Data Analytics in R
Session 4

Maria Kunevich

Homework feedback

- Your feedback for the course and course assignments will be greatly appreciated!
- You can add your thoughts to our Miro board here (and do so anonymously if you wish):
- *https://miro.com/welcomeonboard/SUw3RHpXWjBpUDF2V0dwTWhkOFVVUnlUTnZ4Qm1oVGtSTVN0SmNCUmJyODhydU9hUzA3VUpNZVZHRnNBenhqVHwzMDc0NDU3MzY2NDE0OTE3Njc4?share_link_id=308470562965*
- Please let me know if you find any typos/errors in the lecture slides, R scripts, assignments, etc

Homework assignments - preparation

- Booking **A-303** for study sessions, you need:

1) to check the room A-303 availability in ASIO,

2) email Kristi (kristi.oikimus@tlu.ee) with the date/time,

3) provide her with the name of the **contact person** for the session

- If you work on assignments in groups, please mention your group members in your R scripts

Homework assignment 1: feedback

Question: How do you feel about the openness of your submissions?

Update: we'll leave everything as it is now (gists are open in comments), but if you want to submit to me directly, send me an email (before the deadline) with the link to your gist, only I will see it and provide my feedback directly to you

After everybody submits the task, I'll upload the HW1_Solution file in the "/Assignments" folder

How I'm checking the assignments:

1. If your comment is hidden with the 'Resolved' status, it means I can see/access your homework and will provide feedback shortly (when everybody submits the assignment)
2. I'm copy pasting your script into my RStudio environment, run it and provide some feedback in comments: **#Maria**:
3. I create a gist based on the script with comments
4. I'll reply to your submission in a comment with a link to the Gist I created based on your R script

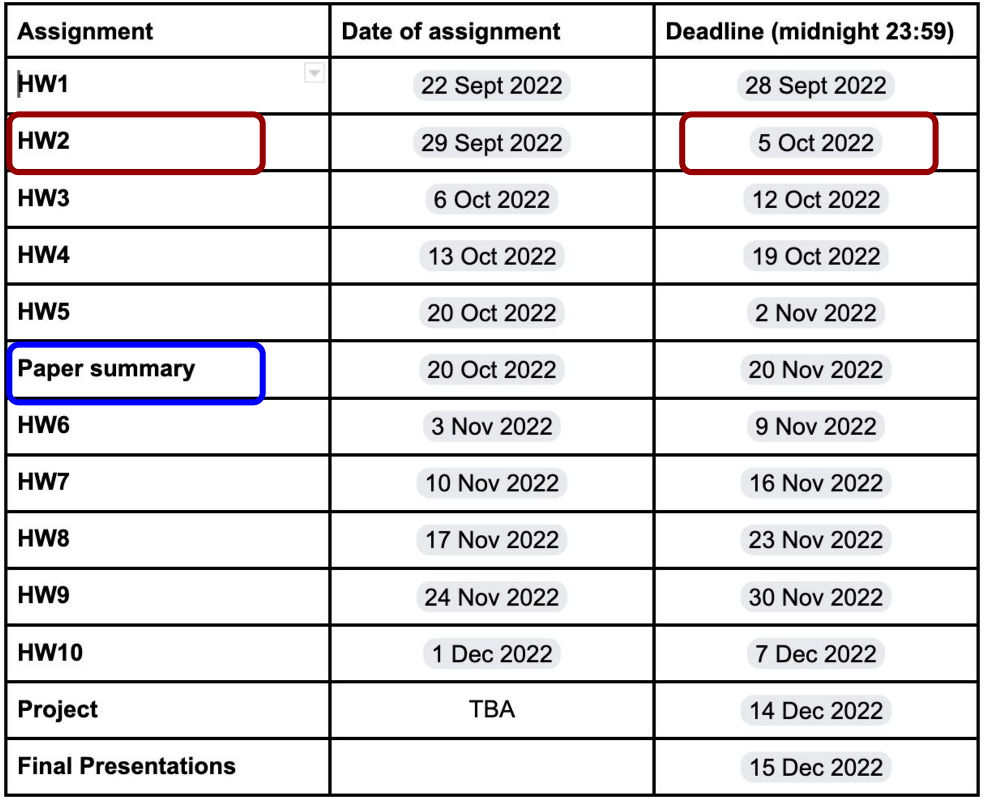
   NB! Be careful with how you comment/organise your answers and please **don't delete** my original comments/task descriptions

Homework assignment 2

- This time we will do it a bit differently:
- Part 1: the same procedure for copy pasting the task (HW2) and creating your own Gist, submit the link on **Wednesday evening** by 23:59
- Part 2: we'll practise in the **DataCamp** environment:

1) Log in to DataCamp (https://www.datacamp.com) with your **TLU** Google account
2) Join the group "Data Analytics in R" through the link I'll email to all of you or here: https://www.datacamp.com/groups/shared_links/aa89574f778134ef99010b5a6ea7e3f41c413cf0b924d09626b06def38011717
3) Complete **Chapters 1, 2, 3, 4** from '**Introduction to R**' course to get 4 points for each chapter

Course information: assignments deadlines



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	10 Nov 2022	16 Nov 2022
HW8	17 Nov 2022	23 Nov 2022
HW9	24 Nov 2022	30 Nov 2022
HW10	1 Dec 2022	7 Dec 2022
Project	TBA	14 Dec 2022
Final Presentations		15 Dec 2022

Plan for today

1. Revision of the basic concepts in R - **Kahoot** game

   Go to: https://kahoot.it enter the game Pin

2. Practice session: working with matrices in R
3. Sample vs Population
4. Basic concepts in descriptive statistical analysis

Revision of the basic concepts in R

Kahoot game

(now you can practise the questions on your own if you wish **till Oct 6**)

https://kahoot.it/challenge/06525185?challenge-id=fedce431-8dda-405f-bde5-fb9a9ee0f0d6_1664534177873

(most questions are from online resources like https://www.w3schools.com https://www.r-exercises.com/start-here-to-learn-r/)

Matrices in R

- Array in R - **multi-dimensional** generalisations of vectors
- **Matrix** - a special case of arrays, namely a **two-dimensional** matrix, which is generally used to represent tabular data (aka **tables**)
- Several ways to create a matrix:
  - using function **matrix()**
  - combining vectors by using the functions **cbind()** and **rbind()**
- Let’s practise in R!

Matrices in R: accessing elements

We can assign/fill values in an empty vector by using “**[]**” operator which is known as the **index operator**: [1] square brackets - Indexing starts from 1

- A matrix uses two indices: row - the first one and column - the second one
- **Comma after** the index operator refers to **rows** [1,]
- **Comma before** the index operator refers to **columns** [,1]
- Check R script for examples how to access elements and subset the matrix

The standard matrix operations (+, -, *, -, ^, %) are carried out **elementwise**

- **rowMeans(a)** - a function to get **row** means of a matrix a
- **colMeans(a)** - a function to get **column** means of a matrix a

Descriptive statistics

Descriptive statistics summarises and organises characteristics of **a data set**.

A data set is a collection of responses or observations from **a sample** or **entire population**.

- It's important to understand **the difference** between a sample and population
- To revise these concepts - check here:
  - ➢ https://www.scribbr.com/methodology/population-vs-sample/
  - ➢ https://www.investopedia.com/terms/s/sample.asp
  - ➢ https://explorable.com/population-sampling - look at **different types of sampling**

Sample vs Population

Some quick tests to check if you understand the difference::

- https://www.scribbr.com/methodology/population-vs-sample/ - Section: Practice questions
- https://www.khanacademy.org/math/ap-statistics/gathering-data-ap/sampling-observational-studies/e/identifying-population-sample

Descriptive and inferential statistics

After collecting the data, the first step of statistical analysis is to **describe** and **summarise the data,** for example, look at frequency distributions, compute the average of some variables, explore the relation between two variables - all these refer to **descriptive statistics.**

The next step is **inferential statistics**, where you can **test your hypotheses** and decide if the results are **generalisable to a larger population**.

Data type classification in statistics

Watch an introductory video from DataCamp on **what statistics is**:

https://campus.datacamp.com/courses/introduction-to-statistics-in-r/summary-statistics?ex=1

Make sure you understand **the difference between different data types** in statistics!

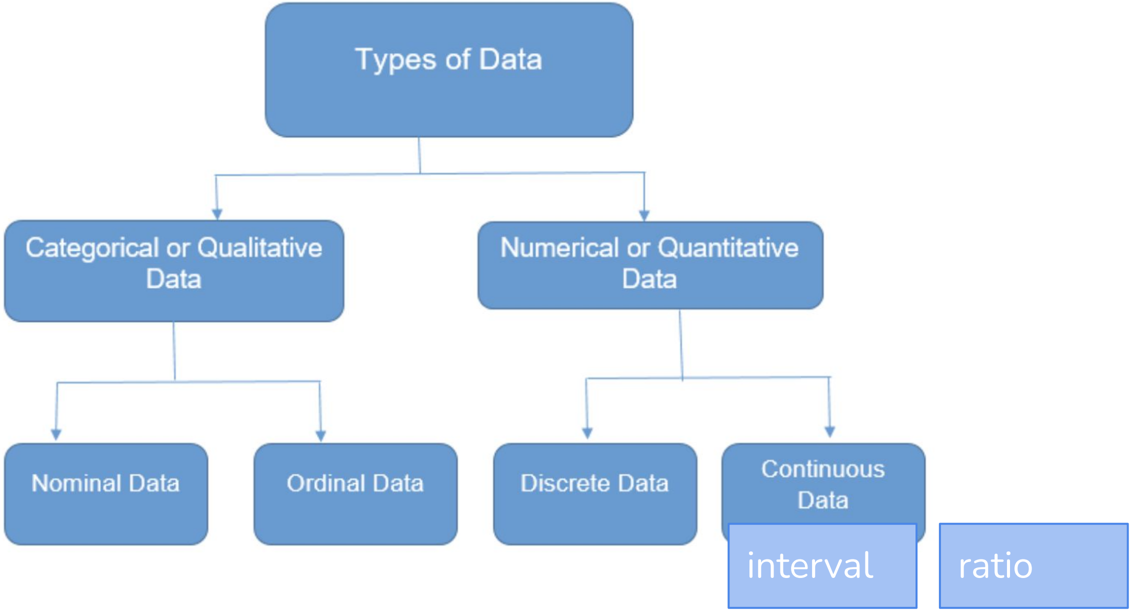
More explanations here:

- ➢ https://www.analyticsvidhya.com/blog/2021/06/complete-guide-to-data-types-in-statistics-for-data-science/
- ➢ https://www.scribbr.com/statistics/levels-of-measurement/
- ➢ https://www.chi2innovations.com/blog/discover-data-blog-series/data-types-101/

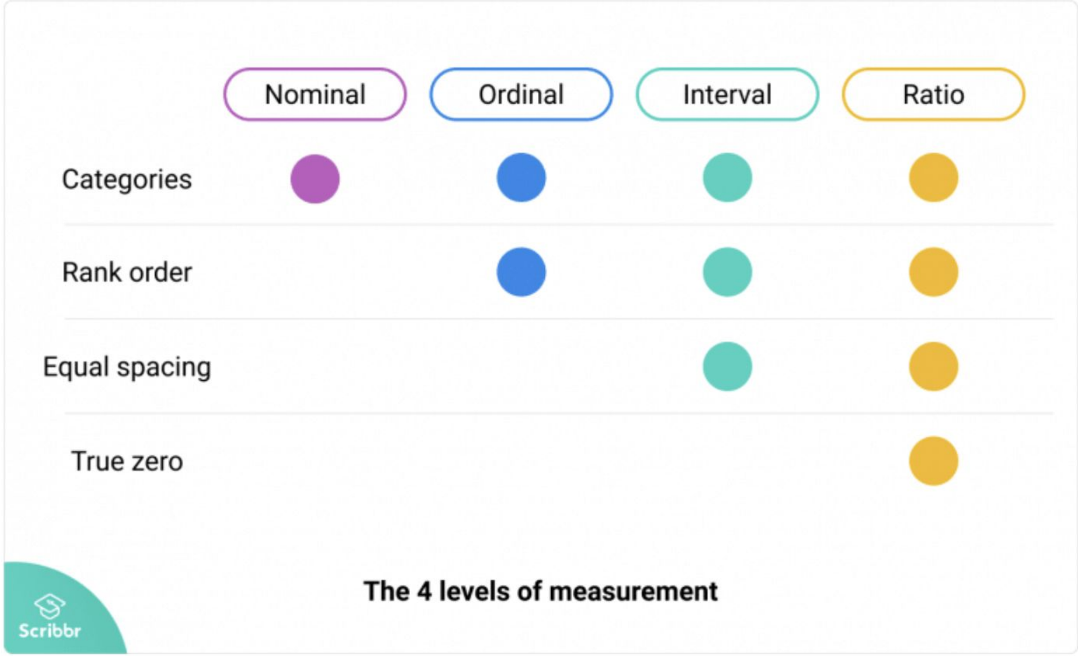
Check your understanding:

- ➢ https://campus.datacamp.com/courses/introduction-to-statistics-in-r/summary-statistics?ex=3

Data type classification



Data type classification



	Nominal	Ordinal	Interval	Ratio
Categories	●	●	●	●
Rank order		●	●	●
Equal spacing			●	●
True zero				●

The 4 levels of measurement

Why data types are important?

Data types are an extremely **important** concept in statistics and data analysis as certain statistical methods can **only** be used with **certain data types.**

Continuous data is analysed differently from **categorical** data, so you need to know the types of data you are dealing with to apply the correct method of analysis.

Next sessions

- **In Session 5** we will discuss **Data frames** in more detail
- Revise types of descriptive statistics
- And we will explore different packages for **importing** and **visualising** data in R

Thank you!