

Nice to see you!

- Connect to **WiFi** with your Eduroam account or hit me up!
- **Slides and hands-on materials** are accessible on Github:

<https://github.com/Data-Science-Center-UB/Intro-Quantitative-Analysis-R>

Please download the folder **Materials** to your local machine

- You need a local **R/R Studio** installation:
 - R : <https://cran.r-project.org>
 - RStudio: <https://www.rstudio.com/products/rstudio/>



No Fear of Numbers

Introduction to Quantitative Data Analysis in R

CONTACT

Dr. Susanne de Vogel

Data Scientist | Help-desk

devogel@uni-bremen.de

dsc-ub.de

Schedule For Today

- 🕒 09:00 **Welcome and Introduction**
- 🕒 09:30 Basic Statistical Concepts
- 🕒 09:50 First Steps in R: Exploring Data
- 🔍 11:00 Univariate Analysis
- 🍽️ 12:30 **Lunch break**
- 📊 13:30 Bivariate Analysis
- 🌐 15:00 Multivariate Analysis
- 👋 16:25 **Farewell**

When Non-Statistics
Students use
Statistics



Idea

We'll go through four practical parts today:

- 1 First Steps in R: Exploring data
- 2 Univariate analysis
- 3 Bivariate analysis
- 4 Multivariate analysis

→ Each part starts with a short input introducing the theory and some basic R concepts.

→ Then you'll have time to work on the exercises independently.

-  **Exercise sheet (without solutions)** – for those who are already more proficient in R
-  **Exercise sheet (with solutions and detailed explanations)** – for those who prefer to follow along conceptually

→ After each hands-on block, we'll go through the solutions together and discuss different approaches.

→ Don't worry if you can't finish everything — all materials remain available afterwards.

Housekeeping

- Slides and hands-on materials are accessible on Github:

<https://github.com/Data-Science-Center-UB/Intro-Quantitative-Analysis-R>

Please download the folder **Materials** to your local machine.

- You need a local R/R Studio installation:

- R : <https://cran.r-project.org>
- RStudio: <https://www.rstudio.com/products/rstudio/>

- Please sign the attendance sheet.
- Certificates will be sent after the training.
- Questions welcome anytime ☺

Who am I?

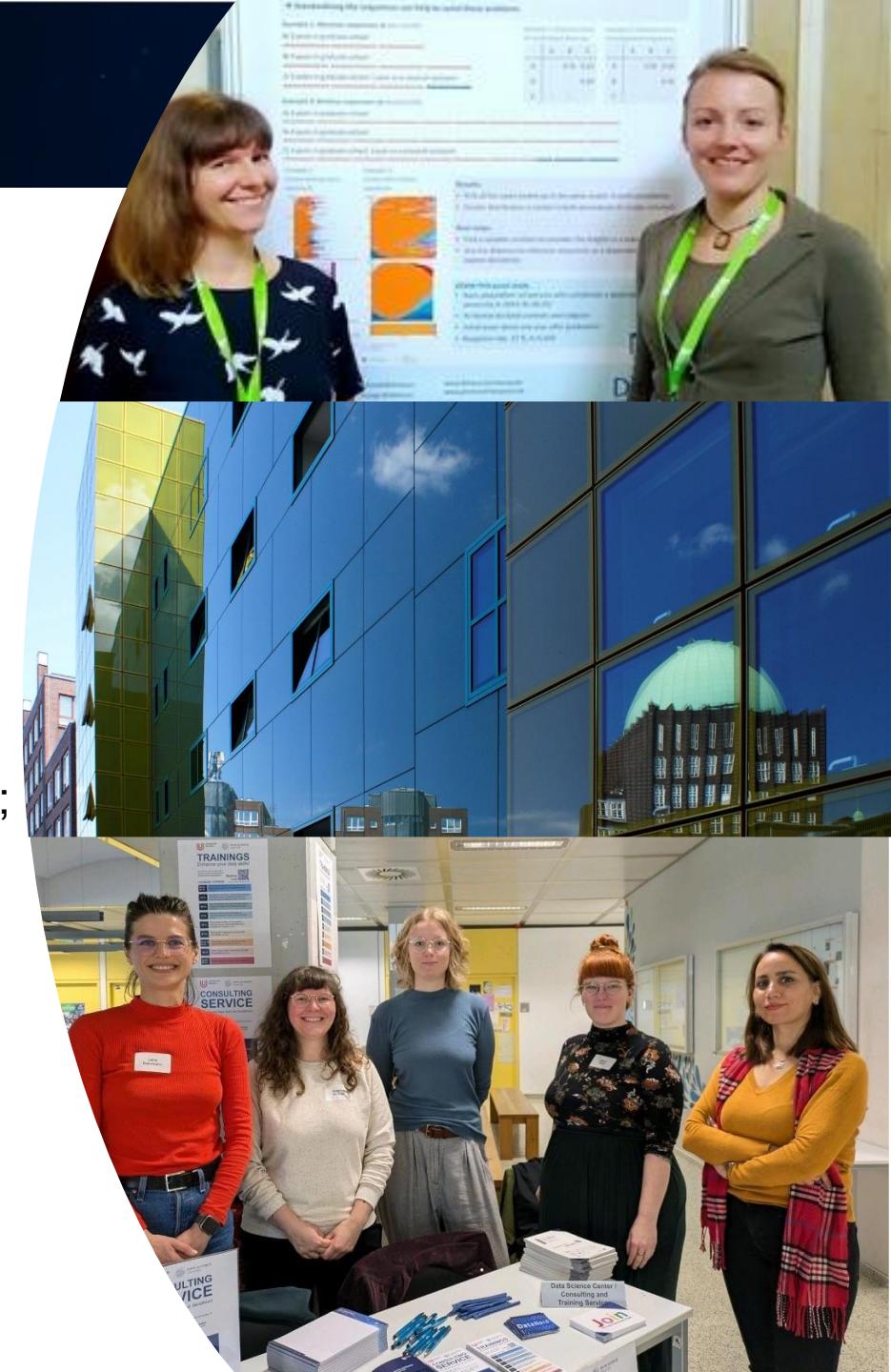
Susanne de Vogel (she/her)

Social Scientist

- Diploma in Social Sciences (2013), University of Cologne and Utrecht University
- PhD in Sociology (2019), Martin-Luther-University Halle-Wittenberg
- Research Associate at DZHW German Centre for Higher Education Research and Science Studies (2013-2020), Hannover
- National Academics Panel Study (Nacaps) and PhD Graduate Panel; survey design, panel data analysis, research data management
- Research on educational inequalities, early career researchers, academia

Since May 2024 Data Scientist at the Data Science Center

- Training and consulting for researchers
- Data science and research data management



Brief round of introductions



- Name (pronouns)
- Institution
- Discipline
- Previous experience with quantitative data & programming software

Basic Statistical Concepts

Qualitative vs. Quantitative Research

Qualitative Research

Focus: Explores individual experiences, subjective perspectives.



Data type: Non-numerical data (e.g., words, narratives, observations).

Format: Text, audio, video.

Method: Interviews, case studies, focus groups, observations.

Analysis: Exploratory, interpretative, context-specific.

Quantitative Research

Focus: Tests hypotheses, looks for patterns and (causal) relations, makes predictions.



Data type: Numerical data (e.g., statistics, scores, ratings).

Format: Tabular data sets.

Method: Surveys, experiments, standardized tests.

Analysis: Descriptive and inferential statistics.



What is statistics?

Statistics is a branch of [applied mathematics](#) that involves

- Collection
- Description
- Interpretation
- Inference of conclusions of quantitative data

[Two main types](#) of statistics:

1. Descriptive statistics
2. Inferential statistics

Population (Grundgesamtheit)

The complete set of individuals that share common characteristics and are the **focus of a quantitative study**.

→ Group of people you are interested in.



Every individual who registered as a Ph.D. student at a German HEI in the 2024 academic year



Persons living in private households in Germany and aged 16-75 at the time of recruitment.



Full-time academic and artistic staff at German universities and equivalent institutions at the time of recruitment.



All people aged 18 to 49 living in Germany in 2020.

Sample (Stichprobe)

The specific group of individuals **chosen from the target population** to participate in the study.

- The smaller group we actually collect data from.
- Selected based on certain criteria or sampling methods to ensure that it **accurately represents the larger population**.



Full census.



Random sample of 60,002 addresses, disproportionately stratified (with an oversampling of female professors and postdocs).

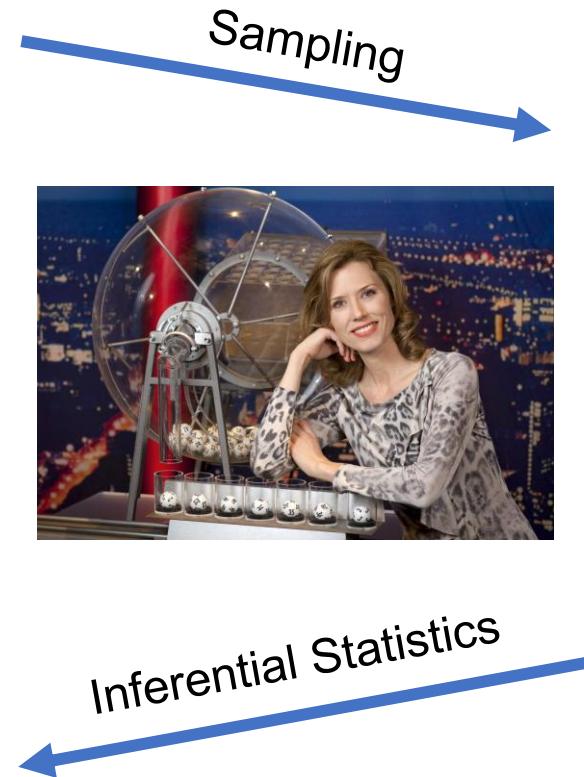
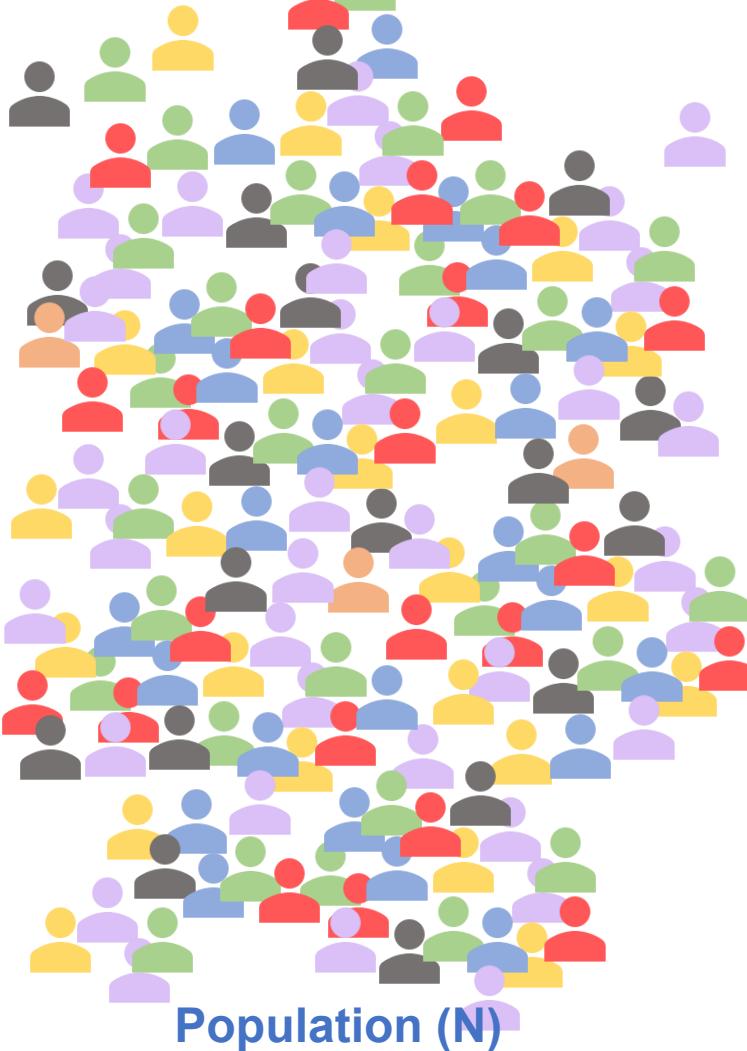


Multi-stage random samples which are regionally clustered, 2012 and 2014 samples are additionally clustered in households.



Random two-stage sampling from the population registers of selected municipalities in Germany.

What has sampling to do with statistics?



Sample (n)

Descriptive Statistics

What is the difference?

Descriptive statistics

Goal: Describing the data already available (sample data).

Example:

“What is the average age of our survey participants?”

“Does the average age of our participants differ by gender?”

Method:

- Absolute and relative frequencies
- Central tendency (Mean, Mode, Median), Shape
- Variability (Variance, Standard Deviation)



Inferential Statistics

Goal: Make inferences (draw conclusions) from the sample and generalize them to the population.

Example:

“Do women and men differ in their average income?”

“Are males more likely to gain a high salary than females?”

Method:

- Hypothesis testing (Chi²-Test, t-test, ANOVA, Correlation)
- Linear- or logistic regression



What do I need to consider?

Descriptive statistics

Tell you **what your data look like.**

↳ “What is happening in my data?”

→ Works well with small samples and non-random sampling.

Inferential statistics

Tell you **whether your findings are likely a real phenomenon in your population or just coincidence.**

↳ “Is it statistically significant?”

→ Only works well with **sufficient sample size**

→ Small samples mean more uncertainty and less power to detect effects.

→ Requires **random sampling** to ensure the sample is representative of the population.

How do I know that my sample size is sufficient?



#...i think this meme demonstrated the importance of sample size better and more efficiently than any science class i've ever taken

Tools and Guidelines

[Australian Bureau of Statistics Sample Size Calculator](#)

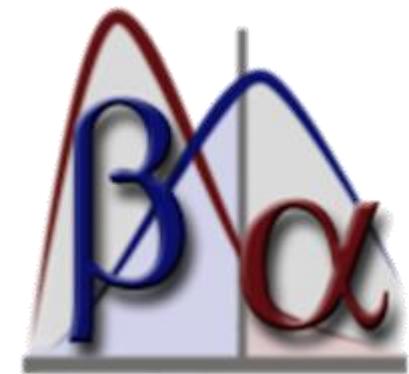
[UK Sample Size Calculator](#)

[Gesis Sample Size Calculation For Complex Sampling Designs](#)

Programmes and packages

[G*Power](#) Desktop Application

R [pwr](#) and [practools](#) packages



What is a variable?

In quantitative analysis, we investigate **attributes** (like characteristics) of our observations (one person, case), e.g. age, gender, income, self-efficacy score, opinions

- Variables are the attributes we want to investigate
- Variables can vary between people and/or over time

Example:

What does the distribution of **age** look like in my sample?

What does the **age** distribution of the population in Germany look like on 2nd December 25?

Are **women** more likely to reach the **age** of 100 in Germany than men?

Does the mean **age** differ between **gender** groups in my sample?

Which **characteristics** influence the **age** people reach in Germany?

Definition

How is each variable **measured**?

What is the **mathematical nature of the values** assigned to each variable?

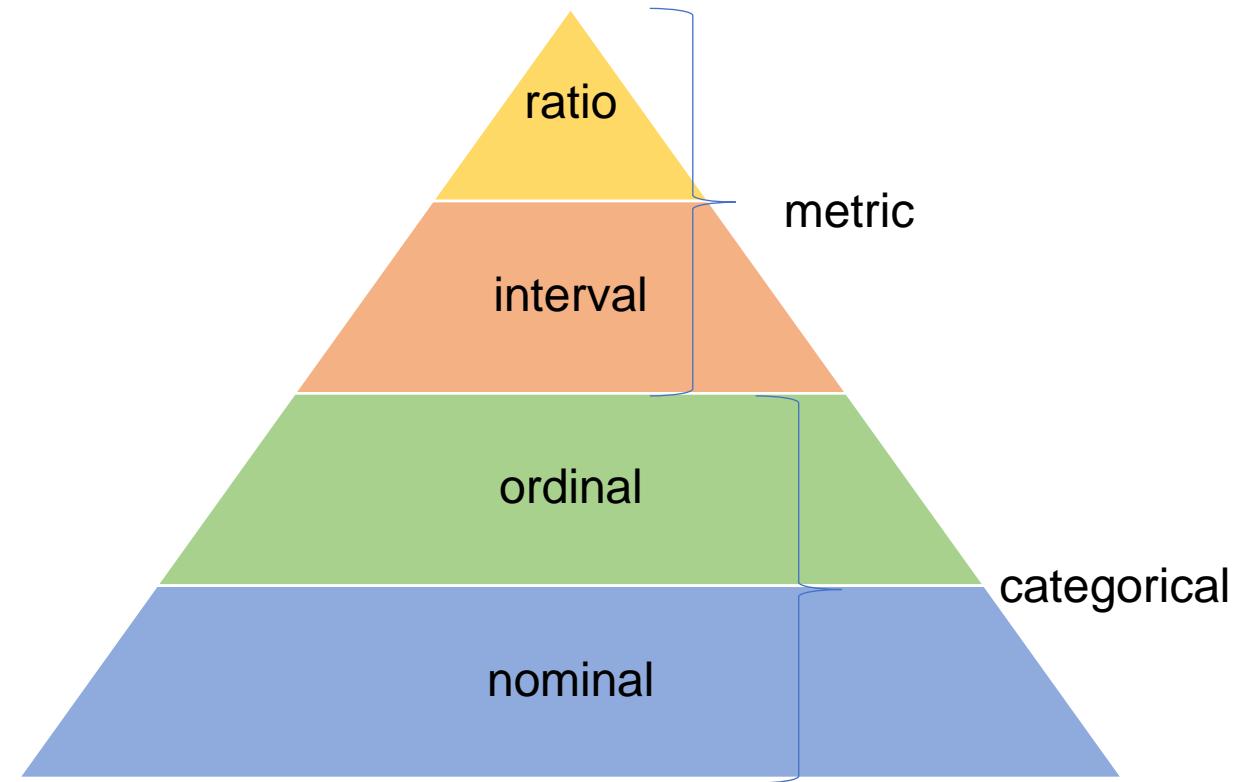
Importance

Level of measurement determines the type of **statistical analysis** you can carry out on the outcomes of your questions.

These levels reflect how data can be:

- Quantified
- Analyzed
- Interpreted

Levels of Measurement



Nominal data

Nominal data divides variables into mutually exclusive, labeled categories.

Examples

Eye color



Blue Brown Green

Smartphone



iPhone Samsung Moto

Transport



Bus Train Car

How is nominal data analyzed?

Descriptive statistics:
Frequency distribution
and mode

Non-parametric
statistical tests

Ordinal data

Ordinal data classifies variables into categories which have a natural order or rank.

Examples

School grades



Education level



Seniority level



How is ordinal data analyzed?

Descriptive statistics:
Frequency distribution,
mode, median, and range

Non-parametric
statistical tests

Interval data

Interval data is measured along a numerical scale that has equal intervals between adjacent values.

Examples

Temperature



IQ score



Income ranges



How is interval data analyzed?

Descriptive statistics: Frequency distribution; mode, median, and mean; range, standard deviation, and variance

Parametric statistical tests (e.g. t-test, linear regression)

Ratio data

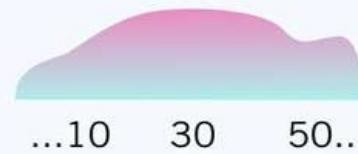
Ratio data is measured along a numerical scale that has equal distances between adjacent values, and a true zero.

Examples

Weight in KG



Number of staff



Income in USD



How is ratio data analyzed?

Descriptive statistics: Frequency distribution; mode, median, and mean; range, standard deviation, variance, and coefficient of variation

Parametric statistical tests (e.g. ANOVA, linear regression)

Comparison

THE FOUR LEVELS OF MEASUREMENT:

	Nominal	Ordinal	Interval	Ratio
Categorizes and labels variables	✓	✓	✓	✓
Ranks categories in order		✓	✓	✓
Has known, equal intervals			✓	✓
Has a true or meaningful zero				✓

First Steps in R

What is R?

R is a programming language designed for

-  **Data manipulation** – cleaning, transforming, and preparing datasets.
-  **Statistical analysis** – running summaries of your data, models and tests of your hypotheses.
-  **Data visualization** – creating clear and beautiful graphs and plots.



Why do we use R?

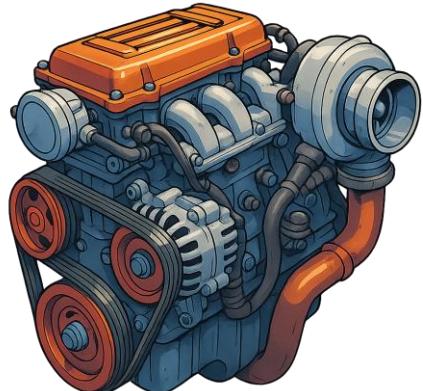
- Free and open source.
- Great for transparency and reproducibility
- Powerful and state of the art

What is the difference between R and RStudio?



R is the **programming language** itself.

- Executes the code and does the calculations;
- Runs in the background.



R Studio is a **development environment** (IDE) to make R easier to use.

- Provides the graphical interface to write, run and visualize code;
- Provides windows for code, output and plots.



You

You as a researcher use R working in R Studio.

- You have control ☺

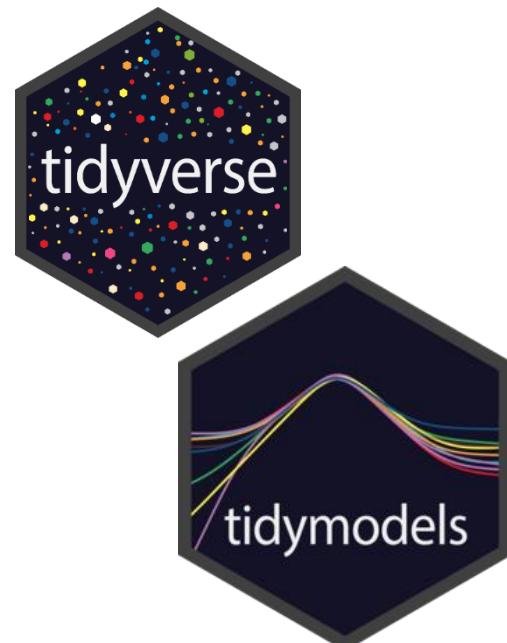


How can I perform quantitative data analysis in R?

Base R

Core set of functions and tools that are included with R by default.

It allows you to do many data analysis tasks, but the syntax is inconsistent and can be confusing for beginners.



Tidyverse & tidymodels package

To make data work easier and more consistent, the R community has developed many **packages**.

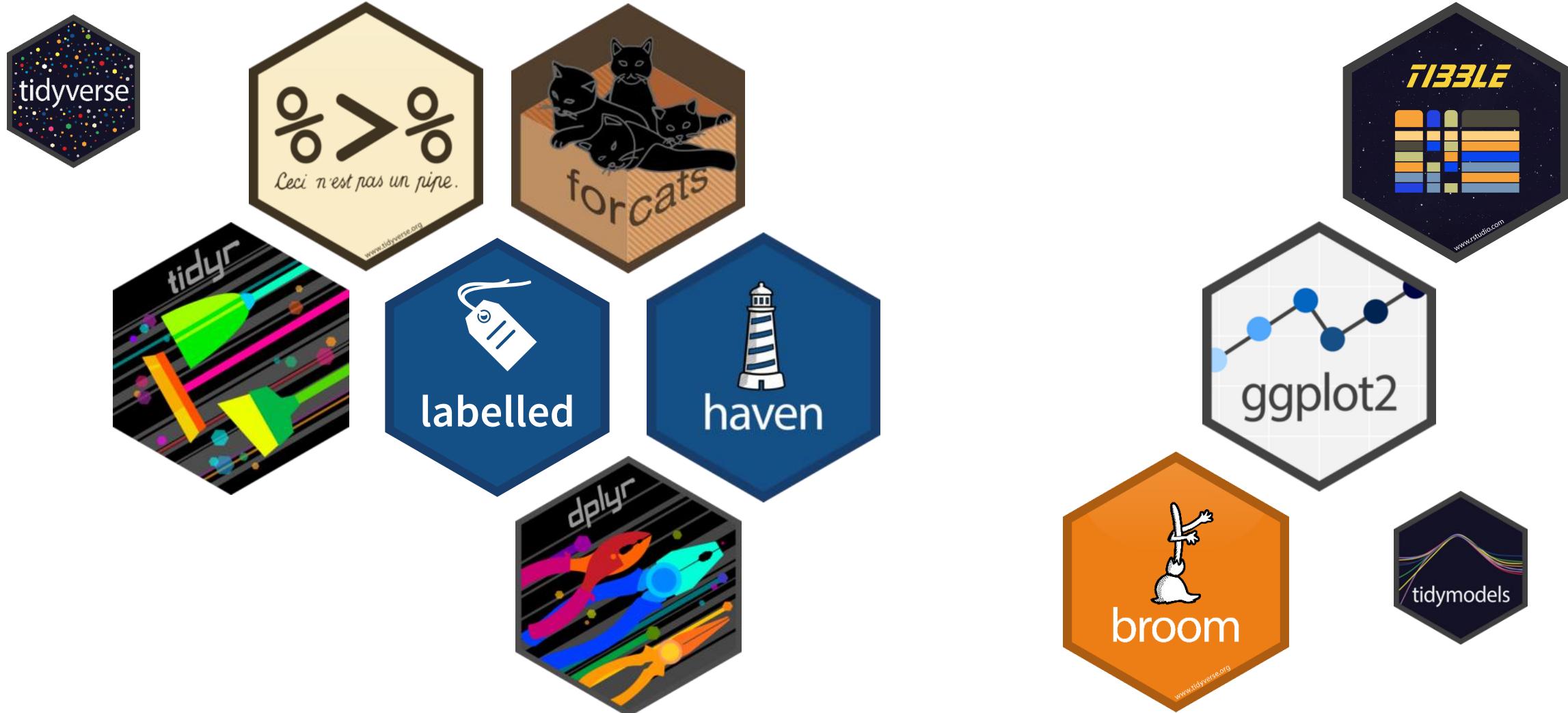
Packages are collections of functions designed to extend the capabilities of base R.

Tidyverse and tidymodels are a collection of R meta-packages designed for **data science workflows**.

- Tidyverse for data wrangling, tidying and visualization.
- Tidymodels for modeling and statistical analyses.

→ **Code has a consistent syntax and resembles natural language.**

Tidyverse and Tidymodels in R



Basic Vocabulary

Structure in R

Observation/Case/Record:

One row in the dataset, usually one respondent or unit.

Observation/case/
record

Variable:

One measured characteristic/attribute

Value:

The specific entry for a variable in one case (e.g., 1)

Data frame:

Dataset, the full table:

- rows = cases
- columns = variables
- cells = values

Data frame

pid	adbi01	adbi04c	variable
1	2	2017	
2	9	2014	
3	10	2018	
4	15	2018	
5	17	2017	
6	21	2018	value
7	23	2017	
8	24	2017	
9	26	2016	
10	27	2012	
11	31	2018	
12	39	2017	
13	43	-989	
14	49	2016	
15	54	2018	
16	55	2017	
17	59	2018	
18	63	2015	

Example: Data frame in data viewer using `View()` function

Elements of a variable

Variable name:

The short technical name used in the dataset (e.g. adem01).

Variable label:

A descriptive text for the variable (e.g. "Gender").

Value:

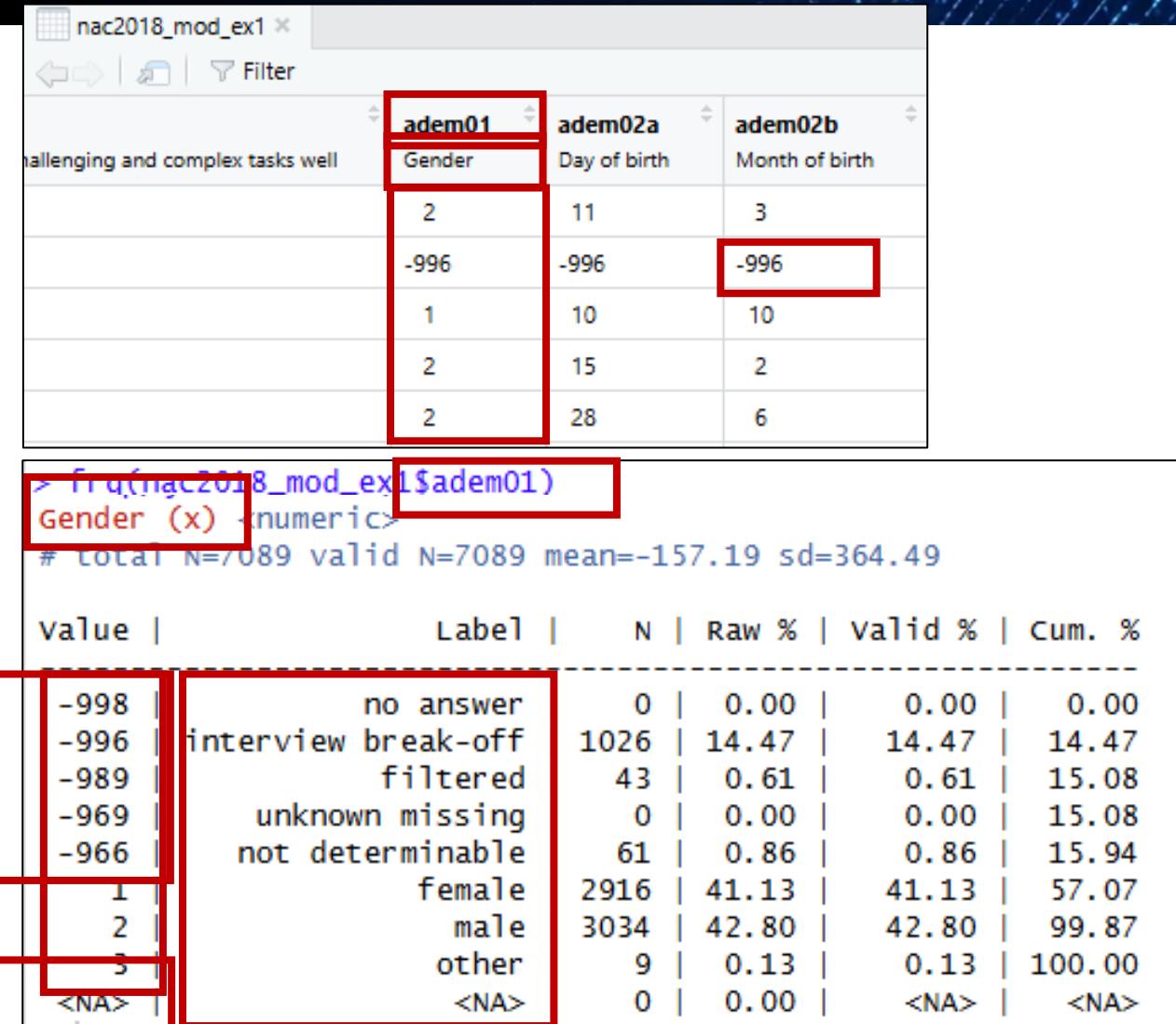
The specific entry for a variable in one case (e.g., 1).

Value label:

A description of what a value means (e.g., 1 = female).

Missing value:

Data not available, either left empty or coded (e.g. -996 = interview break-off).



Example: Frequency table of variable `adem01` measuring gender using `freq()` function

Levels of Measurement & data types in R

The basic storage mode of an object in R:

Nominal, Ordinal, Interval, Ratio	<code>double, integer</code>
Strings	<code>character</code>
Binary (special variables with two categories)	<code>logical (TRUE/FALSE) (booleans)</code>

- R often imports data as `character`, `double`, or `integer` – other variable classes need to be created explicitly e.g. for nominal (`factors`), ordinal (`ordered factors`) or date/time (`Date`) variables.
- `Characters` usually need to be `converted` into another data type so that R can handle the variables correctly in analyses.

Levels of measurement & data types in R

We can use the `typeof()` function to check the data type of a variable:

```
> typeof(mydata$adem01)
[1] "double"
```

```
> typeof(mydata$adbi10a)
[1] "character"
```

Data classes in R

The class tells R how to treat an object and which methods apply.

Factor:

Nominal variable, internally stored as `integer` + labels.

Ordered factor:

Ordinal variable.

`data.frame`:

Rectangular table, base R's main data structure.

`tibble`:

Data frame in `tidyverse`.

`matrix`:

2D array (same type only).

`array`:

N-dimensional generalization of matrix (same type only).

`list`:

Generic container for mixed objects, can hold vectors, data frames, etc.

→ **Classes** = higher-level structures or behaviors **built on types**.

Data classes in R

We can use the `class()` function to check the class of an object:

```
> class(mydata4)
[1] "tbl_df"      "tbl"        "data.frame"
> |
```

Dataset;

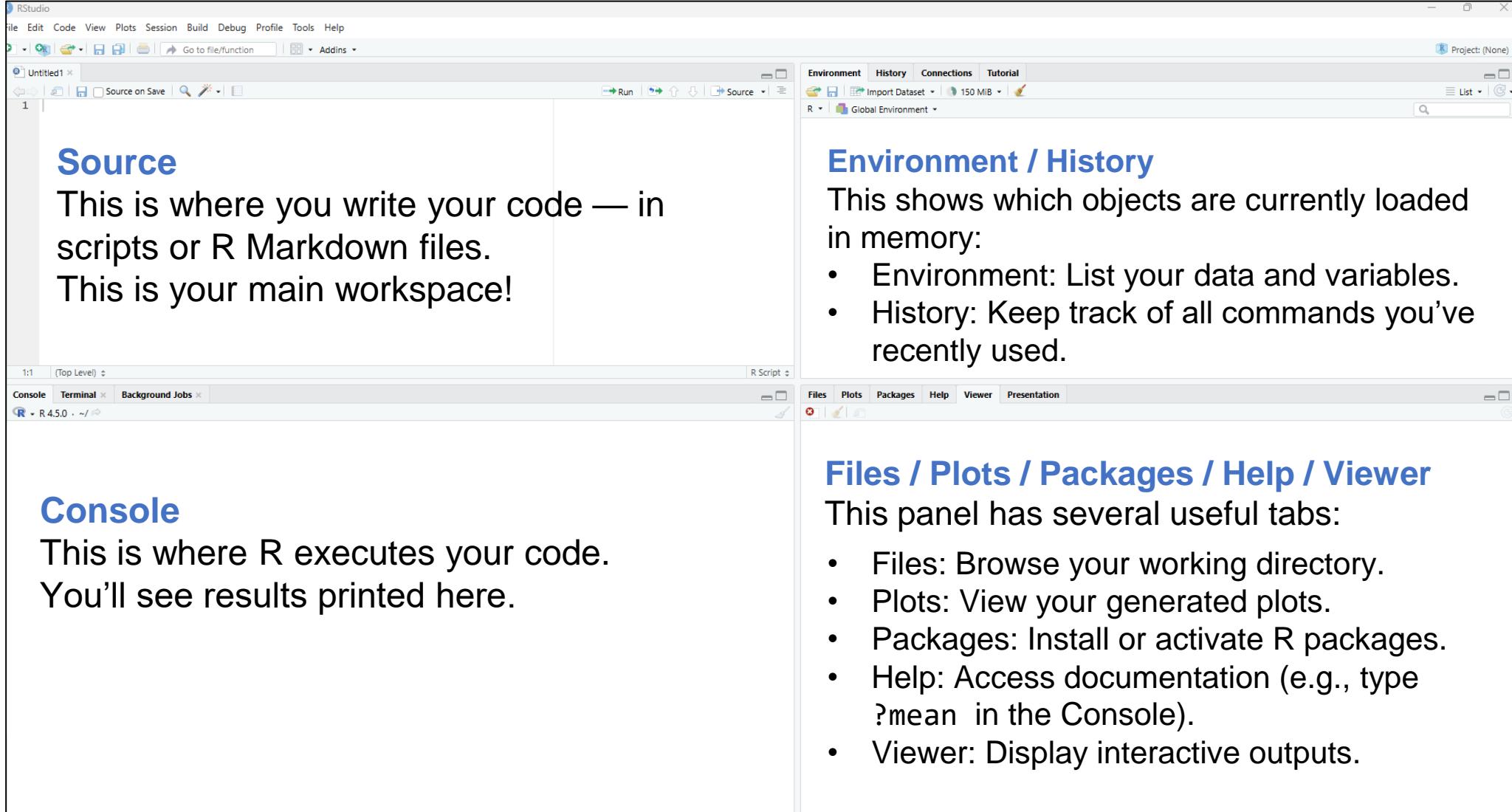
```
> class(mydata4$adbi01)
[1] "haven_labelled_spss" "haven_labelled"      "vctrs_vctr"       "double"
> |
```

PhD status (ongoing, completed., etc..)
not defined as factor yet;

```
> class(mydata4$adem02f_g1)
[1] "ordered" "factor"
> typeof(mydata4$adem02f_g1)
[1] "integer"
> |
```

Age categories
defined as factor.

How does RStudio look like?



Source
This is where you write your code — in scripts or R Markdown files.
This is your main workspace!

Environment / History
This shows which objects are currently loaded in memory:

- Environment: List your data and variables.
- History: Keep track of all commands you've recently used.

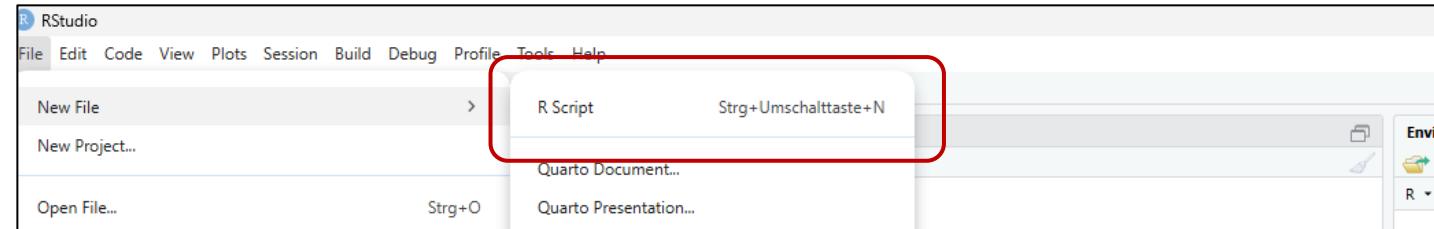
Console
This is where R executes your code.
You'll see results printed here.

Files / Plots / Packages / Help / Viewer
This panel has several useful tabs:

- Files: Browse your working directory.
- Plots: View your generated plots.
- Packages: Install or activate R packages.
- Help: Access documentation (e.g., type `?mean` in the Console).
- Viewer: Display interactive outputs.

Initial Steps

- 0.1. Open Rstudio and open new script (File → New File → R script), save.



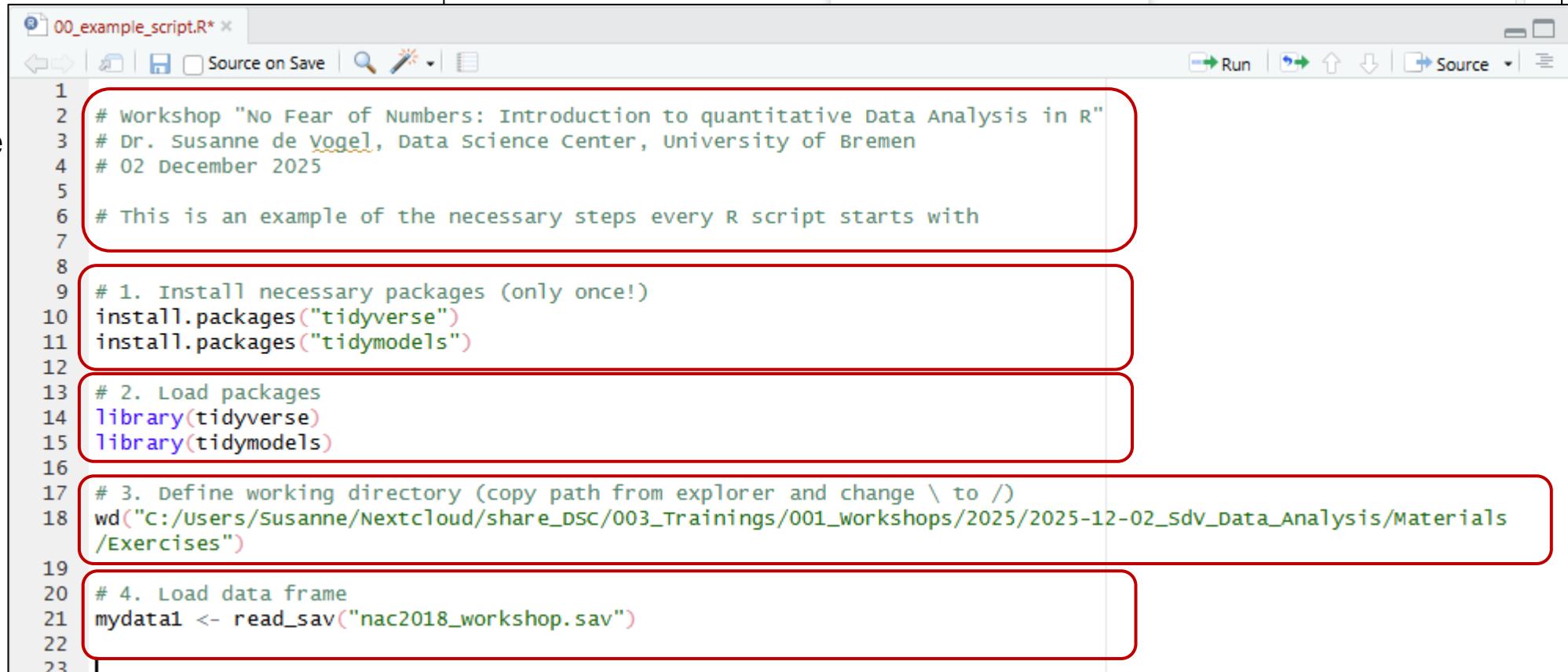
- 0.2. Create a header with title, author, date and description.

1. Install packages.

2. Load packages.

3. Set working directory.

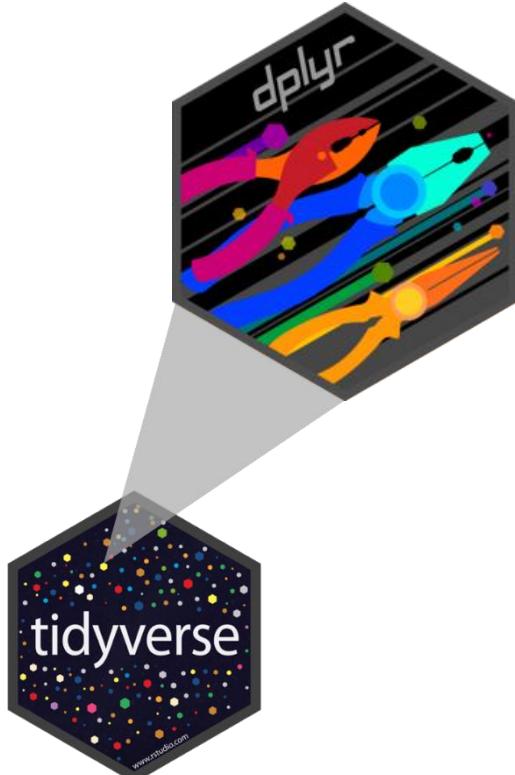
4. Import dataset as data frame.



```
1 # Workshop "No Fear of Numbers: Introduction to quantitative Data Analysis in R"
2 # Dr. Susanne de Vogel, Data Science Center, University of Bremen
3 # 02 December 2025
4
5 # This is an example of the necessary steps every R script starts with
6
7
8
9 # 1. Install necessary packages (only once!)
10 install.packages("tidyverse")
11 install.packages("tidymodels")
12
13 # 2. Load packages
14 library(tidyverse)
15 library(tidymodels)
16
17 # 3. Define working directory (copy path from explorer and change \ to /)
18 wd("C:/Users/Susanne/Nextcloud/share_DSC/003_Trainings/001_workshops/2025/2025-12-02_sdv_data_Analysis/Materials/Exercises")
19
20 # 4. Load data frame
21 mydata1 <- read_sav("nac2018_workshop.sav")
```

Rules of functions

- Functions say **what to do with the data frame**.
- First argument calls the **data frame** you want to work with (and, separated with a \$-sign, the **variable** you want to work with).
- Always return a **data frame**.
- **Don't modify** in place:
 - save as **new object**, needs to be explicitly saved.

A screenshot of an RStudio session showing two code snippets. The first snippet uses `names` to check variable names in a dataset, with the entire command highlighted by a red box. The second snippet uses `frq` to check frequencies of a specific variable, also with the entire command highlighted by a red box. Both snippets show the resulting output below them.

```
```{r names}
Check variable names in the dataset
names(mydata1)
```
[1] "pid"      "adbi01"   "adbi10a"  "adbi04c"  "adbi06c"  "afin04a"  "afin04b"  "afin04c"  "afin04d"  "afin04e"  "afin04f"
[12] "afin04g"  "afin04h"  "afin04i"  "afin04j"  "afin04k"  "alcd07"   "apsy05a"  "apsy05b"  "apsy05c"  "adem01"   "adem02a"
[23] "adem02b"  "adem02c"  "adem03"   "adbi04b"  "adbi06b"  "adbi04a"  "adbi06a"
```



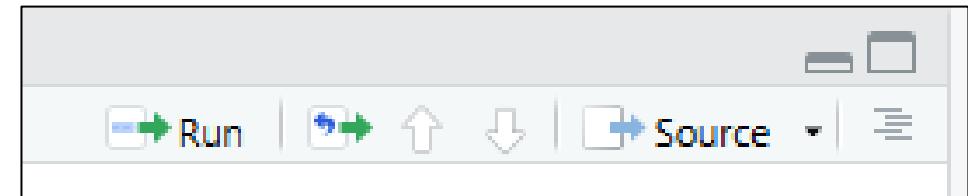
```
```{r freq pid, results='hide'}
checking frequencies of pid (output suppressed)
frq(mydata1$pid)
```
unique identifier (x) <numeric>
# total N=7089 valid N=7089 mean=14183.45 sd=8225.83

value	N	Raw %	valid %	Cum. %
2 | 1 | 0.01 | 0.01 | 0.01
1 | 1 | 0.01 | 0.01 | 0.02
```

Running code

Run one line of code:

1. Place the cursor on the line you want to run.
2. Press **Ctrl + Enter (Windows/Linux)** or **Cmd + Enter (Mac)** or click “Run” at the top of your script editor.



Run several lines:

1. Highlight the lines you want to run.
2. Press **Ctrl/Cmd + Enter** again or click “Run” at the top of your script editor.

Run the whole script:

1. Click “Source” at the top of the script editor.
2. Or press **Ctrl + Shift + Enter**.

See the results:

The output appears in the **Console** or in the **Environment** pane.

→ For example, after `mydata <- read_sav("data.sav")`, you'll see mydata listed in your Environment.

Running code: Troubleshooting

What to do if nothing happens?

→ Check that you're typing in the Script Editor or Console, not somewhere else.

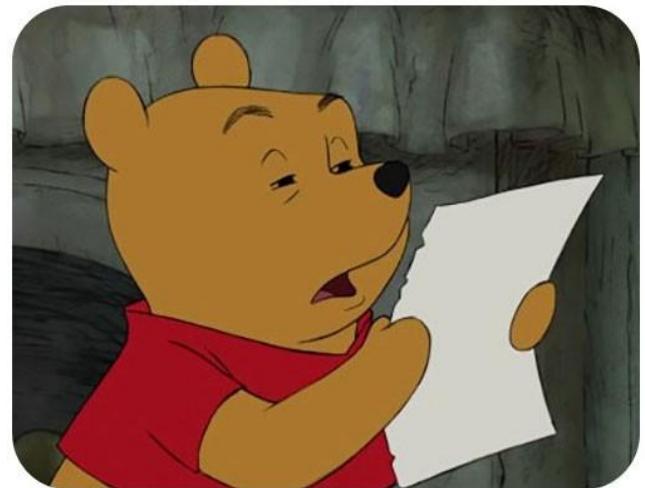
What to do if there is an error message?

- Misspellings, typos, too many/too few brackets, wrong quotation marks.
- R is case-sensitive.
- Make sure you run your code in the right order (e.g., load packages before using their functions).
- You try to use an object that doesn't exist (yet).
- Wrong file path.
- Wrong data type or missing values.
- Function arguments missing or incorrect.

→ Try running the code line by line to find the issue.

→ Restart R Session (Session → Restart R) and run your code again from the beginning.

Error at line 132 but to fix it you add a parenthesis at line 120



Is there a life beyond the PhD?

In this workshop, we will work with one coherent example throughout all four parts.

We focus on **two outcomes of interest**:

- **Overall life satisfaction (0-10 scale)**
- **Children (yes/no)**

We will examine how these outcomes are related to different aspects of doctoral researchers' lives and backgrounds:

PhD and Work conditions

- Status of the doctorate
- Discipline
- Emotional support during PhD
- Perceived scientific pressure
- Gross income

Attitudes and Well-Being

- Satisfaction with work-life balance
- Relationship with parents
- Self-Efficacy

Demographics

- Gender
- Age in years
- Highest vocational degree of parents
- Country of birth
- Relationship status

National Academics Panel Study (Nacaps)

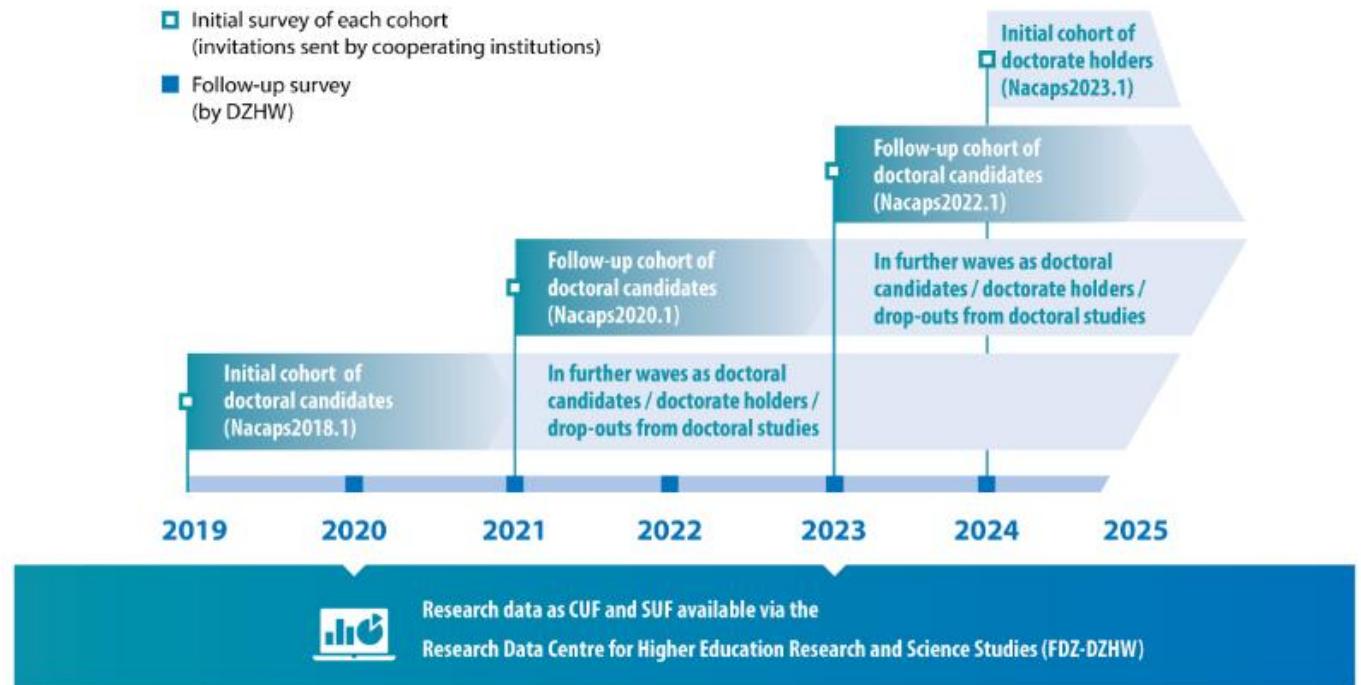
Project duration: Januar 2017 – ongoing (funded by BMBF)

Target population: full census survey

- All persons who started/completed a doctorate at a German HEI
- 66 participating universities

Survey design: multicohort longitudinal study

- Initial wave about one year after starting/graduating PhD
- Annual follow-up surveys
- New cohort every other year (doctoral candidates since 2018; doctorate holders since 2024)
- Online



Your research strategy

To explore these questions, we will:

1. start by exploring the data and getting to know our sample
2. look at univariate distributions of key variables (e.g. life satisfaction, number of children, income),
3. examine bivariate relationships (e.g. does life satisfaction differ between people of different phd status and gender)
4. finally fit multivariate models:
 - a linear regression with life satisfaction
 - a logistic regression on probability of having children

Your turn!

1. Copy folder **Materials** from to your local machine.

→ <https://github.com/Data-Science-Center-UB/Intro-Quantitative-Analysis-R>
→ Path to subfolder **Exercises** will be your **working directory**.
→ Folder content: data sets, exercises without & with solution

2. Do exercise “**01_qa_exploring_ex**”.

Version with solution: **01_qa_exploring_ex_solution**



Univariate Statistics

What is it?

Univariate statistics deal with the **analysis of a single variable** at a time.

The goal is to describe and summarize its **distribution**.

Why univariate analyses are important:

- **Describe your sample**, that means summarize who or what you studied, e.g., age, gender, education.
- **Check data quality**, e.g. detect missing values, outliers.
- Check the **shape of the distribution**, e.g. skewness and kurtosis as a foundation for further analyses;
→ Many parametric tests, e.g. t-test or ANOVA, assume normally distributed variables.
- Provide **clear and interpretable** tables and charts for publications.

What decisions to make?

1. ⚡ Which variables to analyze?

Depends on your **research question**, for example:

- Your outcome(s) of interest,
- Attributes that might affect outcome(s) of interest, e.g. group differences, influencing factor,
- Demographics.

2. ☰ Which type of analysis and visualization to use?

Depends on the variable's **level of measurement**.

| Level of measurement | Analysis | Visualization |
|----------------------|---|--------------------|
| min. nominal | Absolute/relative frequencies, mode | Bar chart |
| min. ordinal | cumulative frequencies,
median, percentiles | Bar chart |
| Interval/ratio | Mean, confidence intervals
Variability (Standard, Deviation)
Shape (Skewness, Kurtosis) | Boxplot, Histogram |

Absolute and relative frequencies

Absolute frequencies (n): The number of cases in each category or value.

→ *How many?*

Relative frequencies (%): The proportion or percentage of cases in each category.

→ *How much out of the total?*

Example:

| Gender | Absolute frequency (n) | Relative frequency (%) |
|---------------|------------------------|------------------------|
| Female | 250 | 50 |
| Male | 200 | 40 |
| Diverse/Other | 50 | 10 |
| Total | 500 | 100 |

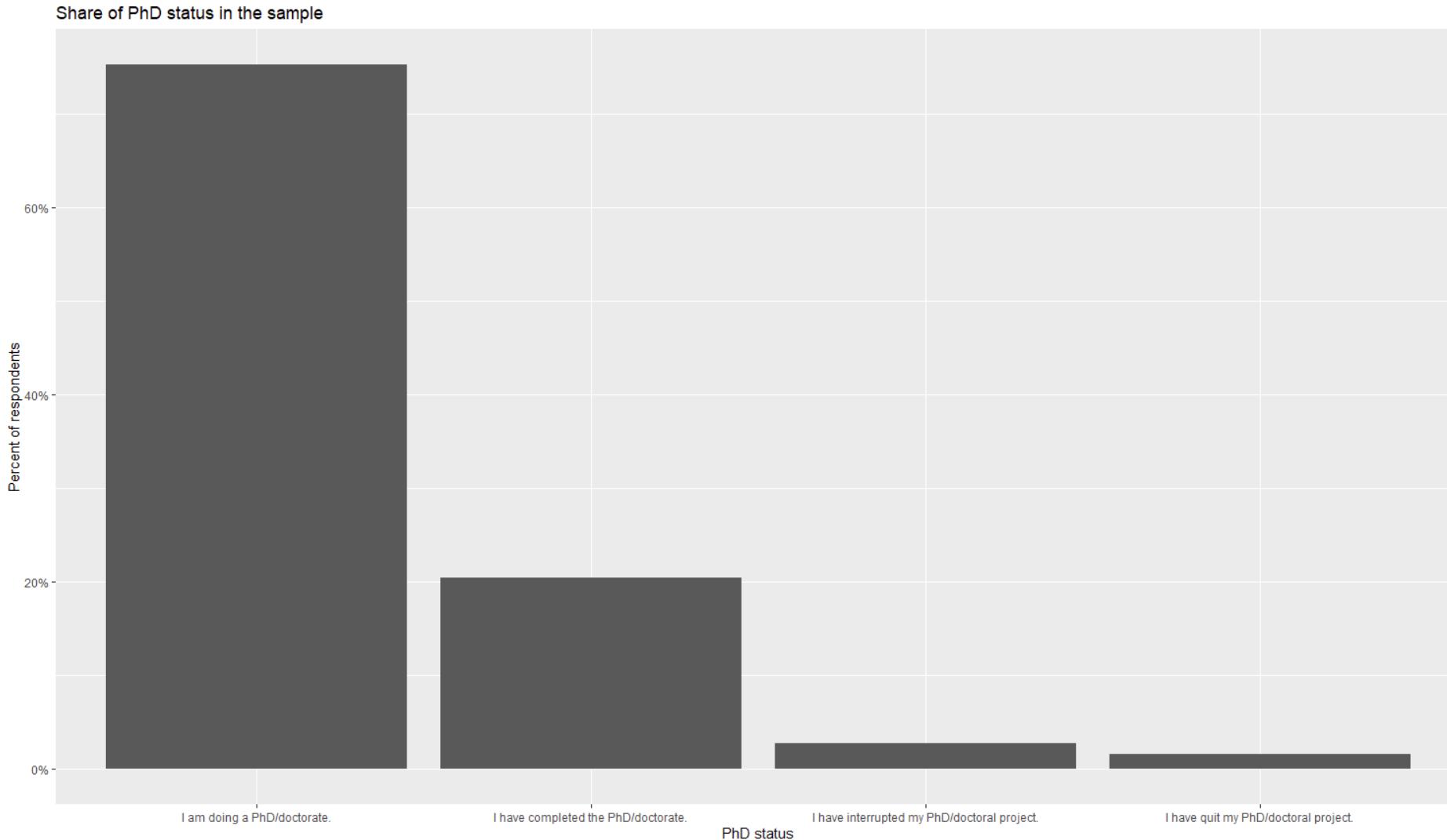
Frequency table reporting

Summarize absolute and relative frequencies in frequency table:

- Absolute (n) and relative frequencies (%) in columns;
- Variable categories in row;
- Cells contain absolute numbers and percentages in each category;
- Total row at the bottom to show the total sample corresponding to 100%.
→ Categories should add up to 100% ☺

| Gender | n | % |
|---------------|------------|------------|
| Female | 250 | 50 |
| Male | 200 | 40 |
| Diverse/Other | 50 | 10 |
| Total | 500 | 100 |

Visualization of absolute and relative frequencies: Bar chart



Example: Bar chart of variable `bdbi01` measuring Status of the doctorate using `ggplot()`

Measure of central tendency: Mode

The mode (in German: Modus) is the [most frequent value](#) of a variable.

→ Helps to describe the most typical or common category of a variable in your sample.

Example:

| Gender | Absolute frequency (n) | Relative frequency (%) |
|---------------|------------------------|------------------------|
| Female | 250 | 50 |
| Male | 200 | 40 |
| Diverse/Other | 50 | 10 |
| Total | 500 | 100 |

When to use:

Works well for [nominal](#) or [categorical](#) data.

Cumulative frequencies

Cumulative frequencies: Running total of absolute or relative frequencies.

- *How many cases fall up to including a certain category?*
- Useful only for variables where categories have a natural order.

Example:

| Highest vocational degree | Absolute frequency (n) | Relative frequency (%) | Cumulative frequency (%) |
|-----------------------------|------------------------|------------------------|--------------------------|
| no vocational qualification | 25 | 5 | 5 |
| Apprenticeship | 250 | 50 | 55 |
| Technical/master school | 50 | 10 | 65 |
| Bachelor/master degree | 150 | 30 | 95 |
| PhD | 25 | 5 | 100 |
| Total | 500 | 100 | |

65 % of respondents finished technical/master school or lower, while 35 % have an university degree.

Percentiles

Percentiles divide an ordered distribution into 100 equal parts.

Each percentile shows the **value below which a certain percentage of cases fall**.

→ *How is the variable distributed?* — not just the average, but also *how spread out* it is.

→ They are based on cumulative frequencies, so they only make sense for ordinal or metric variables.

10th percentile (P10)

→ 10 % of cases are below or equal this value

25th percentile (P25)

→ 25 % of cases are below or equal to this value

50th percentile (P50 = Median)

→ half of the data below, half above

75th percentile (P75)

→ 75 % of cases fall below this value

Minimum (min)

→ smallest observed value in a variable

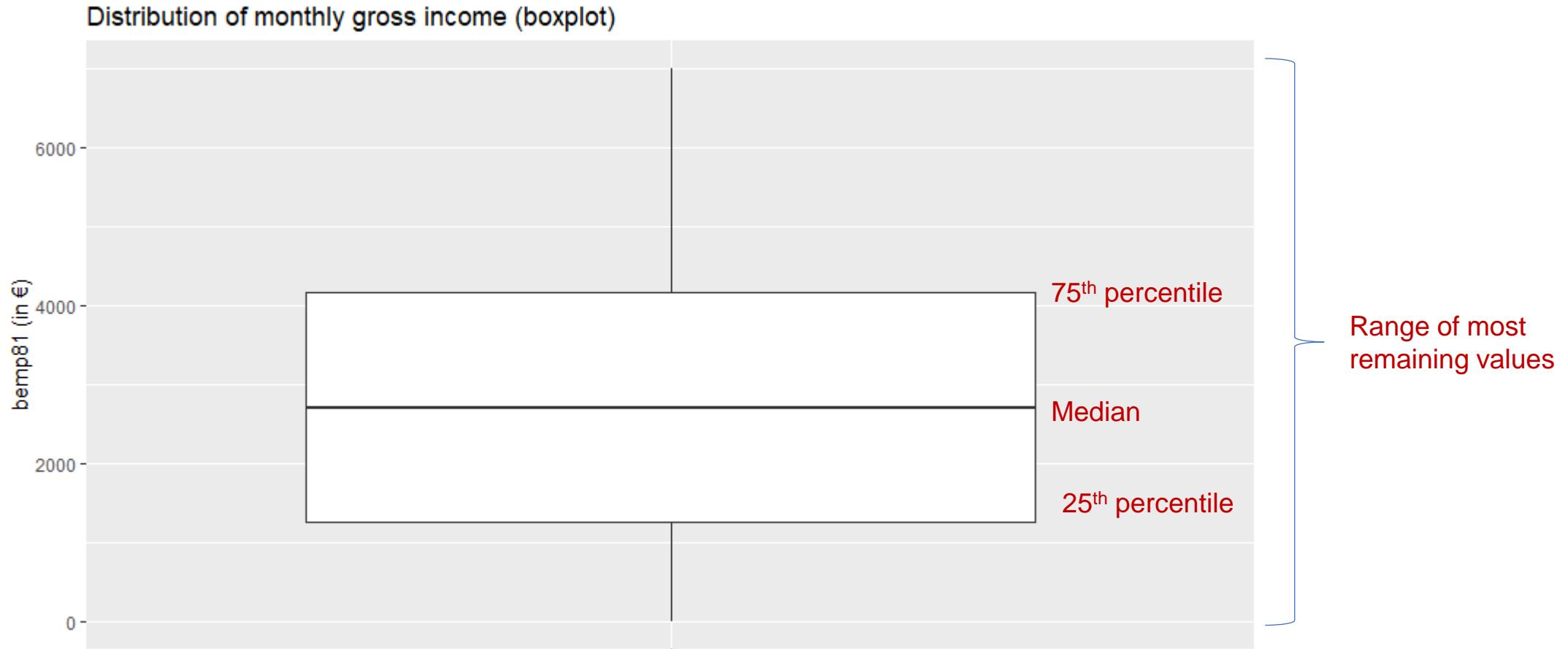
Maximum (max)

→ largest observed value in a variable

} quantils (because they're $\frac{1}{4}$)

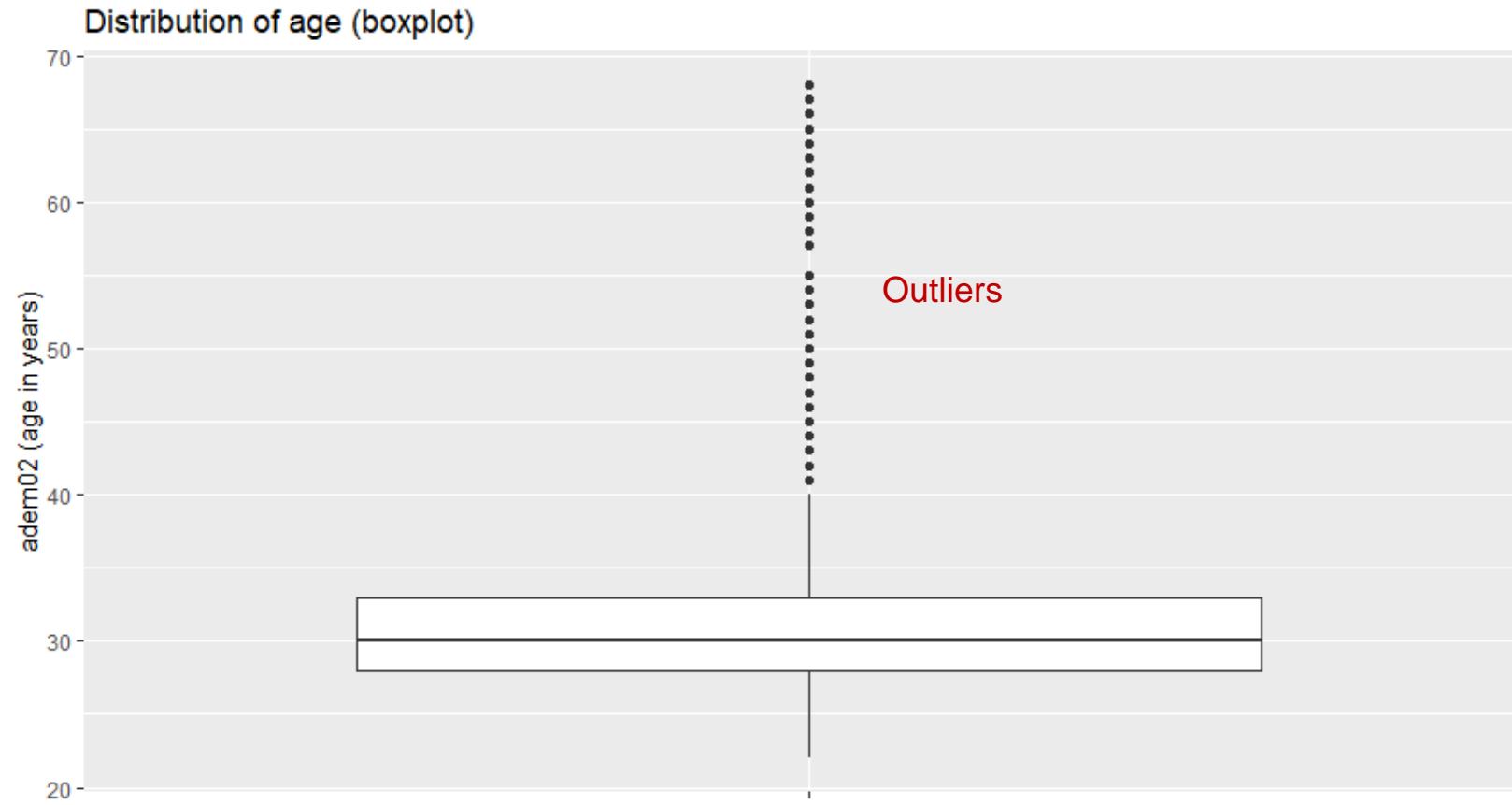
} range of possible values

Visualization of percentiles: Boxplot



Example: Boxplot of variable `bemp81` measuring monthly gross income using `ggplot()`

Visualization of percentiles: Boxplot



Example: Boxplot of variable `adem02` measuring age in years using `ggplot()`

Measure of central tendency: Mean

The Mean (in German: arithmetisches Mittel, Mittelwert) is the [arithmetic average](#) of all values:

$$\bar{x}_{arithm} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$

$$\text{Mean} = \frac{\text{Sum of all values}}{\text{Number of cases}}$$

Example:

| Income (€) | Absolute frequency (n) |
|------------------|------------------------|
| 2000 | 1 |
| 2200 | 1 |
| 2800 | 1 |
| 3000 | 1 |
| 10000 | 1 |
| Total (n) | 5 |

$$\text{Mean} = \frac{2000 + 2200 + 2800 + 3000 + 10000}{5} = 4000$$

→ Careful: The mean is [sensitive to outliers](#).

Measures of central tendency: When to use which measure?

| Measure | Works for | Sensitive to outliers? | What it tells you |
|---------|-----------------------------------|------------------------|------------------------------------|
| Mode | Nominal, ordinal, interval, ratio | ✗ No | The most frequent value |
| Median | Ordinal, interval, ratio | ⚠ Slightly | The middle value (50% below/above) |
| Mean | Interval, ratio | ✓ Yes | The mathematical average |

Variability: Standard Deviation

Standard deviation (s or SD)

Typical deviation from the mean.

It shows, on average, **how far each value lies from the mean**, expressed in the **same unit** as the data (e.g. €)

$$S = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n - 1}}$$

| Symbol | Meaning | Explanation |
|---------------------|---------------------------|---|
| s | standard deviation | The result we want — shows how much the values vary. |
| x_i | each individual value | Every data point in your dataset. |
| \bar{x} | mean (average) | The “center” of all values. |
| $x_i - \bar{x}$ | deviation from the mean | How far each value is from the mean (can be negative or positive). |
| $(x_i - \bar{x})^2$ | squared deviation | We square it so negatives don’t cancel positives. |
| \sum | “sum of” | Add up all the squared deviations. |
| (n - 1) | number of cases minus one | We divide by (n-1) to get an unbiased estimate for samples (→ “sample standard deviation”). |
| $\sqrt{ }$ | square root | Brings the result back to the same unit as the data. |

Standard deviation: Interpretation

Example

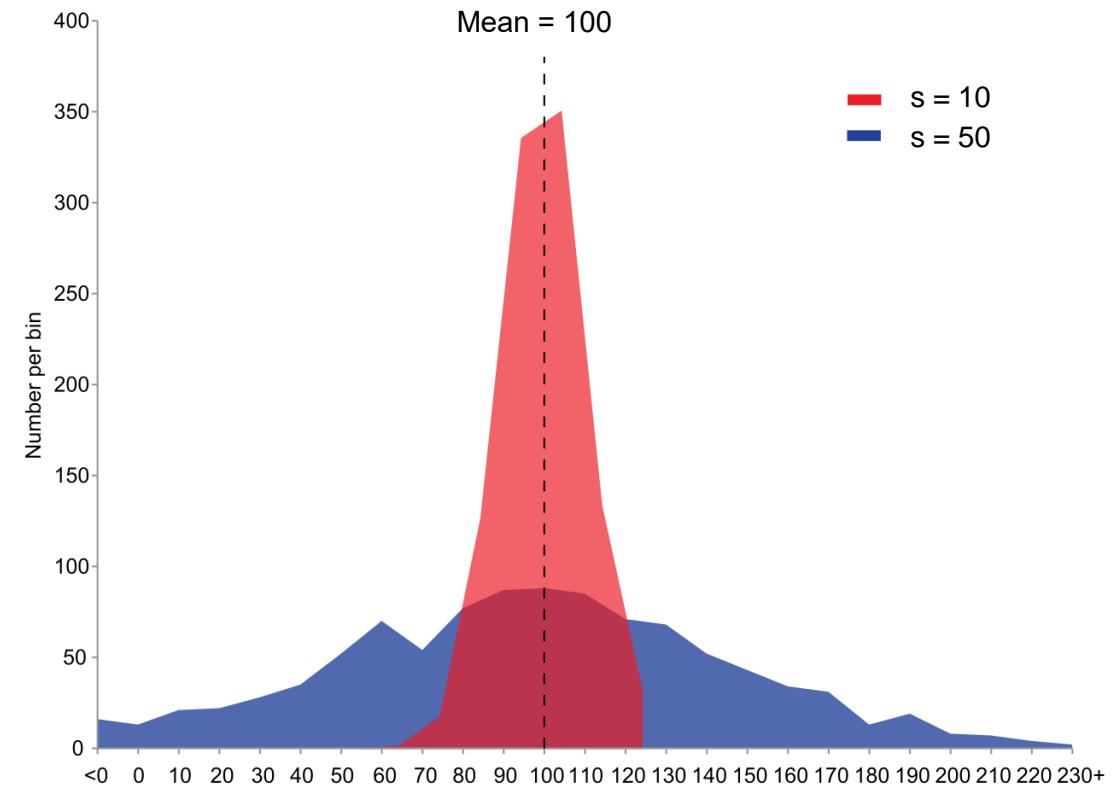
Sample of two population with the same mean but different standard deviations:

- red: small standard deviation
- blue: large standard deviation

→ Easy to interpret: “On average, values deviate from the mean by about s units.”

Small $s \rightarrow$ values are close to the mean
→ data are consistent / homogeneous

Large $s \rightarrow$ values are spread out widely
→ data are variable / heterogeneous



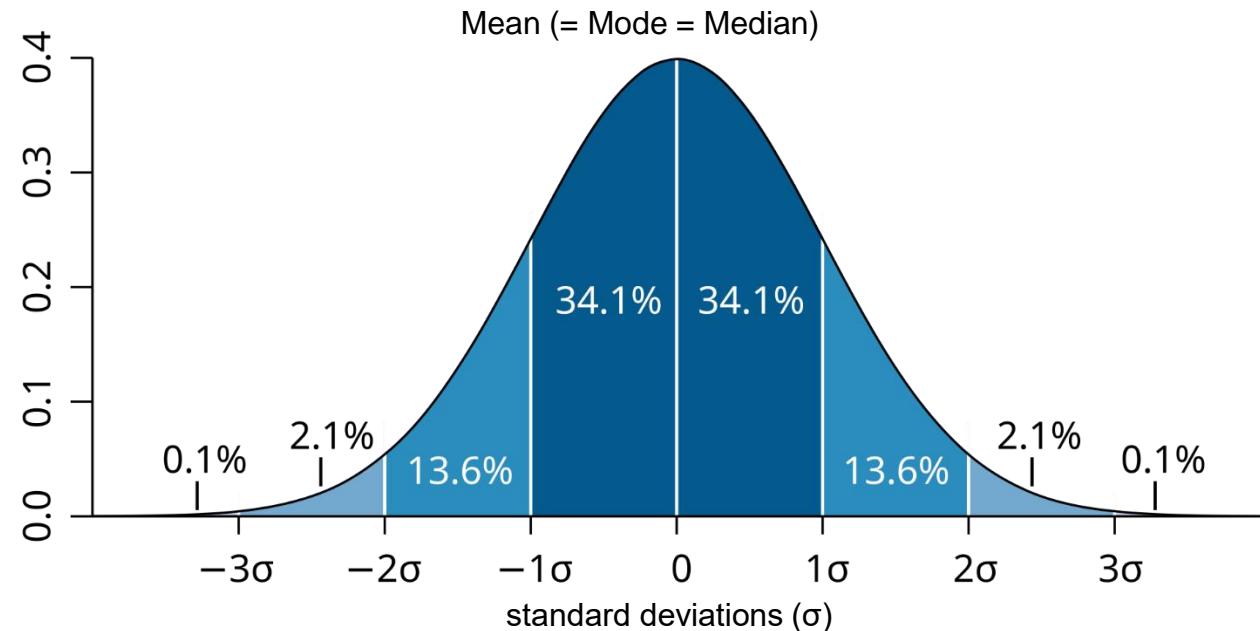
Normal distribution

The **normal distribution curve (bell curve, Normalverteilung)** represents how data are distributed when:

- most values are close to the **mean** and,
- fewer are very high or very low.

The peak in the middle marks the **mean**, which is also the **median** and **mode** in a perfect normal distribution.

- About 68% of all values lie within ± 1 standard deviation of the mean.
- About 95% lie within ± 2 standard deviations.
- About 99,7% cases fall within ± 3 standard deviations.



Why do I need to hear that?

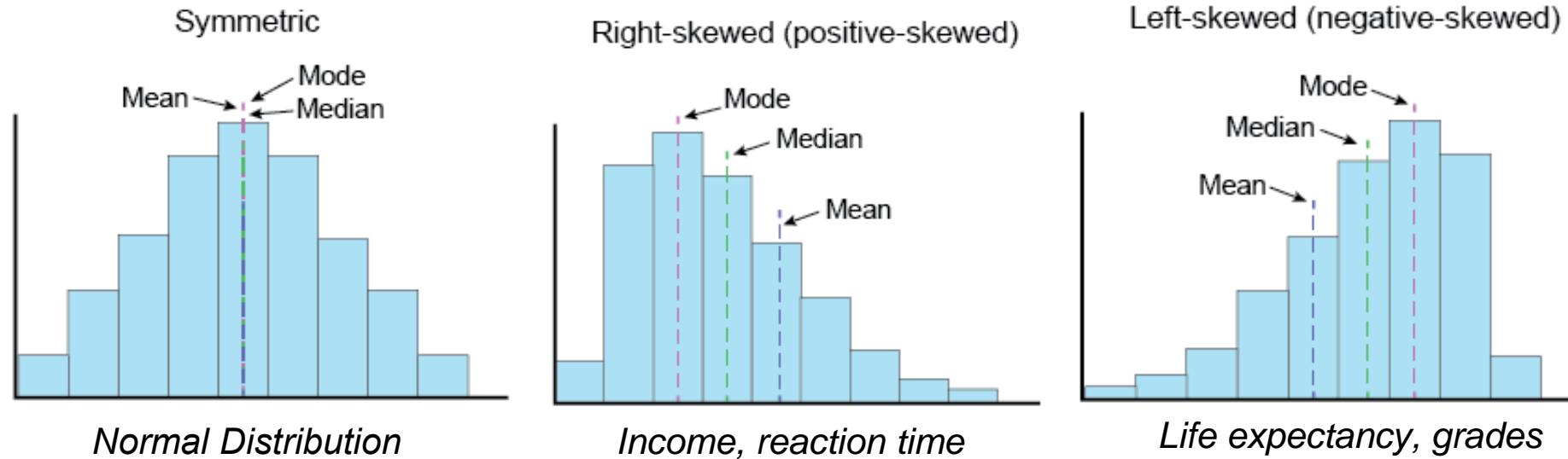
Many statistical inference tests (e.g. t-tests, ANOVA) assume normality



Skewness

Skewness (Schiefe) measures the **asymmetry** of a distribution:

- Symmetric
- Right-skewed
- Left-skewed



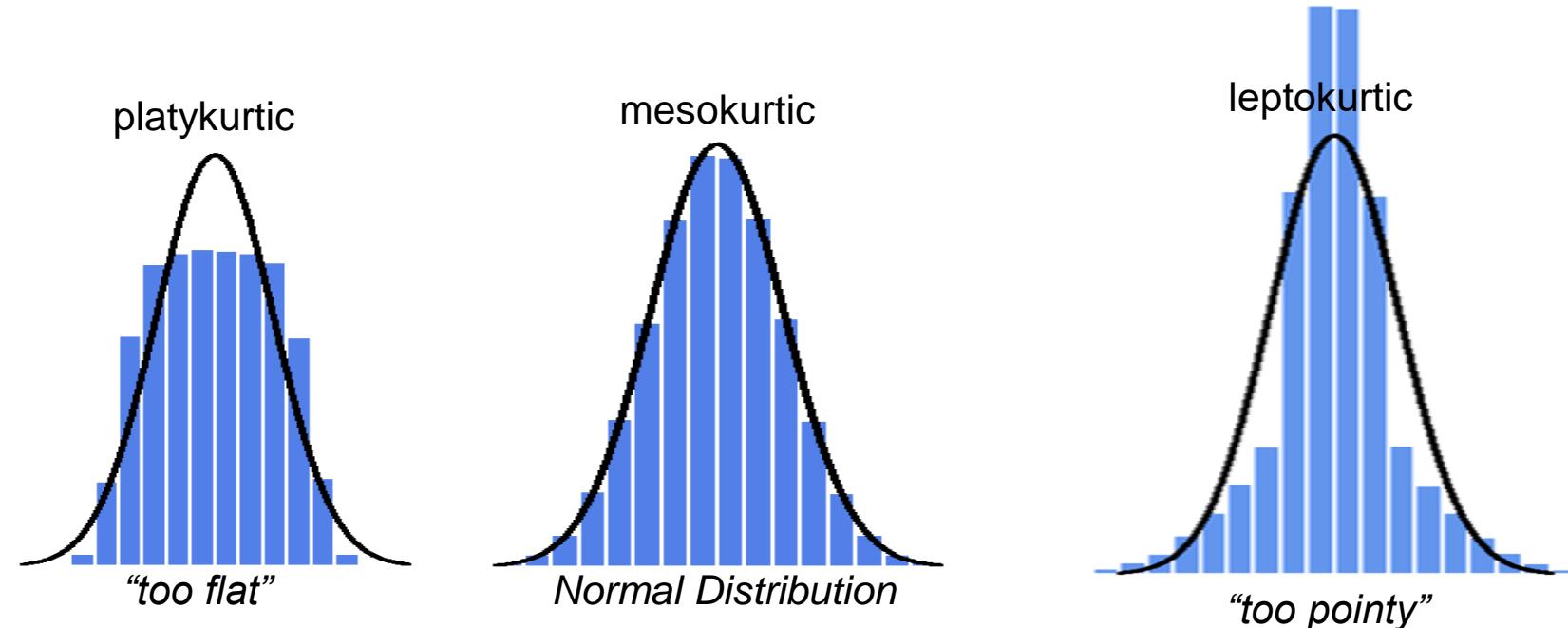
Interpretation

→ Tells you whether the values are more spread out on one side of the mean.

| Skewness value | meaning |
|-----------------------|--|
| 0 | Symmetric, perfectly normal distribution |
| -0.5 – 0.5 | Approximately symmetric |
| -1 – -0.5 and 0.5 – 1 | Moderately skewed |
| -1 – 1 | General acceptable for normal distribution |
| < -1 or > 1 | Highly skewed |

Kurtosis

Kurtosis (Wölbung) describes peakedness or flatness of a distribution compared to a normal distribution.



Interpretation

→ It describes how concentrated or spread out the values are around the mean.

| kurtosis value | meaning |
|----------------|---|
| 3.0 | mesocurtic, perfectly normal distribution |
| < 3.0 | platykurtic |
| > 3.0 | leptokurtic |

A kurtosis value between 2 to 4 (or -1 to +1 for excess kurtosis) is generally acceptable for normal distribution.

Variability: Standard Deviation and Variance

Standard deviation (s)

Typical deviation from the mean.

It shows, on average, how far each value lies from the mean, expressed in the same unit as the data (e.g. €).

→ But: negative deviations cancel each others out.

Example

values 2, 4, 6, with a mean of 4

Deviations from the mean:

- $(2 - 4) = -2$
- $(4 - 4) = 0$
- $(6 - 4) = +2$

If you simply sum them up: $-2 + 0 + 2 = 0$

Variance (s^2)

Average of squared deviations from the mean.

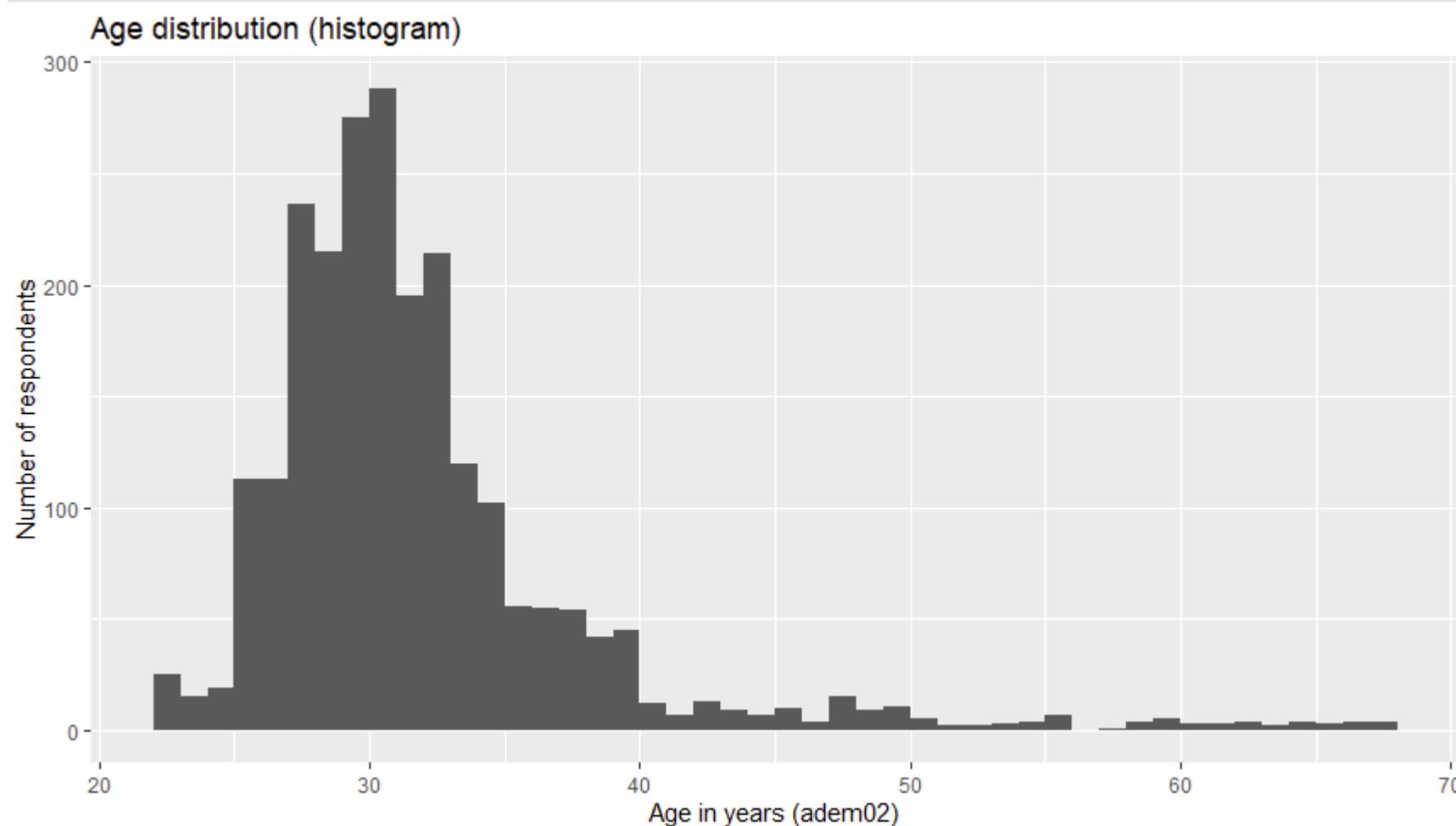
Measures how much the values differ from the mean **on average**, expressed as squared unit (e.g. €²).

$$s^2 = \frac{\sum (X - \bar{X})^2}{n-1}$$

→ Squaring prevents negative deviations from cancelling each other out.

→ Even if it's not easy to interpret directly (because of squared units), it gives us a **precise measure of overall variability** in a dataset.

Visualization: Histogram



Example: Histogram of variable `adem02` measuring age in years using `ggplot()`

Inference Statistics: Standard Error of the Mean ($SE_{\bar{x}}$)

How precise is our sample mean as an estimate of the population mean?

→ Smaller $SE_{\bar{x}}$ → more precise estimate

$$SE_{\bar{x}} = \frac{SD}{\sqrt{n}}$$

- If **SD is large** → SE is large (more variability → less precision).
- If **n is large** → SE is small (more data → more precision).

Standard deviation (SD, s)

Standard deviation describes how much **individual values vary** around the mean.

→ Spread of the data.



Standard Error (SE)

If we **repeated the study many times**, the sample means would vary. The standard error describes this variation.

→ Uncertainty of the estimate.

Inference Statistics: Confidence Interval (CI)

Instead of a single estimate, what is a plausible range of values for the population mean?

→ Narrow CI → more precise estimate

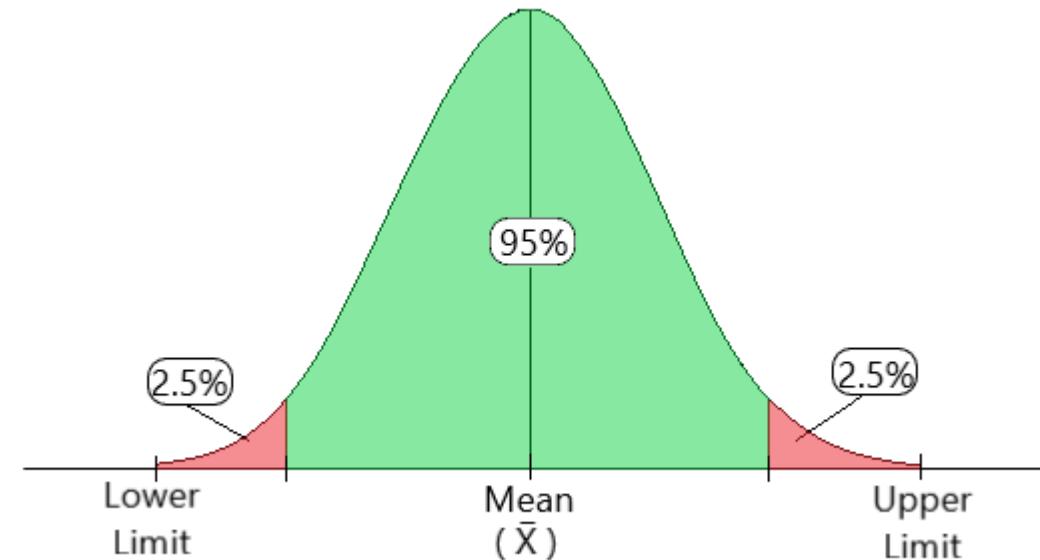
For a 95% CI of the mean:

$$\bar{x} \pm 1.96 * SE_{\bar{x}}$$

(or using a t-value for smaller samples from the t-distribution table)

Larger **SE** → wider CI.

Larger **n** → smaller SE → narrower CI.



There is no absolute rule for when a CI is ‘narrow’ or ‘wide’.

It always has to be judged relative to the measurement scale and the research question.

How to report SD and CI?

█ Report always both, the mean and SD (standard deviation).

- Standard deviation is preferred over variance because it's easier to interpret
- Variance (s^2) is usually only reported when it's directly needed for another calculation or method (e.g., ANOVA).

⊕ Whenever the mean is an important result, report it together with the SD and a 95% confidence interval to show how precise the estimate is.

- Standard error (SE) is mainly a technical quantity used to construct the CI and is rarely reported on its own.

Example: $M = 13.2$ ($SD = 0.8$, 95% CI [13.04, 13.36], $n = 100$)

→ The average wage/hour is 13,30 €. On average, values differ from the mean by 0,80 €.

→ A 95% confidence interval for the population mean ranges from about 13.0 € to 13.4- €.

Your turn!

1. Copy folder Materials from to your local machine

→ <https://github.com/Data-Science-Center-UB/Intro-Quantitative-Analysis-R>

→ Path do subfolder Exercises will be your
working directory

→ Folder content: data sets, exercises without &
with solution

2. Do exercise “**02_qa_uni_ex**”

Version with solution: **02_qa_uni_ex_solution**



Bivariate Statistics

What is it?

Bivariate statistics deal with the analysis of the **relationship between two variables at a time**.

Why bivariate analyses are important:

- **Explore relationships**, how one variable changes with another (e.g., age and income, education and job satisfaction).
- **Compare groups**, e.g. mean differences between gender groups, income groups.
- **Check assumptions** before running multivariate models (e.g., linearity, homoscedasticity, independence).
→ Many inferential methods (e.g., regression, ANOVA, chi-square tests) **start with bivariate exploration**.
- Provide **clear and interpretable** tables and charts for publications.

What decisions to make?

1. ⚙ Which variables to analyze together?

Depends on your **research question**, for example:

- **Does one variable influence another?** e.g., education → income, age → life satisfaction
- **Are there differences between groups?** e.g., gender → mean income
- **Do two variables move together?** e.g., correlation between age and political interest
- **Is there an association between categories?** e.g., education level × employment status

2. ☀ Which type of analysis to use?

Depends mainly on the **level of measurement**

| Variable X + Variable Y | Suitable analysis | Visualization |
|---------------------------|--------------------------------------|---------------------|
| Categorical + categorical | Cross tabulations, Chi-square test | Clustered bar chart |
| Categorical + metric | Mean group comparison, t-test, ANOVA | Grouped box plot |
| Metric + metric | Correlation, linear regression | Scatterplot |

Cross tabulation (Crosstab)

A **cross tabulation** (or contingency table, Kreuztabelle/Kontingenztabelle) summarizes the relationship between two categorical variables by showing **how often each combination of categories occurs**.

- Identifying patterns and associations between two categorical variables e.g., *Is education level related to employment status?*
- Comparing group differences → e.g., *Does voting preference vary by gender?*

| Gender | Left | Greens | SPD | CDU/CSU | Other | Total |
|---------|-----------------------|------------------------|------------------------|------------------------|-----------------------|-------------|
| Women | 70
(10.8% / 50.7%) | 200
(30.8% / 59.7%) | 180
(27.7% / 54.5%) | 150
(23.1% / 35.5%) | 50
(7.7% / 40.0%) | 650
100% |
| Men | 60
(9.2% / 43.5%) | 120
(18.5% / 35.8%) | 140
(21.5% / 42.4%) | 260
(40.0% / 61.6%) | 70
(10.8% / 56.0%) | 650 |
| Diverse | 8
(16.0% / 5.8%) | 15
(30.0% / 4.5%) | 10
(20.0% / 3.0%) | 12
(24.0% / 2.8%) | 5
(10.0% / 4.0%) | 50 |
| Total | = 138
100% | 335 | 330 | 422 | 125 | 1,350 |

Example: Voting behaviour and gender; absolute and relative frequencies absolute count (row % / column %).

Bivariate Analysis: Categorical x Categorical

Crosstab reporting

Ask yourself which variable is the **grouping variable**.

1. If you want to understand **effects or group differences**:

group = row → use row %

2. If you want to understand **composition**:

group = column → use column %

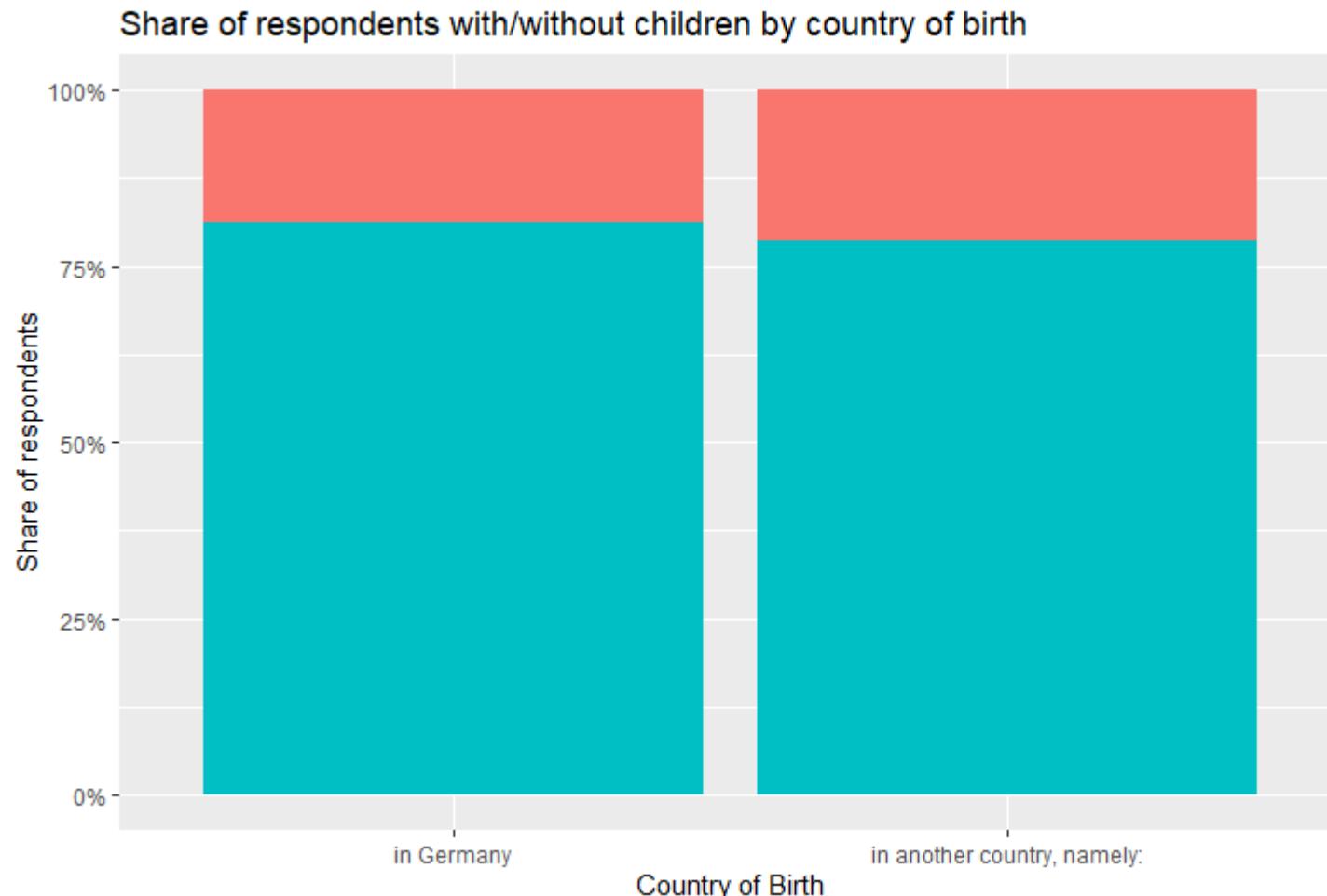
Example: gender differences in voting behaviour

| Gender | Left | Greens | SPD | CDU/CSU | Other | Total |
|---------|-------------|--------------|--------------|--------------|-------------|-------|
| Women | 70
10.8% | 200
30.8% | 180
27.7% | 150
23.1% | 50
7.7% | 650 |
| Men | 60
9.2% | 120
18.5% | 140
21.5% | 260
40.0% | 70
10.8% | 650 |
| Diverse | 8
16.0% | 15
30.0% | 10
20.0% | 12
24.0% | 5
10.0% | 50 |
| Total | 138
100% | 335
100% | 330
100% | 422
100% | 125
100% | 1,350 |

Example: gender composition of voters

| Gender | Left | Greens | SPD | CDU/CSU | Other | Total |
|---------|-------------|--------------|--------------|--------------|-------------|-------------|
| Women | 70
50.7% | 200
59.7% | 180
54.5% | 150
35.5% | 50
40.0% | 650
100% |
| Men | 60
43.5% | 120
35.8% | 140
42.4% | 260
61.6% | 70
56.0% | 650
100% |
| Diverse | 8
5.8% | 15
4.5% | 10
3.0% | 12
2.8% | 5
4.0% | 50
100% |
| Total | 138 | 335 | 330 | 422 | 125 | 1,350 |

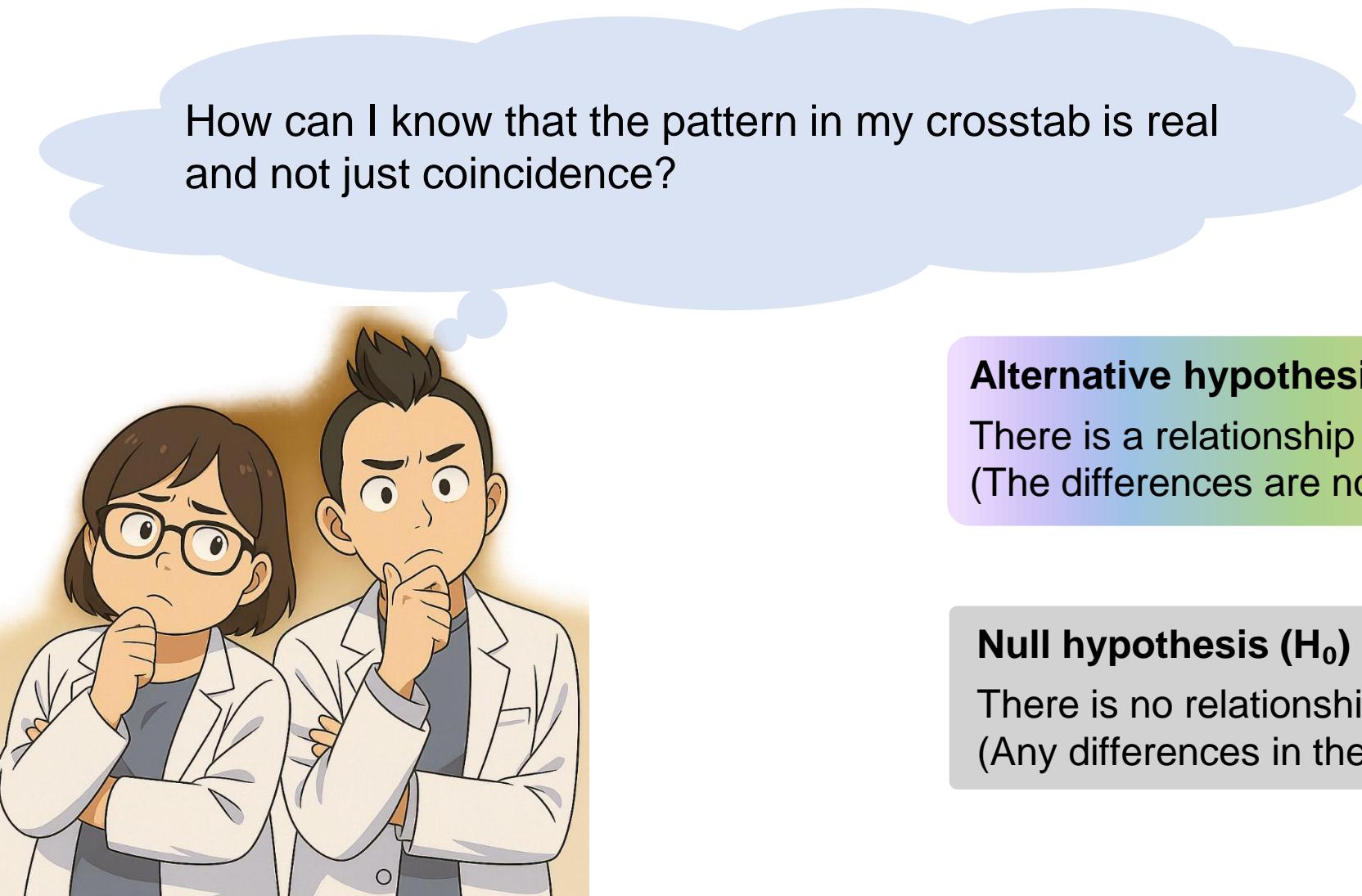
Crosstab Visualization: Clustered bar chart



Children (blcd06)
yes
no

Example: Clustered Bar chart of variable `blcd06` by `adem03` country of birth measuring Status of the doctorate using `ggplot()`

Inference Statistics: Hypothesis testing



How can I know that the pattern in my crosstab is real and not just coincidence?

Alternative hypothesis (H_1)

There is a relationship between the two variables.
(The differences are not random.)

Null hypothesis (H_0)

There is no relationship between the two variables.
(Any differences in the table are just random.)

Chi-square test (χ^2 -test) of independence

The Chi-square test (χ^2 test) tests for **statistical significance** in cross tabulations.

→ The Chi-square test tells us whether we should **reject H_0**

What does the Chi-square test do? (in very simple words)

- The table shows the **observed counts** (the real data).
- The test calculates what the counts **would look like if there were no relationship** between the variables (expected counts).
- Then it compares the two.

→ The **bigger** the difference between observed and expected χ^2 , the more likely it is that the **relationship is real**.

Assumptions

The chi-square test requires:

- Sufficient sample size (at least **20–30** total cases in the table).
- Expected frequency in each cell ≥ 5 (rule of thumb).
→ Otherwise collapse categories.
- Independent observations.
- Categories must be **mutually exclusive**.

Chi-square test: p-value

The p-value tells you how likely it is to see your result just by chance if there is no real effect.

$p < 0.05$ means: There is less than a 5% chance that the pattern in your data is just random.

Small p-value ($< .05$) → very unlikely to be random

→ We reject H_0 , there is a significant relationship.

Large p-value ($> .05$) → easily could be random

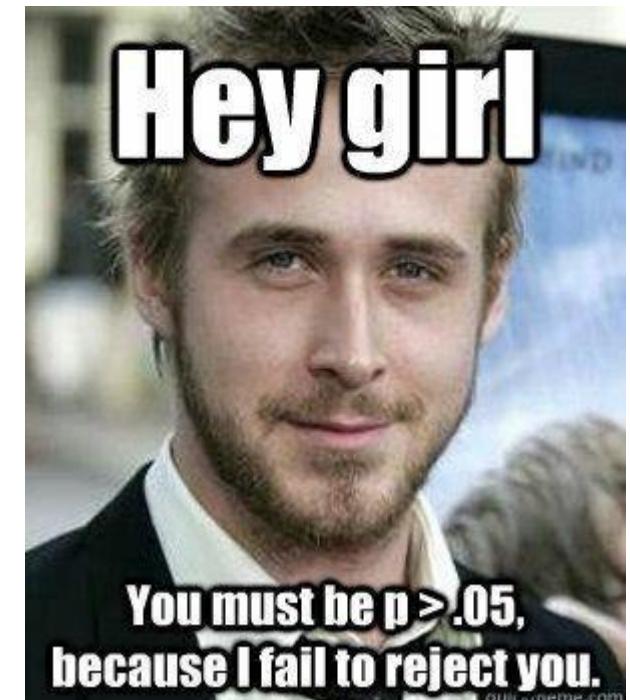
→ No significant relationship

Degrees of freedom (df) show how many independent pieces of information are available for estimating variability.

→ They adjust the test so that p-values are accurate.

Example sentence for reporting:

The association between gender and voting preference was significant, $\chi^2(8) = 21.5$, $p = .006$.



Chi-square test in R

```
> tab_counts <- mydata2 %>% # store results in new data frame tab_counts
+   filter(!is.na(adem03), !is.na(b1cd06)) %>% # include only cases without NA
+   tabyl(adem03, b1cd06) # basic cross tabulation (absolute values)
>
> tab_counts
      adem03 yes   no
      in Germany 348 1503
      in another country, namely: 74  270
> chisq.test(tab_counts)
```

Pearson's chi-squared test with Yates' continuity correction

```
data: tab_counts
X-squared = 1.2039, df = 1, p-value = 0.2726
```

>

Mean group differences

Mean group differences analyze whether the **average value of a metric variable differs across groups of a categorical variable**.

- *Do education groups differ in their voter turnout?*
- *Do gender groups differ in average income in €?*

Example: gender differences in monthly average gross income in €

| Gender | Mean income (€) | SD (€) | n |
|---------|-----------------|--------|-----|
| Women | 2,900 | 650 | 650 |
| Men | 3,200 | 850 | 650 |
| Diverse | 3,050 | 700 | 50 |

Mean group differences reporting

- The direction of comparison depends on your research question and changes how the result is perceived scientifically and socially:

On average, men earn more than women. vs. On average, women earn less than men.

→ If you study group differences, choose the conceptually meaningful reference
e.g., “men vs. women”.

→ If you study inequality, choose the dominant or advantaged group as reference
e.g., “women earn less than men”.

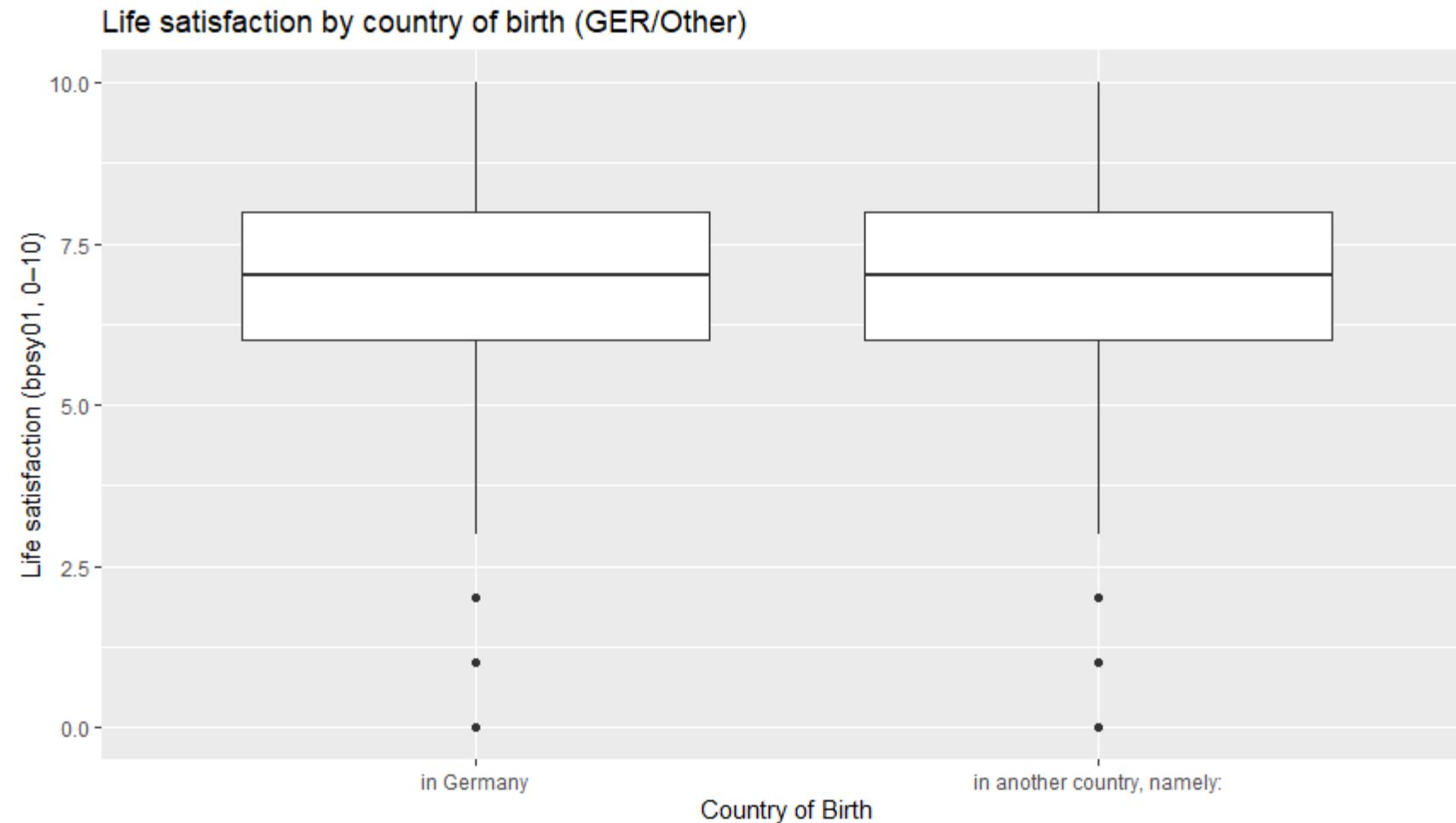
→ If you study policy impact, choose the control group as reference
e.g., “treatment group scores higher than control”.

- State the means (M), standard deviations (SD) and group size (n) for each group.

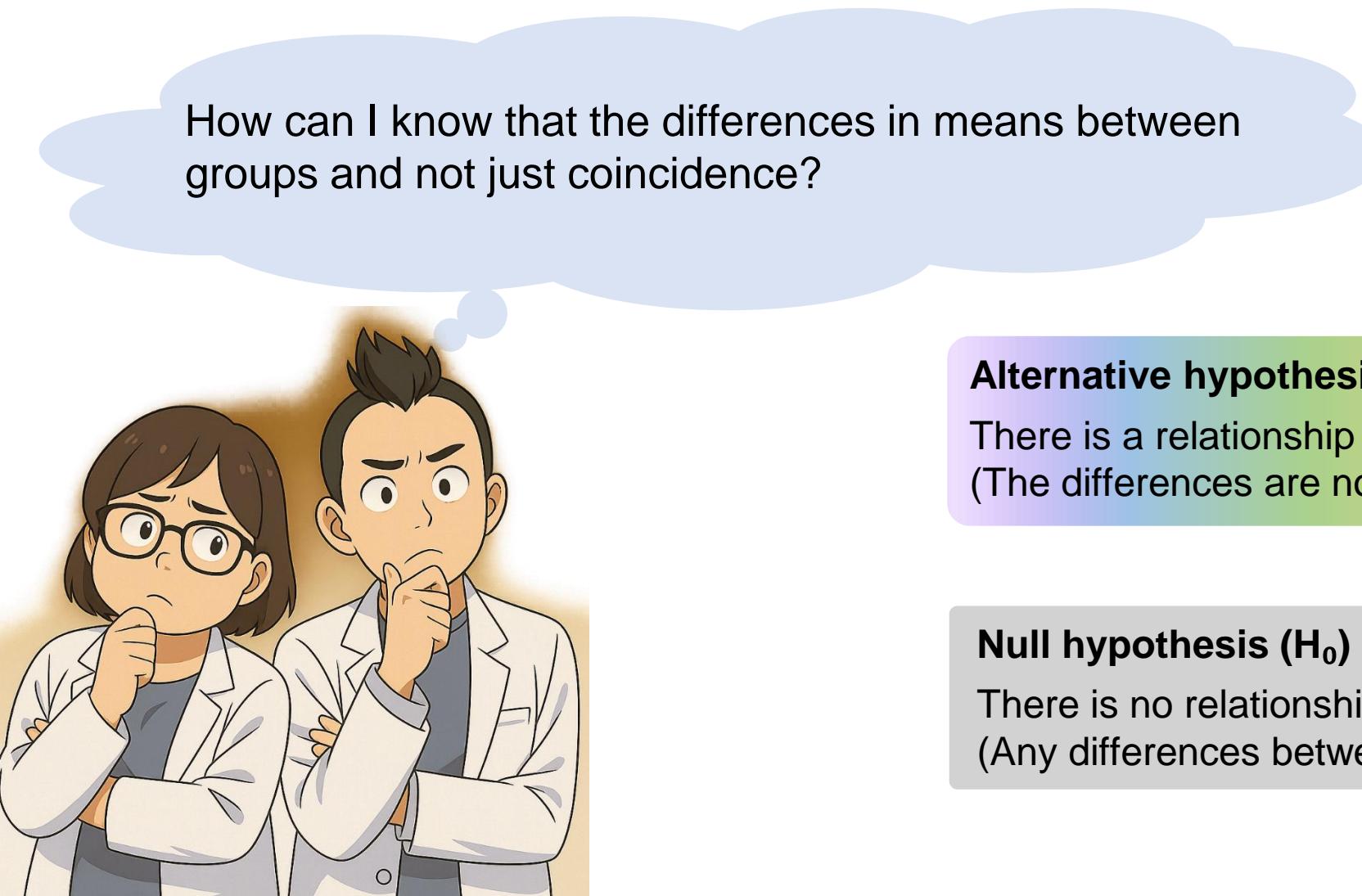
Women ($M = 2,900$, $SD = 650$), Men ($M = 3,200$, $SD = 850$), Diverse ($M = 3,050$, $SD = 700$, $n = 50$)

- If you want to understand how income groups are composed by gender, you need to first create income categories and then run a crosstab (income group × gender).

Mean group differences visualization: Grouped box plots



Inference Statistics: Hypothesis testing



How can I know that the differences in means between groups and not just coincidence?

Alternative hypothesis (H_1)

There is a relationship between the two variables.
(The differences are not random.)

Null hypothesis (H_0)

There is no relationship between the two variables.
(Any differences between groups are just random).

T-test and ANOVA

Testing for statistical significance in mean differences across groups.

T-test

Compares means of **two groups**

Are the two group means different beyond what we would expect by chance?

→ Looks at the **difference between the means** relative to the **variation within the groups**.

Independent t-test

Compares means of **two separate groups** that have **different participants**.

- Employed vs unemployed
- Treatment group vs. control group

Paired t-test

compares means **within the same individuals/paired units**.

- Before vs. after an intervention (same participants)
- Matched pairs (e.g., each patient matched with a control)

(oneway) ANOVA (F-test)

Compares means of three or more groups

Do at least one of these group means differ significantly?

- Male, female, diverse
- Scientific discipline

→ Looks at the **difference between the means** relative to the **variation within the groups**.

→ ANOVA only tells you **that a significant difference exists, not which groups differ**.

T-test and ANOVA: p-value

The p-value tells you how likely it is that the difference we observe between group means is just due to random chance.

→ Test become significant if the **between-group variance is much larger than the within-group noise**.

$p < 0.05$ means: There is less than a 5% chance that the differences in group means in your data is just random.

Small p-value ($< .05$) → very unlikely to be random.

→ We reject H_0 , there is a significant group difference.

Large p-value ($> .05$) → easily could be random.

→ No significant group difference.

Example sentence for reporting:

The mean voting turnout differed significantly between employed and unemployed, $t(128) = 2.45, p = .016$.

t-test: $t(df) = \text{value}, p = \dots$

Example sentence for reporting:

The mean income differed significantly across gender groups, $F(2, 147) = 4.21, p = .017$.

ANOVA: $F(df1, df2) = \text{value}, p = \dots$

T-test in R

```
Welch Two Sample t-test

data: bpsy01 by adem03
t = 3.6655, df = 452.22, p-value = 0.0002761
alternative hypothesis: true difference in means between group in Germany and group in another country, namely: is not equal to 0
95 percent confidence interval:
 0.2035913 0.6742183
sample estimates:
mean in group in Germany mean in group in another country, namely:
 7.107509                      6.668605
```

ANOVA in R

```
> summary(anova_gender)
      Df Sum Sq Mean Sq F value Pr(>F)
adem01     1      0   0.138   0.038  0.845
Residuals 2194 7963  3.629
>
```

T-test and ANOVA: Assumptions

T-test and ANOVA require:

- Approximately **normal distribution** within each group.
- Homoscedasticity: equal variances, that means groups have **similar spread**.
→ If not, use Welch t-test or Welch ANOVA (more robust).
- No strict **minimum sample size**, BUT small samples = unstable results, low power, and sensitivity to non-normality.
→ T-test and ANOVA work best with:
 - At least 20-30 cases per group (independent) or pairs (paired);
 - Similar group size.→ To detect small differences, you often need larger samples.

Pearson's correlation (r)

Correlation analyses examine whether two metric variables move together and if changes in one variable are systematically associated with changes in the other.

- Do older people report higher political interest?
- Do hours studies predict exam scores?

Example: Relationship between age in years and political interest score (0–10)

| Statistic | Value |
|---------------------|-------|
| Correlation (r) | 0.32 |
| p-value | 0.004 |
| n | 1,100 |

Pearson's correlation (r): Visualization

Scatterplot

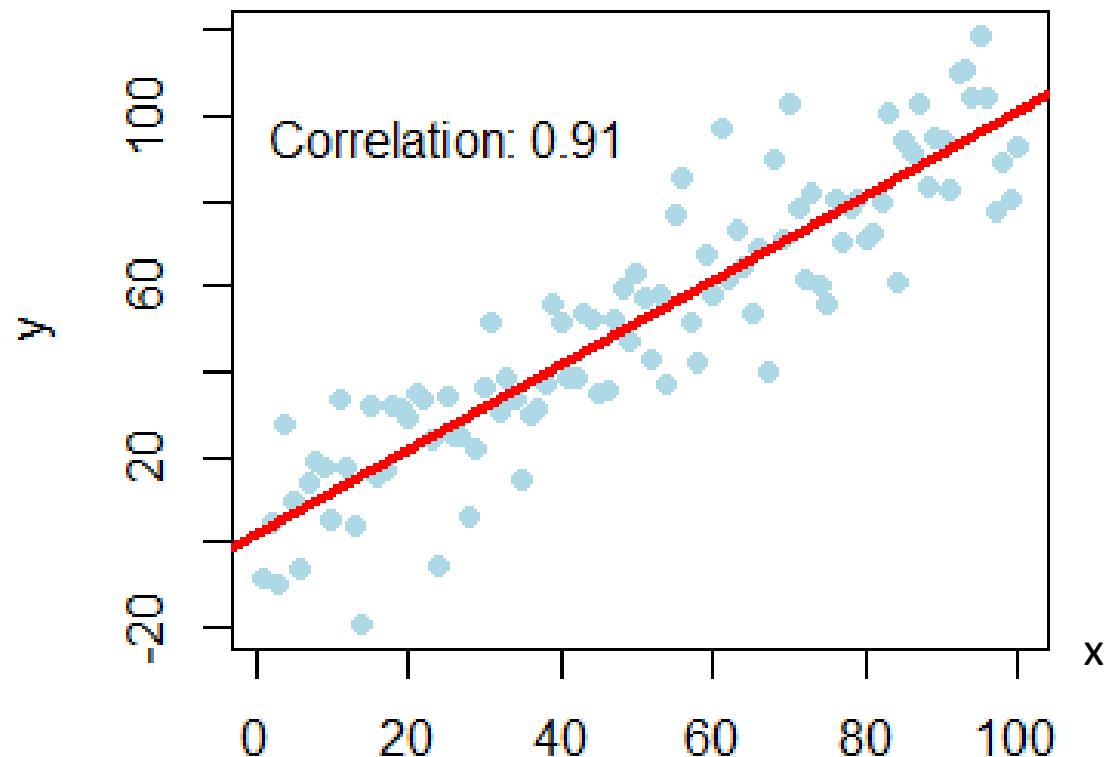
Each point represents one person/case;

X-axis = variable 1;

Y-axis = variable 2;

Shows the pattern: positive, negative, or zero;

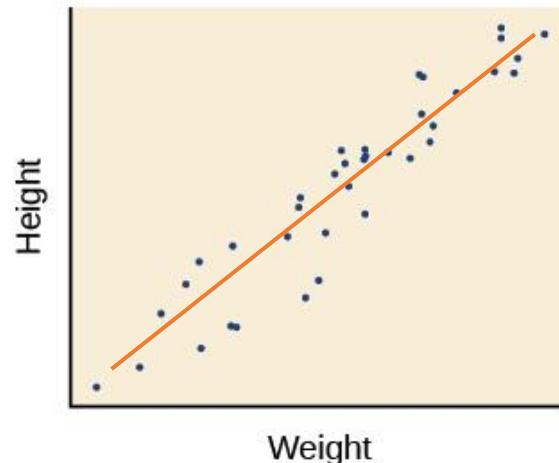
Add a regression line to show the trend.



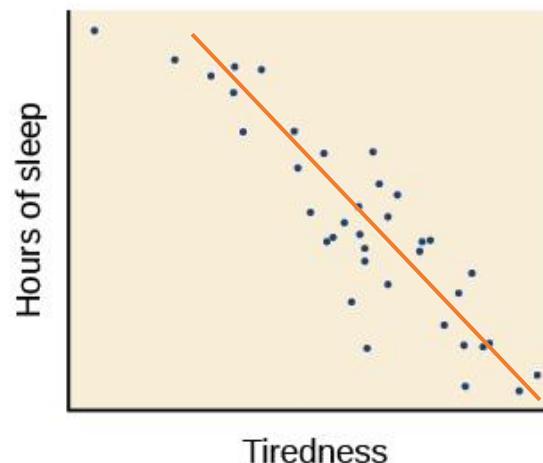
<https://r-coder.com/correlation-plot-r/>

Pearson correlation (r) interpretation

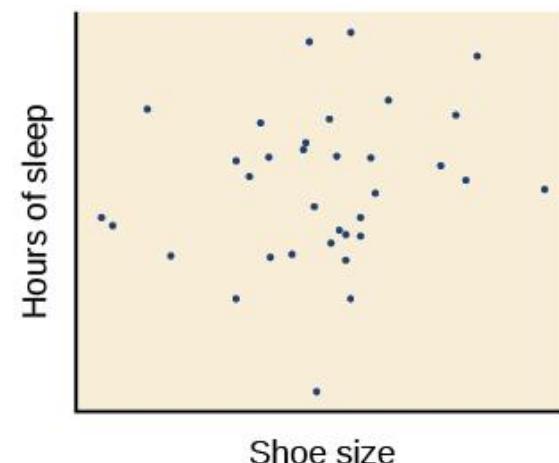
r tells you whether people who have more of X also tend to have more (or less) of Y.



(a) Positive Correlation



(b) Negative Correlation



(c) No Correlation

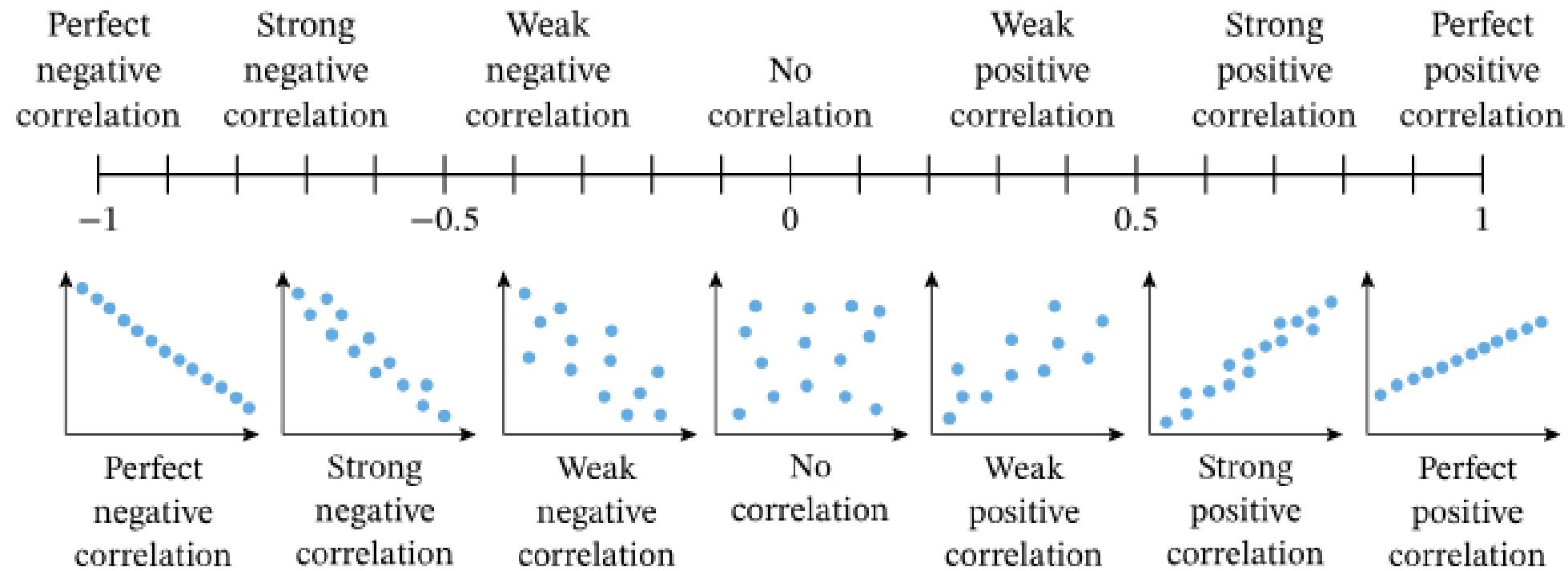
<https://opened.cuny.edu/courseware/lesson/16/student/?section=2>

The correlation coefficient r ranges from -1 to $+1$

- $r > 0$: when X increases, Y tends to increase (positive relationship).
- $r < 0$: when X increases, Y tends to decrease (negative relationship).
- $r \approx 0$: no systematic relationship.

Pearson correlation (r) interpretation

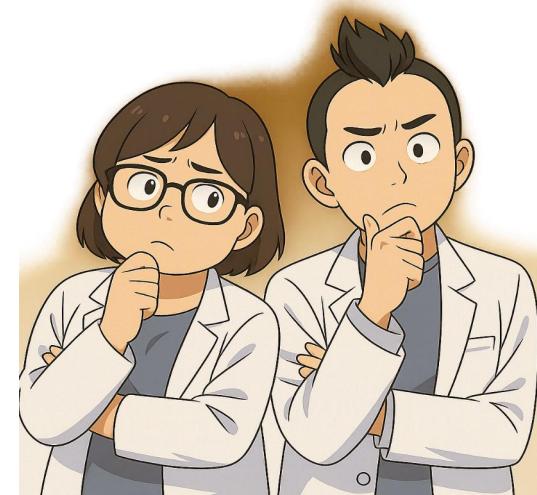
r 's size gives you information about the strength of the relationship:



<https://www.nagwa.com/en/explainers/143190760373/>

Pearson correlation: p-value

How can I know that the relationship between the two variables are not just coincidence?



The **p-value** tells you how likely it is to observe a correlation as big as the one you found, if in **reality there is no relationship between X and Y**.

p < 0.05 means: There is **less than a 5% chance** that the correlation is just random.

Small p-value (**< .05**) → We reject H_0 , there is a **significant correlation**.

Large p-value (**> .05**) → **No significant correlation**.

Example sentence for reporting:

There was a significant positive correlation between study hours and exam scores, $r(248) = .45$, $p < .001$, indicating that students who studied more tended to score higher.

Pearson's r: $r(df) = \text{value}$, $p = \dots$

Pearson correlation (r): Assumptions

Pearson correlation require:

- Both variables are **metric**;
→ Use Spearman correlation for ordinal variables.
- **Linear** relationship;
→ The relationship should look roughly like a straight line in a scatterplot.
- **No strong outliers**;
- Approximately **normal distribution**;
- **Homoscedasticity**: the spread of Y should be fairly similar across values of X;
→ Equal variance" of points along the line.
- **Independent observations**;
→ Each case/person must represent an independent data point.

Your turn!

1. Copy folder Materials from to your local machine

→ <https://github.com/Data-Science-Center-UB/Intro-Quantitative-Analysis-R>

→ Path do subfolder Exercises will be your
working directory

→ Folder content: data sets, exercises without &
with solution

2. Do exercise “**03_qa_bi_ex**”

Version with solution: **03_qa_bi_ex_solution**



Multivariate Analyses

Correlation coefficient vs. Simple linear regression

Correlation measures tell you the **strength and direction of an association**.

- It does not assume one variable predicts the other;
- Correlation **does not mean causation**.

Regression models how one variable (X) **predicts another (Y)**.

- It gives you a mathematical equation ($Y = a + bX$).
- It tells you the size of the effect (slope);
- It includes significance tests and uncertainty (SE, p-value, CI).

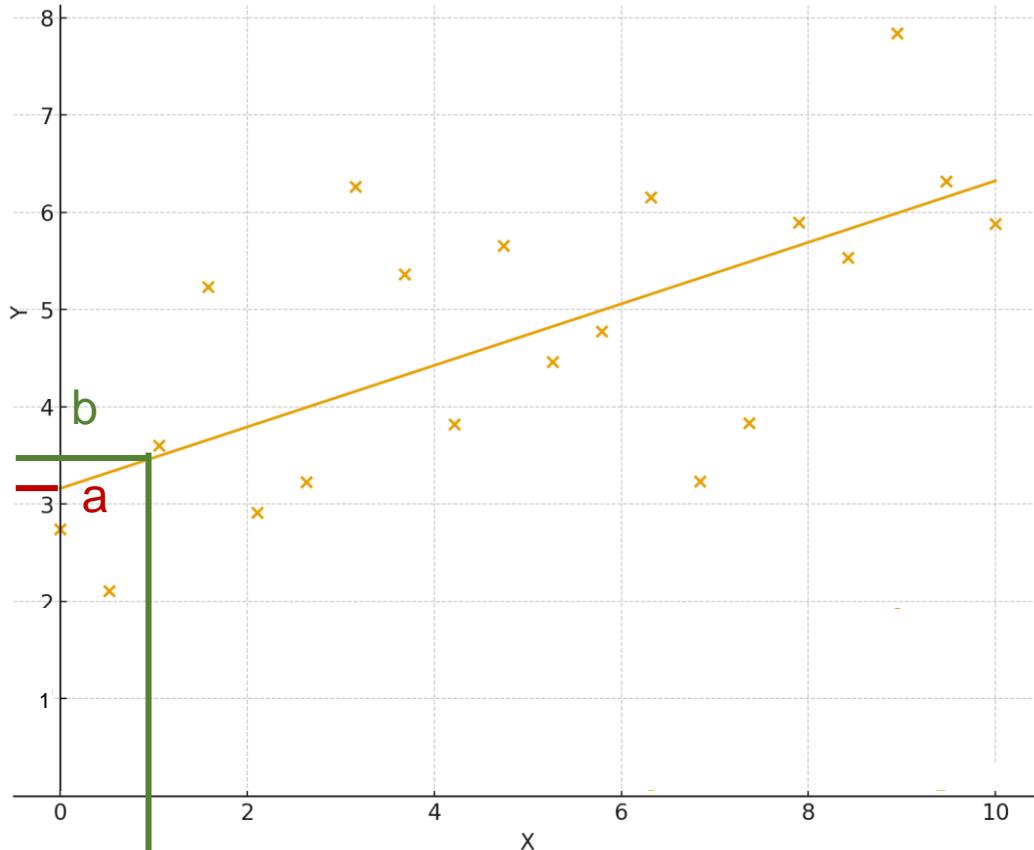
Linear regression models the relationship between two variables with a straight line.

Simple linear regression is the specific case of linear regression with exactly **one predictor variable**.

How does a simple linear regression look like?

Example: We model how **study time** predicts **exam performance**

$$(Y = a + bX).$$



Y = Dependent variable = exam score.

→ Variable you want to explain.

X = Predictor = hours studied.

→ Variable you assume has an effect on Y.

a = Intercept = exam score if hours studied = 0.

b = Slope (Steigung) = Increase in exam score per hour studied.

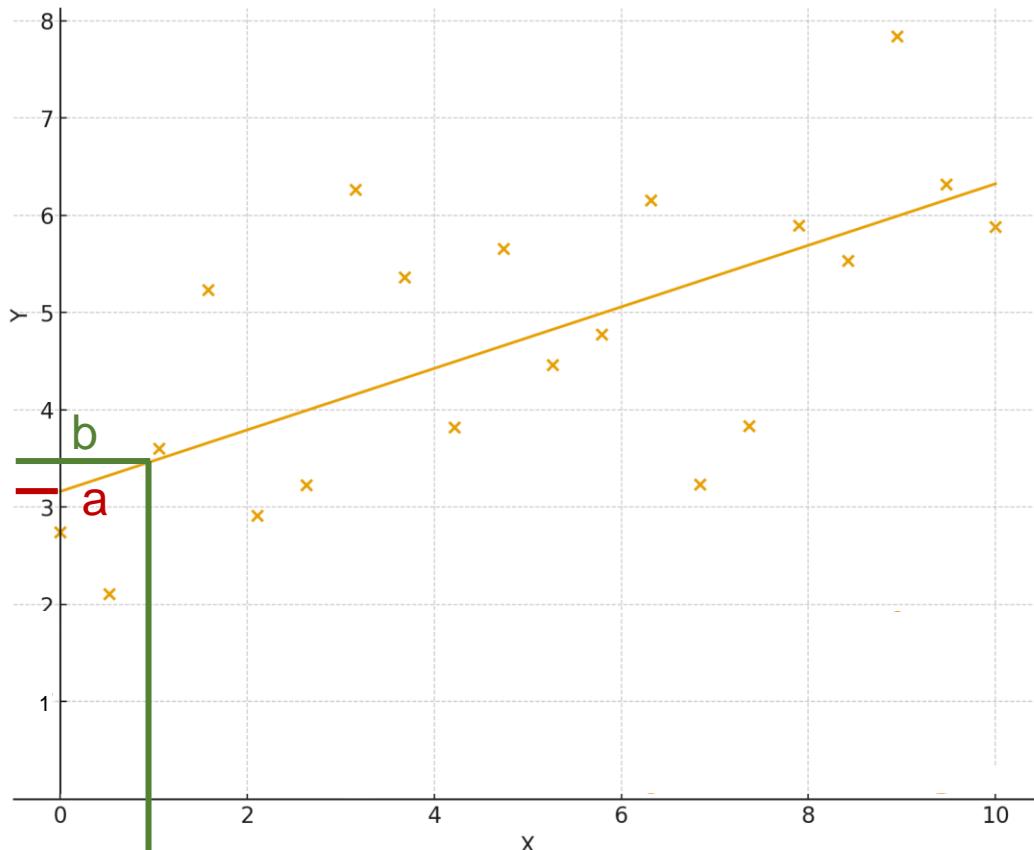
For each 1 unit increase in X, Y increases/decreases on average by b units.

- Positive b = increase, **positive effect on Y**.
- Negative b = decrease, **negative effect on Y**.

Simple Linear Regression

How does a simple linear regression look like?

Example: We model how **study time** predicts **exam performance**



X = Hours of study per week, Y = Exam score

| Term | Estimate | Std. Error | Beta | t | p-value |
|----------------|----------|------------|-------|-------|---------|
| Intercept | 3.162 | 0.503 | — | 6.280 | 0.000 |
| h studied | 0.316 | 0.086 | 0.655 | 3.674 | 0.002 |
| R ² | 0.429 | — | — | — | — |
| n | 20 | | | | |

$$(Y = 3.162 + 0.316X).$$

Without studying, the predicted exam score is 3.16.

For each 1h increase in hours studied, the exam score increases by 0.32.

What do the other outcomes tell me?

Example: We model how **study time** predicts **exam performance**

| Term | Estimate | Std. Error | Beta | t | p-value |
|----------------|----------|------------|-------|-------|---------|
| Intercept | 3.162 | 0.503 | — | 6.280 | 0.000 |
| h studied | 0.316 | 0.086 | 0.655 | 3.674 | 0.002 |
| R ² | 0.429 | — | — | — | — |
| N | 20 | | | | |



Is this effect true or random?

P-value is the probability of getting a result as extreme as (or more extreme than) the observed one, if the null hypothesis were true (e.g. true effect = 0).
→ Small p-value ($< .05$) → reject H_0 , significant effect.

- The **standard error** reflects the uncertainty of b).
→ Smaller SE: more precise estimate.
- **Beta** is the regression coefficient in standard deviation units.
→ It shows how many SDs Y changes when X increases by one SD.
- Useful to compare the strength of predictors measured on different scales.
- The **t value** is the estimate divided by its standard error.
→ Larger t: stronger evidence that true effect is not 0 (H_0 is rejected).

What do the other outcomes tell me?

Example: We model how **study time** predicts **exam performance**

| Term | Estimate | Std. Error | Beta | t | p-value | 95% CI |
|----------------|----------|------------|-------|-------|---------|--------------|
| Intercept | 3.162 | 0.503 | — | 6.280 | 0.000 | [2.10; 4.22] |
| h studied | 0.316 | 0.086 | 0.655 | 3.674 | 0.002 | [0.14; 0.50] |
| R ² | 0.429 | — | — | — | — | |
| N | 20 | | | | | |

From the estimate and its SE, we can build a **95% confidence interval**.

→ If **0 is not inside the 95% CI**, the effect is statistically **significant** at $\alpha = .05$.

What do the other outcomes tell me?

Example: We model how **study time** predicts **exam performance**

| Term | Estimate | Std. Error | Beta | t | p-value |
|----------------|----------|------------|-------|-------|---------|
| Intercept | 3.162 | 0.503 | — | 6.280 | 0.000 |
| h studied | 0.316 | 0.086 | 0.655 | 3.674 | 0.002 |
| R ² | 0.429 | — | — | — | — |
| N | 20 | | | | |

R² shows **how much of the variance in Y is explained by the model.**

→ It is the proportion of variance in the outcome that can be predicted from the predictor(s).

R² ranges from **0 to 1**:

→ 0 = the model explains **none of the variance**.

→ 1 = the model explains **all of the variance**.

Example:

“The model explains 43% of the variance in exam scores.”

Simple Linear Regression

Storks Deliver Babies

(Matthews, 2000)

| Country | Area (km ²) | Storks (pairs) | Humans (10 ⁶) | Birth rate (10 ³ /yr) |
|-------------|-------------------------|----------------|---------------------------|----------------------------------|
| Albania | 28,750 | 100 | 3.2 | 83 |
| Austria | 83,860 | 300 | 7.6 | 87 |
| Belgium | 30,520 | 1 | 9.9 | 118 |
| Bulgaria | 111,000 | 5000 | 9.0 | 117 |
| Denmark | 43,100 | 9 | 5.1 | 59 |
| France | 544,000 | 140 | 56 | 774 |
| Germany | 357,000 | 3300 | 78 | 901 |
| Greece | 132,000 | 2500 | 10 | 106 |
| Holland | 41,900 | 4 | 15 | 188 |
| Hungary | 93,000 | 5000 | 11 | 124 |
| Italy | 301,280 | 5 | 57 | 551 |
| Poland | 312,680 | 30,000 | 38 | 610 |
| Portugal | 92,390 | 1500 | 10 | 120 |
| Romania | 237,500 | 5000 | 23 | 367 |
| Spain | 504,750 | 8000 | 39 | 439 |
| Switzerland | 41,290 | 150 | 6.7 | 82 |
| Turkey | 779,450 | 25,000 | 56 | 1576 |

Table 1. Geographic, human and stork data for 17 European countries

Surprising correlation:
Regions with more storks also have more newborn babies ($r=0.62$).

Simple linear regression:

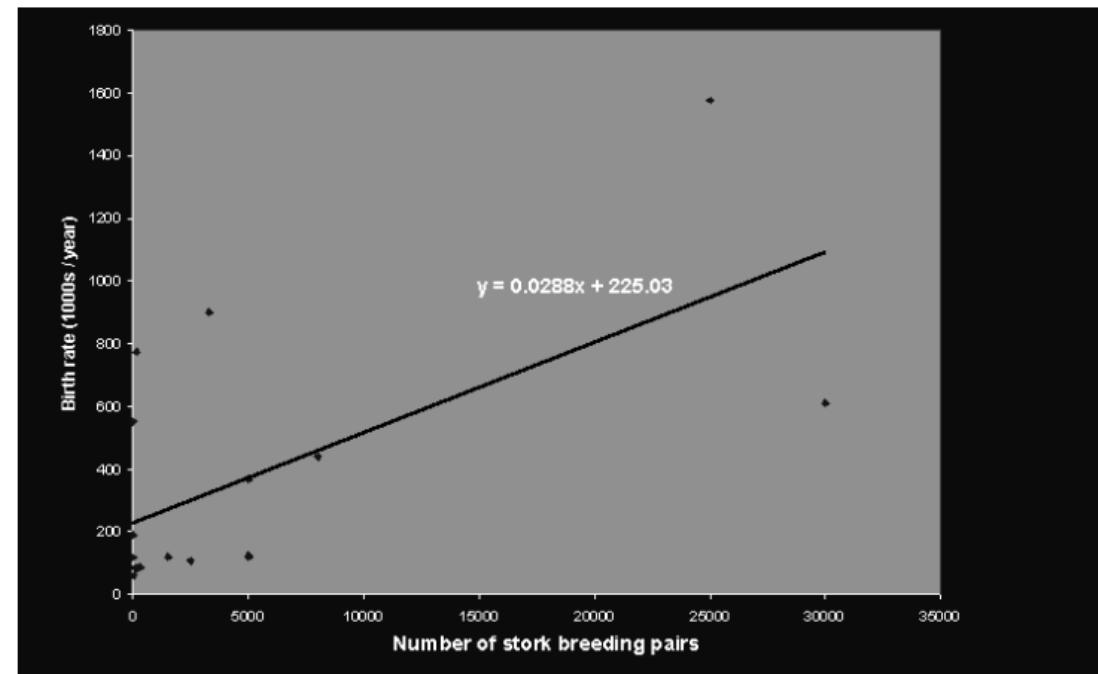


Fig 1. How the number of human births varies with stork populations in 17 European countries.

More storks, more babies.

Simple Linear Regression

Storks Deliver Babies

(Matthews, 2000)



Because both stork numbers and birth rates are higher in [rural regions](#) than in [urban regions](#).

Once you control for urbanicity, the relationship between storks and babies disappears.

→ It's *not* the storks. It's the type of region.

What is it?

Multivariate statistics deal with the analysis of relationships among three or more variables at the same time.

Why multivariate analyses are important

- **Understand complex relationships**, e.g. how an outcome (income) depends on several predictors at once (education, work experience, gender).
- **Control for confounders** and estimate the *unique* effect of one variable while holding others constant.
- **Improve prediction**, because combining several variables usually predicts outcomes better than a single predictor.

What decisions to make?

1. ⚙ Which variables to include in the model?

Depends on your **research question**, for example:

- What is your **outcome variable**? (e.g., income, satisfaction, voting intention)
- Which variables are your **main predictors**?
- Which variables should you **control for**? (e.g., age, gender, education)
- Are there **potential confounders** that might distort the relationship?
- Do you expect **interaction effects**? (e.g., education × gender)

2. ☀ Which type of analysis to use?

Depends mainly on the **level of measurement of the outcome variable**

| Outcome Variable Y | Suitable analysis |
|--------------------|----------------------------|
| Metric | Multiple linear regression |
| Categorial binary | Logistic regression |

How does a multiple linear regression look like?

| Term | Estimate | Std. Error | Beta | t | p-value |
|------------------------------|----------|------------|--------|--------|---------|
| Intercept | 1512.438 | 552.317 | — | 2.738 | 0.007 |
| training (ref. no) | 187.532 | 228.917 | 0.058 | 0.819 | 0.414 |
| Education: medium (ref. low) | 214.759 | 297.531 | 0.067 | 0.722 | 0.472 |
| Education: high (ref. low) | 823.915 | 321.774 | 0.243 | 2.561 | 0.012 |
| Gender: female (ref. male) | -573.284 | 219.847 | -0.232 | -2.608 | 0.010 |
| Gender: diverse (ref. male) | -917.641 | 285.392 | -0.291 | -3.214 | 0.002 |
| Age in years | 41.763 | 12.311 | 0.272 | 3.393 | 0.001 |
| R ² | 0.253 | — | — | — | — |
| Adjusted-R ² | 0.223 | — | — | — | — |
| n | 150 | — | — | — | — |

Example: Linear regression on monthly gross income in €

What does the table tell us?

| Term | Estimate | Std. Error | Beta | t | p-value |
|-----------------------------|----------|------------|--------|--------|---------|
| Intercept | 1512.438 | 552.317 | — | 2.738 | 0.007 |
| Intercept (1512.44) | | | | | |
| Gender: female (ref. male) | -573.284 | 219.847 | -0.232 | -2.608 | 0.010 |
| Gender: diverse (ref. male) | -917.641 | 285.392 | -0.291 | -3.214 | 0.002 |
| Age in years | 41.763 | 12.311 | 0.272 | 3.393 | 0.001 |
| R ² | 0.253 | — | — | — | — |
| Adjusted-R ² | 0.223 | — | — | — | — |
| n | 150 | — | — | — | — |

Example: Linear regression on monthly gross income in €

What does the table tell us?

| Term | Estimate | Std. Error | Beta | t | p-value |
|-----------------------------|----------|------------|--------|--------|---------|
| Intercept | 1512.438 | 552.317 | — | 2.738 | 0.007 |
| training (ref. no) | 187.532 | 228.917 | 0.058 | 0.819 | 0.414 |
| Gender: diverse (ref. male) | -917.641 | 285.392 | -0.291 | -3.214 | 0.002 |
| Age in years | 41.763 | 12.311 | 0.272 | 3.393 | 0.001 |
| R ² | 0.253 | — | — | — | — |
| Adjusted-R ² | 0.223 | — | — | — | — |
| n | 150 | — | — | — | — |

Example: Linear regression on monthly gross income in €

What does the table tell us?

| Term | Estimate | Std. Error | Beta | t | p-value |
|------------------------------|----------|------------|-------|-------|---------|
| Intercept | 1512.438 | 552.317 | — | 2.738 | 0.007 |
| training (ref. no) | 187.532 | 228.917 | 0.058 | 0.819 | 0.414 |
| Education: medium (ref. low) | 214.759 | 297.531 | 0.067 | 0.722 | 0.472 |
| Education: high (ref. low) | 823.915 | 321.774 | 0.243 | 2.561 | 0.012 |

Education: medium (214.76, n.s.), high (823.92) vs. low

Persons with medium education earn on average about 215 € more than those with low education, but this difference is not statistically significant when controlling for the other variables.

Persons with high education earn on average about 824 € more than those with low education, controlling for training, age and gender.

n

150

—

—

—

—

Example: Linear regression on monthly gross income in €

What does the table tell us?

| Term | Estimate | Std. Error | Beta | t | p-value |
|------------------------------|----------|------------|--------|--------|---------|
| Intercept | 1512.438 | 552.317 | — | 2.738 | 0.007 |
| training (ref. no) | 187.532 | 228.917 | 0.058 | 0.819 | 0.414 |
| Education: medium (ref. low) | 214.759 | 297.531 | 0.067 | 0.722 | 0.472 |
| Education: high (ref. low) | 823.915 | 321.774 | 0.243 | 2.561 | 0.012 |
| Gender: female (ref. male) | -573.284 | 219.847 | -0.232 | -2.608 | 0.010 |
| Gender: diverse (ref. male) | -917.641 | 285.392 | -0.291 | -3.214 | 0.002 |

Gender: female (**-573.28**), diverse (**-917.64**) vs. male

- Women earn on average about 573 € less than men, controlling for training, age and education.
- Persons in the “diverse” category earn on average about 918 € less than men, holding training, age and education constant.

Example: Linear regression on monthly gross income in €

What does the table tell us?

| Term | Estimate | Std. Error | Beta | t | p-value |
|--|----------|------------|-------|-------|---------|
| Intercept | 1512.438 | 552.317 | — | 2.738 | 0.007 |
| training (ref. no) | 187.532 | 228.917 | 0.058 | 0.819 | 0.414 |
| Education: medium (ref. low) | 214.759 | 297.531 | 0.067 | 0.722 | 0.472 |
| Education: high (ref. low) | 823.915 | 321.774 | 0.243 | 2.561 | 0.012 |
| Age (41.76) | | | | | |
| <i>Holding training, education and gender constant, each additional year of age is associated with about 42 € higher monthly income.</i> | | | | | |
| Age in years | 41.763 | 12.311 | 0.272 | 3.393 | 0.001 |
| R ² | 0.253 | — | — | — | — |
| Adjusted-R ² | 0.223 | — | — | — | — |
| n | 150 | — | — | — | — |

Example: Linear regression on monthly gross income in €

What does the table tell us?

| Term | Estimate | Std. Error | Beta | t | p-value |
|------------------------------|----------|------------|-------|-------|---------|
| Intercept | 1512.438 | 552.317 | — | 2.738 | 0.007 |
| training (ref. no) | 187.532 | 228.917 | 0.058 | 0.819 | 0.414 |
| Education: medium (ref. low) | 214.759 | 297.531 | 0.067 | 0.722 | 0.472 |
| Age in years | 41.763 | 12.311 | 0.272 | 3.393 | 0.001 |
| R ² | 0.253 | — | — | — | — |
| Adjusted-R ² | 0.223 | — | — | — | — |
| n | 150 | — | — | — | — |

Example: Linear regression on monthly gross income in €

Model Comparison: Nested models

| | Model 1 | | | Model 2 | | | Model 3 | | |
|-------------------------|----------|---------|-------|----------|---------|-------|----------|---------|-------|
| Term | Estimate | Std. E. | p | Estimate | Std. E. | p | Estimate | Std. E. | p |
| Intercept | 3200.845 | 450.231 | 0.000 | 2850.624 | 480.512 | 0.000 | 1512.438 | 552.317 | 0.007 |
| training (ref. no) | | | | 311.420 | 210.531 | 0.070 | 187.532 | 228.917 | 0.414 |
| edu: med. (ref. low) | | | | 260.385 | 295.784 | 0.380 | 214.759 | 297.531 | 0.472 |
| edu: high (ref. low) | | | | 892.174 | 320.987 | 0.004 | 823.915 | 321.774 | 0.012 |
| gender: f (ref. m) | | | | | | | -573.284 | 219.847 | 0.010 |
| gender: d (ref. m) | | | | | | | -917.641 | 285.392 | 0.002 |
| Age in years | | | | | | | 41.763 | 12.311 | 0.001 |
| R ² | 0.060 | — | — | 0.170 | — | — | 0.253 | — | — |
| Adjusted-R ² | 0.054 | — | — | 0.150 | — | — | 0.223 | — | — |
| n | 150 | — | — | 150 | — | — | 150 | — | — |

Example: Linear regression on monthly gross income in €

What does the model comparison tell us?

| Term | Model 1 | | | Model 2 | | | Model 3 | | |
|--------------------|----------|---------|-------|----------|---------|-------|----------|---------|-------|
| | Estimate | Std. E. | p | Estimate | Std. E. | p | Estimate | Std. E. | p |
| Intercept | 3200.845 | 450.231 | 0.000 | 2850.624 | 480.512 | 0.000 | 1512.438 | 552.317 | 0.007 |
| training (ref. no) | | | | 311.420 | 210.531 | 0.070 | 187.532 | 228.917 | 0.414 |

Model 1 (training only):

Training has a clearly positive and significant effect on income

Model 2 (training + education):

When we add education, the training effect becomes smaller and only marginally significant.

Model 3 (training + age + gender + education):

In the full model, the training effect is further reduced and clearly non-significant.

→ The “training effect” in the simple model was at least partly a confounded effect of education, age and gender.

n

150

—

—

150

—

—

150

—

—

Example: Linear regression on monthly gross income in €

What does the model comparison tell us?

| | Model 1 | Model 2 | Model 3 |
|-------------------------|----------------------|----------------------|----------------------|
| Training | Explanatory variable | Explanatory variable | Explanatory variable |
| Education | Explanatory variable | Explanatory variable | Explanatory variable |
| Age in years | Explanatory variable | Explanatory variable | Explanatory variable |
| Gender | Explanatory variable | Explanatory variable | Explanatory variable |
| R ² | 0.060 | — | 0.170 |
| Adjusted-R ² | 0.054 | — | 0.150 |
| n | 150 | — | 150 |

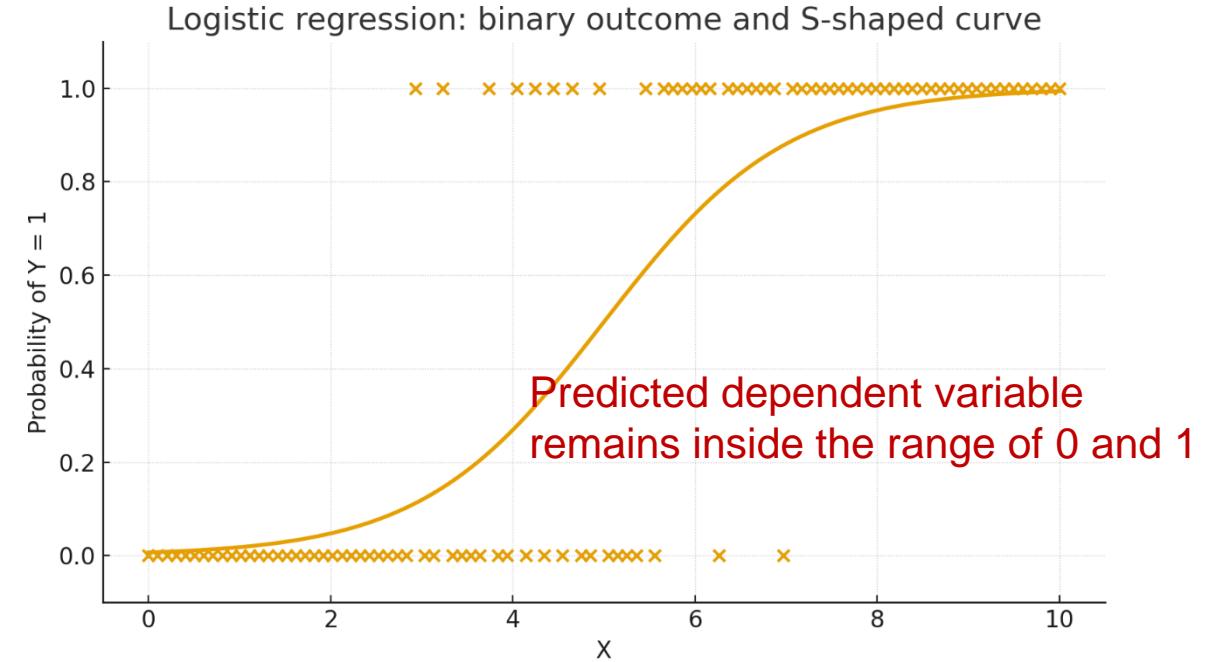
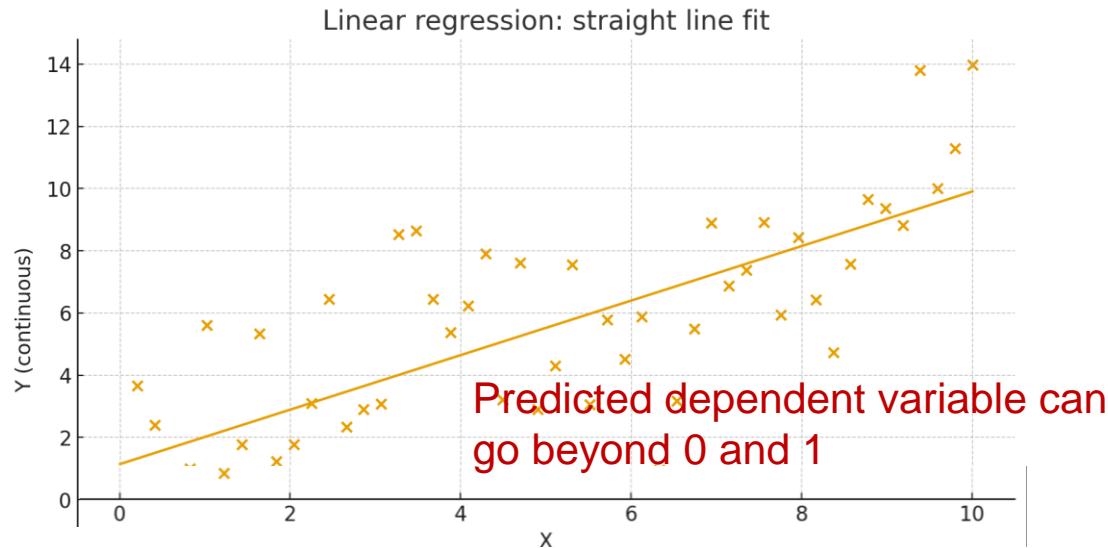
Example: Linear regression on monthly gross income in €

Linear regression: Assumptions

Linear regressions require:

- Dependent variable is **metric**.
→ Use logistic regression for categorial binary variables.
- Relationship between predictors and outcome is (approximately) **linear**.
- Residuals have **constant variance** (homoscedasticity).
- Residuals are approximately **normally distributed**.
- No strong **multicollinearity**, that means predictors are not (highly) redundant (e.g. age and years of education).

Linear vs logistic regression



Linear regression

Outcome Y is **metric/continuous** (e.g. income, test score).

Models a **straight-line relationship** between X and Y.

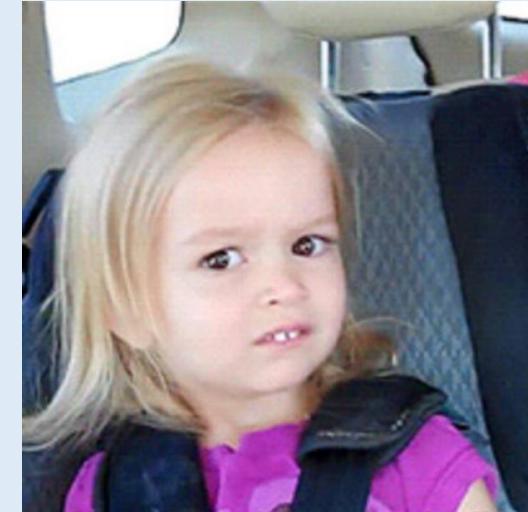
Logistic regression

Outcome Y is **categorical binary** (0/1, e.g. PhD started: yes/no).

Models a **curved, S-shaped relationship** between X and the **probability** of Y = 1.

How does a logistic regression look like?

| Term | Estimate | Std. Error | z | p-value |
|---|----------|------------|-------|---------|
| Intercept | -2.00 | 0.80 | -2.63 | 0.000 |
| Estimate for a predictor | | | | |
| How much the log-odds of $Y = 1$ change, if X_j increases by 1 unit, holding other variables constant. | | | | |
| Odds of an event are $odds = \frac{p}{1-p}$, where p is the probability that the event happens. | | | | |
| Log-odds are the natural logarithm of the odds
$\text{log-odds} = \log\left(\frac{p}{1-p}\right)$. | | | | |
| Age at graduation in years | -0.047 | 0.021 | -2.26 | 0.024 |
| Pseudo-R ² | 0.18 | — | — | — |
| n | 450 | — | — | — |



Example: Logistic regression on PhD started (0 = no, 1 = yes)

Linear regression: Odds Ratios

Odds ratio (OR) compares how likely an event is in one group (or at one value of a predictor) to how likely it is in another.

$$OR = \frac{\text{Odds in group 1}}{\text{Odds in group 2}}$$

For [categorical predictors](#), group 1 is the category you are estimating the effect for (e.g. “yes”, “female”), and group 2 is the reference category. The odds ratio compares the odds in group 1 to the odds in group 2.

For [continuous predictors](#), the odds ratio compares the odds at $X + 1$ unit (group 1) to the odds at X (group 2), e.g. With each additional year of age, the odds of starting a PhD are multiplied by 0.95 (i.e. they decrease by about 5%).

$OR = 1 \rightarrow$ no difference in odds between the groups.

$OR > 1 \rightarrow$ higher odds in the first group (event is more likely).

$OR < 1 \rightarrow$ lower odds in the first group (event is less likely).

What does the table tell us?

| Term | OR | Std. Error | z | p-value |
|------------------------|------|------------|-------|---------|
| Intercept | 0.06 | 0.80 | -3.63 | 0.000 |
| Worked as RA (ref. no) | 2.51 | 0.27 | 3.41 | 0.001 |

Worked as RA (OR = 2.51)

Graduates who worked as a research assistant (RA) have about **2.5 times higher odds** of starting a PhD than graduates who did not work as an RA, **controlling for discipline, final grade, gender and age**.

→ RA experience is associated with a **significant higher chance** of starting a PhD.

| | | | | |
|-----------------------------|------|-------|-------|-------|
| Gender: female (ref. male) | 0.71 | 0.24 | -1.46 | 0.144 |
| Gender: diverse (ref. male) | 0.33 | 0.40 | -2.75 | 0.006 |
| Age at graduation in years | 0.95 | 0.021 | -2.26 | 0.024 |
| Pseudo-R ² | 0.18 | — | — | — |
| n | 450 | — | — | — |

Example: Logistic regression on PhD started (0 = no, 1 = yes)

What does the table tell us?

| Term | OR | Std. Error | z | p-value |
|---|------|------------|-------|---------|
| Intercept | 0.06 | 0.80 | -3.63 | 0.000 |
| Worked as RA (ref. no) | 2.51 | 0.27 | 3.41 | 0.001 |
| Final grade (“very good” vs “worse”) | 2.34 | 0.25 | 3.40 | 0.001 |
| Discipline: social sciences (ref. nat.) | 0.74 | 0.29 | -1.03 | 0.302 |
| Discipline: humanities (ref. nat.) | 0.43 | 0.33 | -2.58 | 0.000 |
| Discipline: health sciences (ref. nat) | 1.16 | 0.34 | 0.44 | 0.662 |

Social sciences (OR = 0.74), Humanities (OR = 0.43)

Graduates from the **social sciences** have about **0.74 times the odds** of starting a PhD compared to graduates from the **natural sciences**, controlling for RA experience, grade, gender and age (this difference is not statistically significant).

Graduates from the **humanities** have only about **0.43 times the odds** of starting a PhD compared to **natural sciences** graduates – that is, their odds are clearly lower, even after controlling for RA experience, grade, gender and age....

Example: Logistic regression on PnD started ($0 = \text{no}$, $1 = \text{yes}$)

What does the table tell us?

| Term | OR | Std. Error | z | p-value |
|---|------|------------|-------|---------|
| Intercept | 0.06 | 0.80 | -3.63 | 0.000 |
| Worked as RA (ref. no) | 2.51 | 0.27 | 3.41 | 0.001 |
| Final grade (“very good” vs “worse”) | 2.34 | 0.25 | 3.40 | 0.001 |
| Discipline: social sciences (ref. nat.) | 0.74 | 0.29 | -1.03 | 0.302 |
| Discipline: humanities (ref. nat.) | 0.43 | 0.33 | -2.58 | 0.000 |

Age at graduation (OR = 0.95)

For each additional year of age at graduation, the odds of starting a PhD are multiplied by about **0.95**, that is, they **decrease by roughly 5% per year**, controlling for discipline, RA experience, grade and gender.

| | | | | |
|----------------------------|------|-------|-------|-------|
| Age at graduation in years | 0.95 | 0.021 | -2.26 | 0.024 |
| Pseudo-R ² | 0.18 | — | — | — |
| n | 450 | — | — | — |

Example: Logistic regression on PhD started (0 = no, 1 = yes)

Pseudo-R²

In [linear regression](#), R² is based on [explained variance](#) in a continuous outcome and a clean decomposition of total variance ($SS_{\text{total}} = SS_{\text{explained}} + SS_{\text{residual}}$).

In [logistic regression](#), the outcome is [binary](#), the model works and we do not have the same variance decomposition.

→ Instead, we use [pseudo R² measures](#) (e.g. McFadden, Cox–Snell, Nagelkerke) that compare the fitted model to a [null model](#) (with no predictors) based on [likelihoods](#).

- They show **how much better** the model fits compared to a model without predictors, **not** the exact “percentage of variance explained”.
- Pseudo R² values are usually **much smaller** than R² in linear regression (e.g. 0.10–0.30 can already indicate a useful model).

What does the table tell us?

| Term | OR | Std. Error | z | p-value |
|---|------|------------|-------|---------|
| Intercept | 0.06 | 0.80 | -3.63 | 0.000 |
| Worked as RA (ref. no) | 2.51 | 0.27 | 3.41 | 0.001 |
| Final grade (“very good” vs “worse”) | 2.34 | 0.25 | 3.40 | 0.001 |
| Discipline: social sciences (ref. nat.) | 0.74 | 0.29 | -1.03 | 0.302 |
| Discipline: humanities (ref. nat.) | 0.43 | 0.33 | -2.58 | 0.000 |

Pseudo-R²

The pseudo-R² of 0.18 suggests that the logistic regression model provides a modest improvement in explaining who starts a PhD compared to a model without predictors.

| | | | | |
|----------------------------|------|-------|-------|-------|
| Age at graduation in years | 0.95 | 0.021 | -2.26 | 0.024 |
| Pseudo-R ² | 0.18 | — | — | — |
| n | 450 | — | — | — |

Example: Logistic regression on PhD started (0 = no, 1 = yes)

Your turn!

1. Copy folder Materials from to your local machine

→ <https://github.com/Data-Science-Center-UB/Intro-Quantitative-Analysis-R>

→ Path do subfolder Exercises will be your **working directory**

→ Folder content: data sets, exercises without & with solution

2. Do exercise “**04_qa_multi_ex**”

Version with solution: **04_qa_multi_ex_solution**



Wrap-up

What we did not cover today....

- Variable transformation
 - Factor Analysis
 - Missing values: Imputation
 - Interaction effects
 - Multinomial logistic regression
 - Model comparison in logistic regression, Average Marginal Effects
 - Event-History-/ Survival-Analysis: censored data, time-varying covariates
 - Visualization
 - Better documentation with Rmarkdown
-

Resources



Self-Learning Tool

<https://datasciencebox.org/>

Cheat Sheets

<https://posit.co/resources/cheatsheets/>

Literature

Aslam, M., Imbad Ullah, M. (2023). *Practicing R for Statistical Computing*. Singapore: Springer Nature.
<https://doi.org/10.1007/978-981-99-2886-6>

Kronthaler, F., Zöllner, S. (2021). *Data Analysis with Rstudio*. Wiesbaden: Springer VS.
<https://doi.org/10.1007/978-3-662-62518-7>

Zamora Saiz et. a. (2020). *An Introduction to Data Analysis in R*. Springer Cham <https://doi.org/10.1007/978-3-030-48997-7>

Wickham, H., Cetinkaya-Rundel, M., Grolemund, G. (2023). *R for Data Science. Import, Tidy, Transform, Visualize, and Model Data*. <https://r4ds.hadley.nz/>

Course content

- Hello world
- Exploring data
- Data science ethics
- Making rigorous conclusions
- Looking further
- Interactive tutorials
- Project
- Exams

Wrangling and tidying data

Unit 2 - Deck 5: Tidy data

 Slides
 Source
 Video

 Reading:
JSS :: Tidy data

Unit 2 - Deck 6: Grammar of data wrangling

 Slides
 Source
 Video

Unit 2 - Deck 7: Working with a single data frame

 Slides
 Source
 Video

 Reading:
R4DS :: Chp 5 - Data transformation

Slides, videos, and application exercises

- Visualising data
- Wrangling and tidying data
- Importing and recoding data
- Communicating data science results effectively
- Web scraping and programming
- Labs
- Homework assignments

 Edit this page
 Report an issue

Data transformation with dplyr :: : CHEATSHEET

dplyr functions work with pipes and expect **tidy data**. In tidy data:

- Each variable is in its own column
- Each observation, or case, is in its own row
- pipes
- x %> f() becomes: f(x, y)

Summarize Cases

Apply **summary** functions to columns to create a new table of summary statistics. Summary functions take vectors as input and return one value (see back).

summary function

- summarise(data, ...)
- Compute table of summaries.
- group_by(data, ..., .by = c(...))
- group_by(data, ..., .by = c(...), .by = mean(mpg))
- count(data, ..., wt = NULL, sort = FALSE, name = NULL) Count number of rows in each group defined by the grouping variables. Also **tally()**, **add_count()**, **pull()**, **tail()**.
- mtcars %>% count(cyl)

Group Cases

Use **group_by**, **by** and **add = FALSE**, **drop = TRUE** to create a grouped copy of a table grouped by columns in **dplyr** functions will manipulate each "group" separately and combine the results.

grouped table

- mtcars %> group_by(cyl) %> summarise(mpg = mean(mpg))

Alternate grouping syntax with **by** as an argument.

use **rowwise**, ... to group data into individual rows. **dplyr** functions will compute results for each row. Also applies functions to list-columns. See tidycheat sheet for last-column workflow.

grouped table

- starwars %> rowwise() %> mutate(lflim = count %> length(lflim))
- ungroup(..., .by = c(...)) Returns ungrouped copy of table.
- ungroup(..., .by = c(...), .by = drop)
- ungroup(..., .by = drop)
- ungroup(..., mtcars)

Manipulate Cases

EXTRACT CASES

Row functions return a subset of rows as a new table.

- filter(data, ..., preserve = FALSE) Extract rows that meet logical criteria.
- mtcars %> filter(cyl > 20)
- distinct(data, ..., keep = FALSE) Remove rows with duplicate values.
- mtcars %> distinct(cyl)
- slice(data, ..., preserve = FALSE) Select rows by position.
- mtcars %> slice(10)
- slice(data, ..., n, prop, weight = FALSE) Randomly select rows. Use n to select a number of rows and prop to select a fraction of rows.
- mtcars %> slice(1:10, 1/10, replace = TRUE)
- slice_min(data, order_by = ..., n, ties = "TRUE") and **slice_max** Select rows with the lowest and highest values.
- mtcars %> slice_min(mpg, order_by = mpg, n = 2)
- slice_head(data, n) Select the first n rows.
- mtcars %> slice_head(5)

Logical and operator uses to work with filter()

| | | | | | |
|----|----|---------|---------|---|-------|
| < | <= | is.na() | %in% | ! | xor() |
| != | > | >= | is.na() | ! | & |

See **base::isTRUE** and **Comparison for help**.

ARRANGE CASES

- arrange(data, ..., by = group = FALSE) Order rows by values of a column or columns (low to high, use with **desc**) to order from high to low.
- mtcars %> arrange(desc(cyl))

ADD CASES

- add_rows(data, ..., before = NULL, after = NULL) Add one or more rows to a table.
- cars %> add_rows(1, dist = 1)

Manipulate Variables

EXTRACT VARIABLES

Column functions return a set of columns as a new vector or table.

- pull(data, var = c(...), name = NULL) ... Extract column values as a vector, by name or index.
- mtcars %> pull(cyl)
- select(data, ...) Extract columns as a table.
- mtcars %> select(cyl)
- relocate(data, ..., after = NULL, before = NULL) Move columns to new position.
- mtcars %> relocate(gear, cyl, after = last(cyl))

Use these helpers with **select** and **across**

e.g. mtcars %> select(cyl)

- contains(match, ..., name = NULL) ...
- end_with(match, ..., name = NULL) ...
- starts_with(match, ..., name = NULL) ...
- matches(match, ..., everything)

MANIPULATE MULTIPLE VARIABLES AT ONCE

d <- tribble(~x, ~y, ~z, 1, 2, 3, 4, 5, 6, 7)

- across(cols, ..., names = NULL) ... **SUMMARIZE** or mutate multiple columns in the same way.
- d %> across(cyl, ~min(mpg))
- d %> across(cyl, ~mean(mpg))
- d %> across(cyl, ~mutate_all(sum))

row wise data

- d %> mutate_all(~sum_across(1:2))

MAKE NEW VARIABLES

Apply **vectorized functions** to columns. Vectorized functions take vectorized inputs and return a vector of the same length as output. (see back)

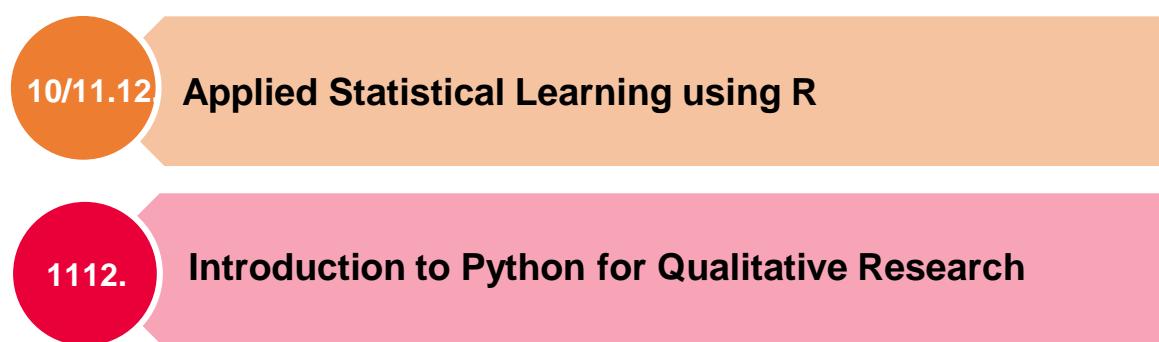
- vectorized function**
- mutate(data, ..., .keep = "all", after = NULL, after = f) **NEW** Compute new columns (also, **add**, **column**).
- mtcars %> mutate(gear = 1 / mpg)
- mtcars %> mutate(gear = 1 / mpg, gear = "none")
- rename(data, ...) Rename columns. Use **rename**, **with** to rename with a function.
- mtcars %> rename(gear, gear = mpg)



DataNord Trainings & Workshops

- **Interdisciplinary and subject-specific workshops**
(esp. in social sciences, marine and environmental sciences, health sciences, and humanities).
- **On-demand customized training** for graduate programs, departments, etc.

... more on dsc-ub.de/en/qualification



And many more...

„Data Snacks“

Data Insights in 30 Minutes!

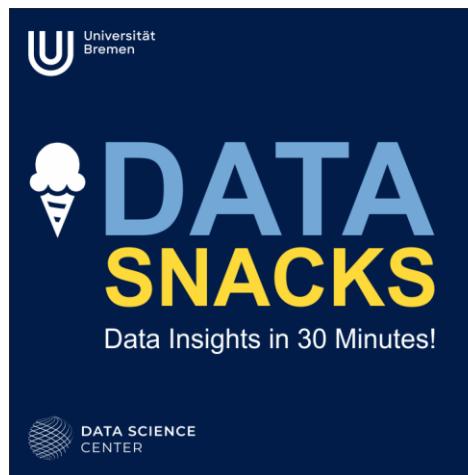
Short and engaging sessions on different “data topics”

Next Data Snack Series: October to December 2025

- Free & open for everyone
- Online
- No registration required, just drop in!

More information at:

https://www.dsc-ub.de/data_science_forum.php



The menu lists six sessions:

| Date | Title | Speaker |
|--------|---|---|
| 23.10. | Interactive Visualization in R with Shiny: Building Your Research Data Skills | Dr. Maryam Movahedifar Uni Bremen |
| 30.10. | Teach Data, Teach Better: Enhancing University Teaching with Research Data | Dr. Susanne de Vogel & Franziska Richter Uni Bremen |
| 06.11. | National High-Performance Computing (NHR) Roadshow on Digital Humanities | Anja Gerbes Uni Göttingen |
| 13.11. | Rethinking Research Data Management: Inclusive Practices for Every Mind | Sarah Büker Uni Bremen |
| 20.11. | Basics of Software Publication | Carina Haupt DLR |
| 04.12. | Beyond ChatGPT: AI Systems for Qualitative Research | Nele Fuchs Uni Bremen |

DataNord Consulting Service



How can I create a data management plan?

How do I prepare my data to apply multivariate statistics?

What do I need to consider when collecting personal data?



- Interdisciplinary Help Desk for researchers from all DataNord institutions
- Support in the application phase and research process
- Free of charge!
- More information: <https://dsc-ub.de/en/consultation.php>

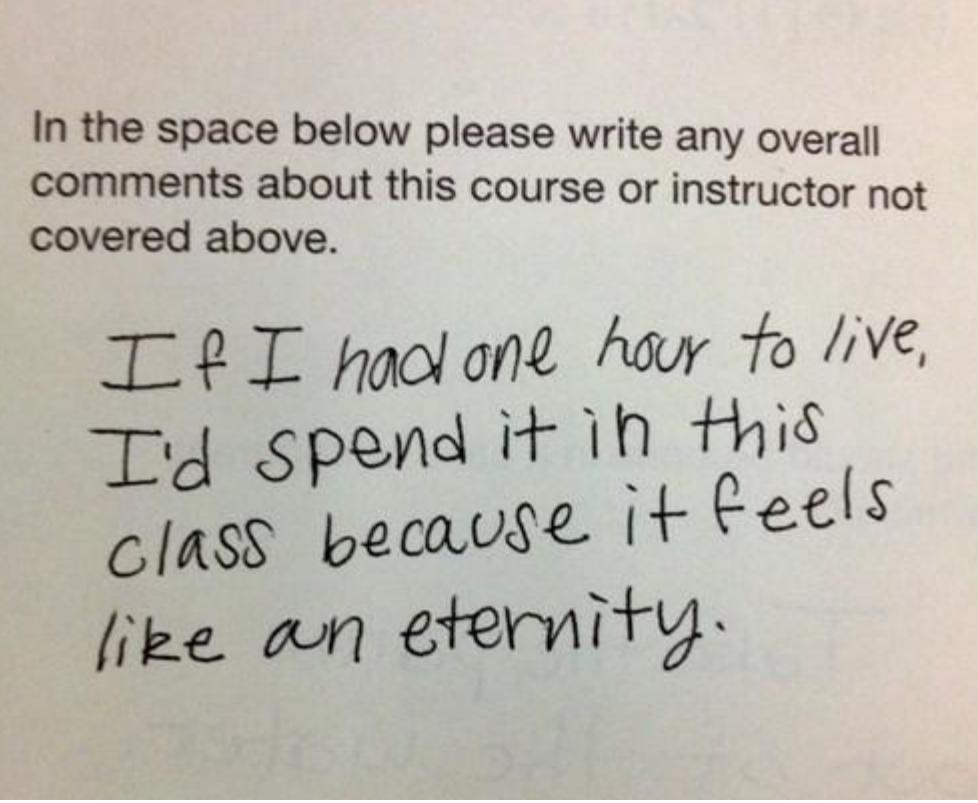
Sign Up Now!



[www.bremen-research.de/
datanord/newsletter](http://www.bremen-research.de/datanord/newsletter)

- **Monthly Updates!**
- Target group: Researchers and anyone who is interested in our offers
- Overview of DataNord **trainings, networking opportunities and other events** in the next month
- Additional announcements

Evaluation



<https://www.dsc-ub.de/evaluation-DSC-2025-26.php>



DATA SCIENCE
CENTER



Thank You for Your Participation!

CONTACT

Dr. Susanne de Vogel
Data Scientist | Help-desk
devogel@uni-bremen.de
 dsc-ub.de

