LMU KLINIKUM
Gemeinsam. Fürsorglich. Wegweisend.

HELMHOLTZ MUNICH

KimHellmuth Lab

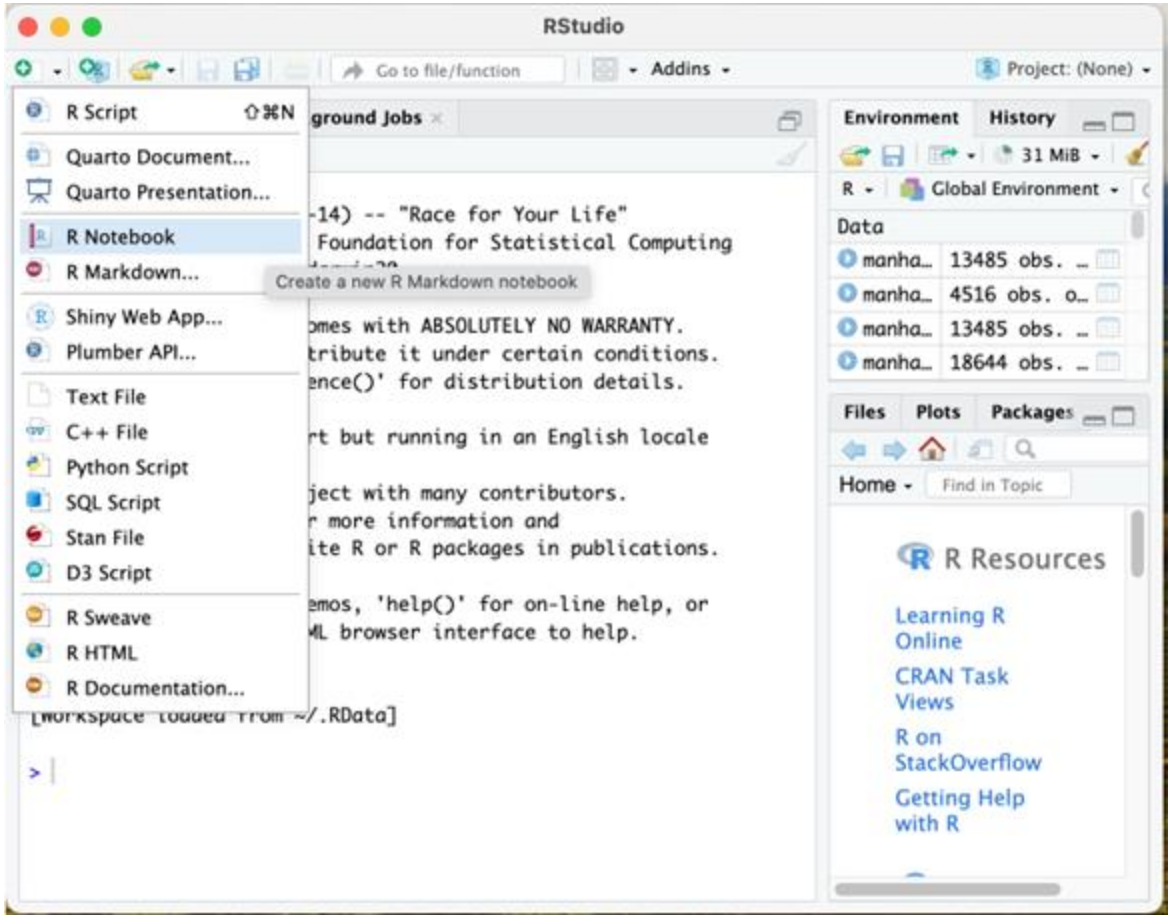
DR.VON HAUNERSCHES KINDERSPITAL
Helfen Forschen Heilen. Seit 1846

Hands-on Biological Data Science with R

Exercise 2 - Statistics in R

Exercise 2
Exercise Upload Format

- Open a new R Notebook in R Studio.
  - The default output is html notebook, don't change that. This means your 'R Notebook' will be saved in two formats, one is Rmd, the format you'll be opening to make edits etc.; and an html format.
- After you are done with it, upload your html file to moodle.
  - Name your html file in the following format **'exercise2_SURNAME_FIRSTNAME'**
  - Make sure it can be opened before you upload.
    - An html file can be opened by any web browser, so either double click and view or, right click and select your browser to open (Chrome, Safari etc.)
- <u>Deadline:</u> **05.12.2025 Fri, 23.59**



```
-14) -- "Race for Your Life"
 Foundation for Statistical Computing

omes with ABSOLUTELY NO WARRANTY.
tribute it under certain conditions.
ence()' for distribution details.

rt but running in an English locale

ject with many contributors.
r more information and
ite R or R packages in publications.

emos, 'help()' for on-line help, or
ML browser interface to help.

[Workspace loaded from ~/.RData]

> 
```

Exercise 2
Data

- In this exercise you will use the data from the first exercise. If you have deleted the data from your computer you can find it in moodle again

Library Requirements

- To do the exercises for this session, we'll use a r statistics package called 'rstatix'. Install and load it to your notebook for this exercise.

Exercise 2

Part 1 - Descriptive Statistics

Height Distribution

- Plot the distribution of height in the dataset using an appropriate plotting format.
- Compute the values of **mean** and **SD** for the data.
  - How much of the people's height lie within +/- 1SD of the data? How much are them in +/- of 2SD data? Calculate the percentages. Is it similar to what would you'd expect?
- Show the mean, and +/- 1 and 2 SD values in your plot. **(Plot1)**

Height Distribution - Sex

- Create a overlapping plot colored for males and females' height distribution. **(Plot 2)**
  - How does it look do you think sex could be a factor affecting height ? How is the general shape of the plot affected ?
- Stratify (subset) the data for both sexes. Calculate the mean and the +/- 1 2 SD for each sex. Calculate the percentage covered in the +/- 1 and 2 SD's of the data.
  - Recreate the plot you've created for the whole data for two sexes separately. Show mean and 1 2 SDs in the plot. **(Plot 3 and 4)**

Exercise 2

Part 2 - Inferential Statistics

Sampling

- Sample 10, 100, 1000 and 10000 data points (rows) from the data.
    - To randomly sample a data in R you use sample_n() function. In order to make sure the example is reproducible (i.e. everyone would get the same sample from that code), run the code ‘set.seed(2024)’ before you sample. Sample the data as shown below for all four sample sizes.
        - set.seed(2024)
        - brfs_sub10 <- brfss_data %>% sample_n(10)

Distribution and Mean of the Height

- Plot the distribution of height for all our subsampled datasets.
    - Have the distribution of height changed ? If yes, how ?
- Based on the subsampled data could you infer the mean of the big unsampled data? How confident are you with these sample sizes ? For the data with 10 samples, run ‘t.test’ with different Confidence Intervals (90-95-99) like below.
    - t.test(brfs_sub10$Height, conf.level = 0.90)
    - How does the resulting interval change with different confidence levels? Is the p-value affected ?

Exercise 2

Part 2 - Inferential Statistics

Height ~ Sex

- You have looked at if and how the distribution of height differs based on sex in the previous steps. Now we’re going to approach this hypothesis scientifically. Run the ‘t_test’ function (pay attention to the underscore) like shown below with all four subsampled datasets.
  - brfs_sub10 %>% t_test(Height ~ Sex, conf.level = 0.95, detailed = TRUE)
  - Look at how to interpret the data frame you get from the function’s web page.
  - How does the effect size, confidence interval and p-value changes with various sample sizes ? Is one result one significant than other ? Is one comparison more effective than other ?

LMU KLINIKUM
Gemeinsam. Fürsorglich. Wegweisend.

HELMHOLTZ MUNICH

KimHellmuth Lab

DR.VON HAUNERSCHES KINDERSPITAL
Helfen Forschen Heilen. Seit 1846

Thanks for your attention!

Contacts:

Ekin Yaman **Dr. med. Sarah**
Kim-Hellmuth
✉: Ekin.Yaman@med.uni-muenchen.de https://www.ccrc-hauner.de/kim-hellmuth-labor

Dr. med. Paula Rothämel
✉: Paula.Rothaemel@med.uni-muenchen.de