - the observed outcome is very unlikely under the null hypothesis

  Low p-value indicates that there is little or no overlap between groups, i.e. it is unlikely that the difference between the groups happened by chance

- **Higher p-value** - indicates that there is high within-group variance and low between-group variance, i.e. it is likely that the difference between groups is caused by chance
- A p-value of **0.05** or lower is generally considered statistically significant

Step 4: reject or fail to reject the null hypothesis

- the p-value is generated by the statistical test
- use the p-value to guide your decision whether to reject or fail to reject the null hypothesis
- predetermined level of significance for rejecting the null hypothesis is 0.05

(there is a less than 5% chance that you would see these results if the null hypothesis were true)

Example: p-value 0.03 is below your cutoff of 0.05, so you decide to reject your null hypothesis of no difference

Step 5: reporting your results

Two ways: statistical results -> report if you can reject the null hypothesis

Academic papers -> report if results support the alternative hypothesis

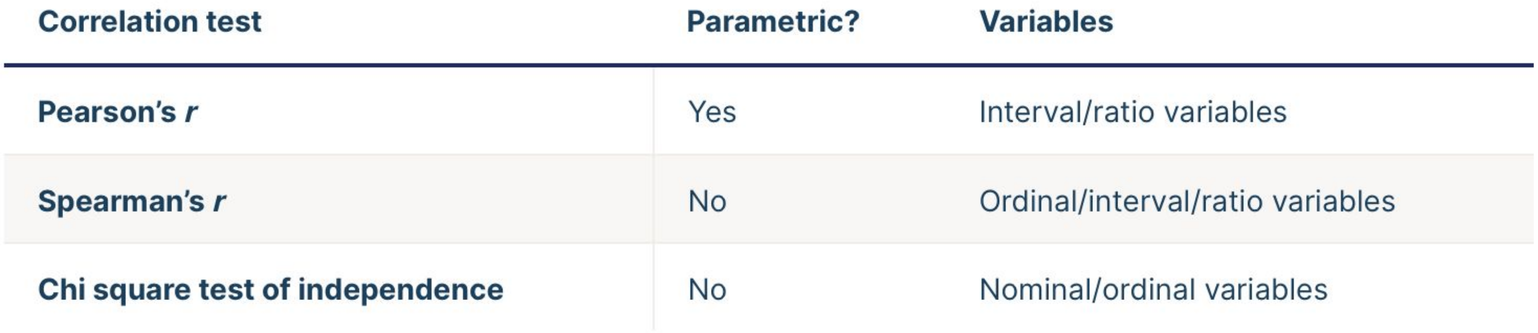
- ❏ reject the null hypothesis -> we can report that the statistical test **supports** our hypothesis
- ❏ fail to reject that null hypothesis -> the difference between the groups can have arisen by chance, i.e. the test is inconsistent with our hypothesis

Example (our exercise):

In our comparison of correlation between salary and expenses we found a linear strong positive correlation, r (8) = .76, p < .009; therefore, we can reject the null hypothesis that there is no relationship between the amount of money people receive as salary and the amount of money they spend and conclude that the more money people receive, the more they tend to spend.

Correlation tests

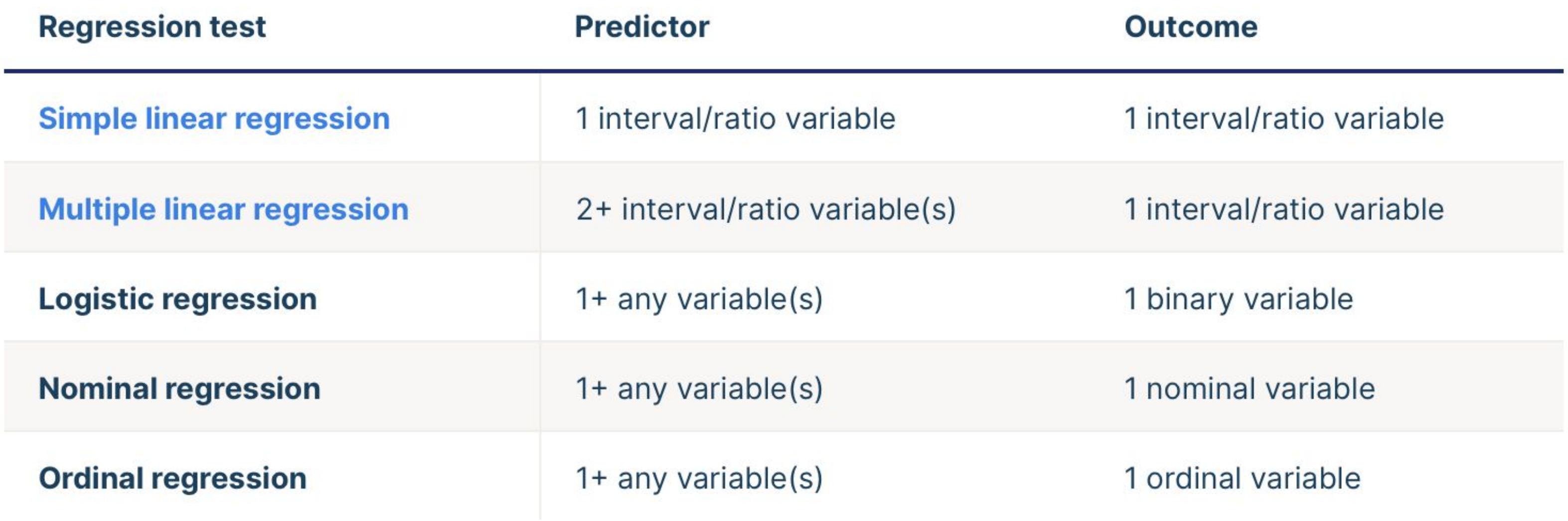
Correlation tests determine the extent to which two variables are associated:



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

Regression tests

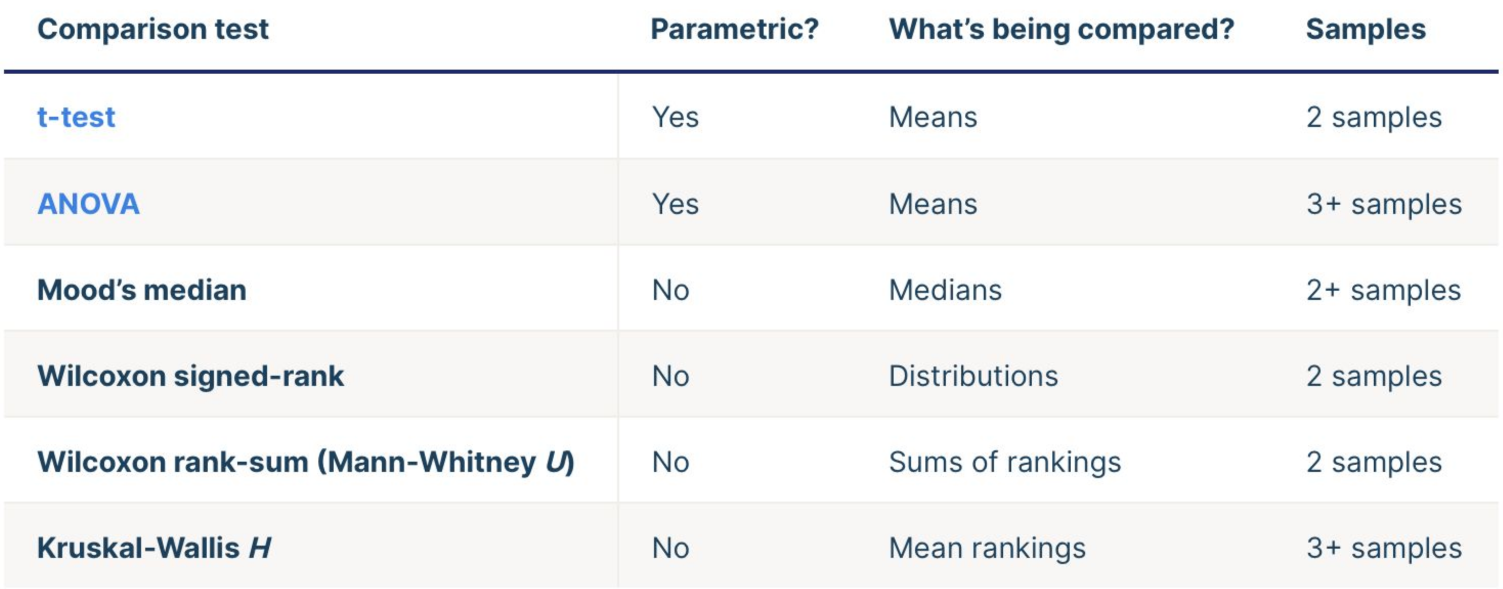
Regression tests demonstrate whether changes in predictor variables cause changes in an outcome variable.



Regression test	Predictor	Outcome
Simple linear regression	1 interval/ratio variable	1 interval/ratio variable
Multiple linear regression	2+ interval/ratio variable(s)	1 interval/ratio variable
Logistic regression	1+ any variable(s)	1 binary variable
Nominal regression	1+ any variable(s)	1 nominal variable
Ordinal regression	1+ any variable(s)	1 ordinal variable

Tests for comparison

Comparison tests compare the difference between two or more groups in means, medians, or rankings (means - interval or ratio data; medians and rankings -ordinal data)



Comparison test	Parametric?	What’s being compared?	Samples
t-test	Yes	Means	2 samples
ANOVA	Yes	Means	3+ samples
Mood’s median	No	Medians	2+ samples
Wilcoxon signed-rank	No	Distributions	2 samples
Wilcoxon rank-sum (Mann-Whitney *U*)	No	Sums of rankings	2 samples
Kruskal-Wallis *H*	No	Mean rankings	3+ samples