Einführung in R und RStudio

Daniela Palleschi

Vorlesung Mi. den 18.10.2023

Inhaltsverzeichnis

Wiederholung

Last week we...

- installed R and RStudio
- created our first R script
- did simple arithmetic with objects and vectors

Wiederholung

```
x <- c(1,2,3)
y <- sum(1,2,3)
```

- what do the vectors `x` and `y` contain?
- The object `x` contains `1, 2, 3`.
- The object `y` contains `6`.

Heutige Ziele

Today we will learn...

- what dataframes are
- the difference between categorical and continuous data
- how to produce plots with `ggplot`

- choose the right plot for our data

Lust auf mehr?

- Chapter 2 (Data Visualisation) in Wickham et al. (2023), up until section 2.4
- Chapter 3 (Data Visualisation) in Nordmann & DeBruine (2022)

Vorbereitung

In your RProject folder...

- create a new folder called `moodle`
  - download the Moodle materials from today and save them there
- create a new folder in `notes` called `02-datenviz1`
- open a new `.R` script
  - save it in the new folder

0.0.1 Packages

- Pakete (installiert und) ladt
  - `tidyverse`
  - `languageR`
  - `ggthemes`
  - `patchwork`

```
# in the CONSOLE: install packages if needed
install.packages("tidyverse")
install.packages("languageR")
install.packages("ggthemes") # for customising our plots
install.packages("patchwork") # plot layouts
```

```
# Pakete laden
library(tidyverse)
library(languageR)
library(ggthemes)
library(patchwork)
```

1 Data frames

- data frames are a collection of variables, where
  - each variable is one column
  - each row is a single observation/data point
  - each cell in a row is linked
- data frames are just like spreadsheets, but are rectangular
- different words for data frames:
  - data frame
  - dataset
  - tibble (in the `tidyverse`)

1.1 Talking about datasets

- when we talk about our data, we use certan words to refer to different parts:
- a **variable**: a quantity, quality, or property you can measure
- a **value**: the state of a variable when you measure it
- an **observation**: set of measurements made under similar conditions
  - will contain several values each associated with a variable
  - an observation for a single variable is sometimes called a *data point*
- **tabular data** is a set of values, each associated with a variable and an observation
  - tabular data is *tidy* if each value is placed in its own *cell*, each variable in its own column, and each observation in its own row

1.2 Categorical and continuous variables

- how we visual the distribution of a variable depends on what type of data it represents: *categorical* or *numerical*
- a variable is *categorical* if it can take a small set of values that can be grouped together
- e.g., old/young, short/tall, grammatical/ungrammatical, L1/L2-speaker
- a variable is *numerical* (i.e., quantitative) if it can take on a wide range of numerical values
  - and it would make sense to add, subtract, compute the mean, etc.
  - can be *continuous* (decimal points make sense, e.g., 1.5cm)

  - or *discrete* (decimal points do *not* make sense, e.g., 1.5 children doesn't make sense)
  - age, height, reaction times, format frequencies
- we produce different plots depending on what type of variables we want to visualise

2 Lexical Decision Task

- our first dataset contains data from a lexical decision task (LDT)
- in the LDT, participants press a button to indicate whether a word is a real word or a pseudoword

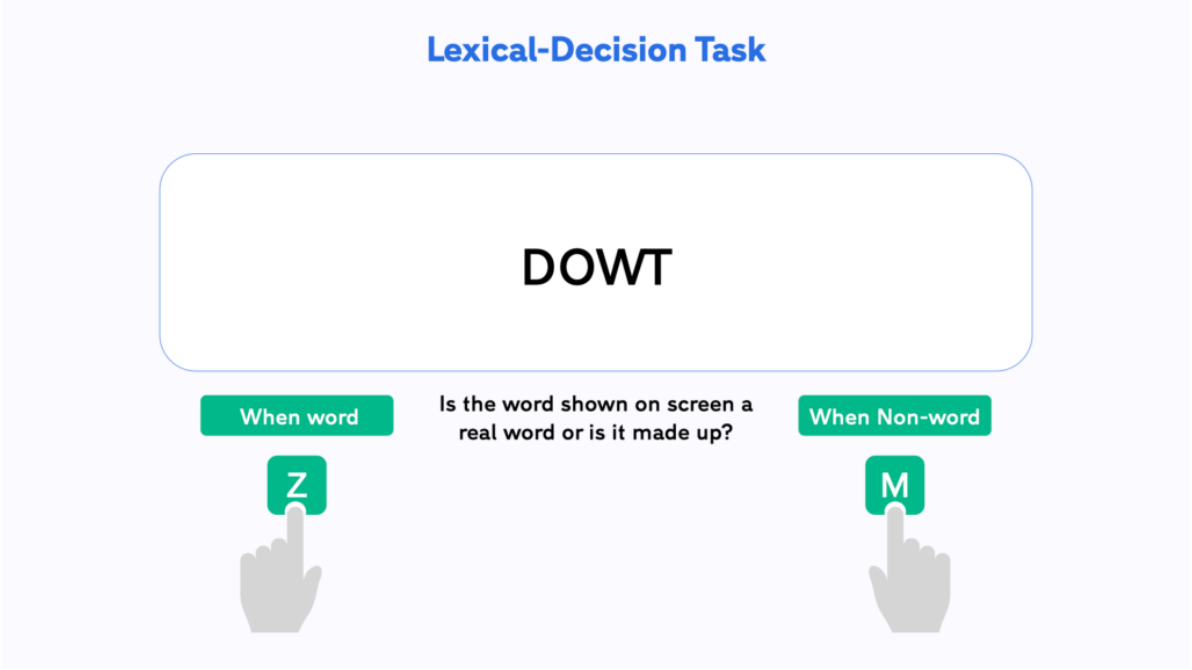
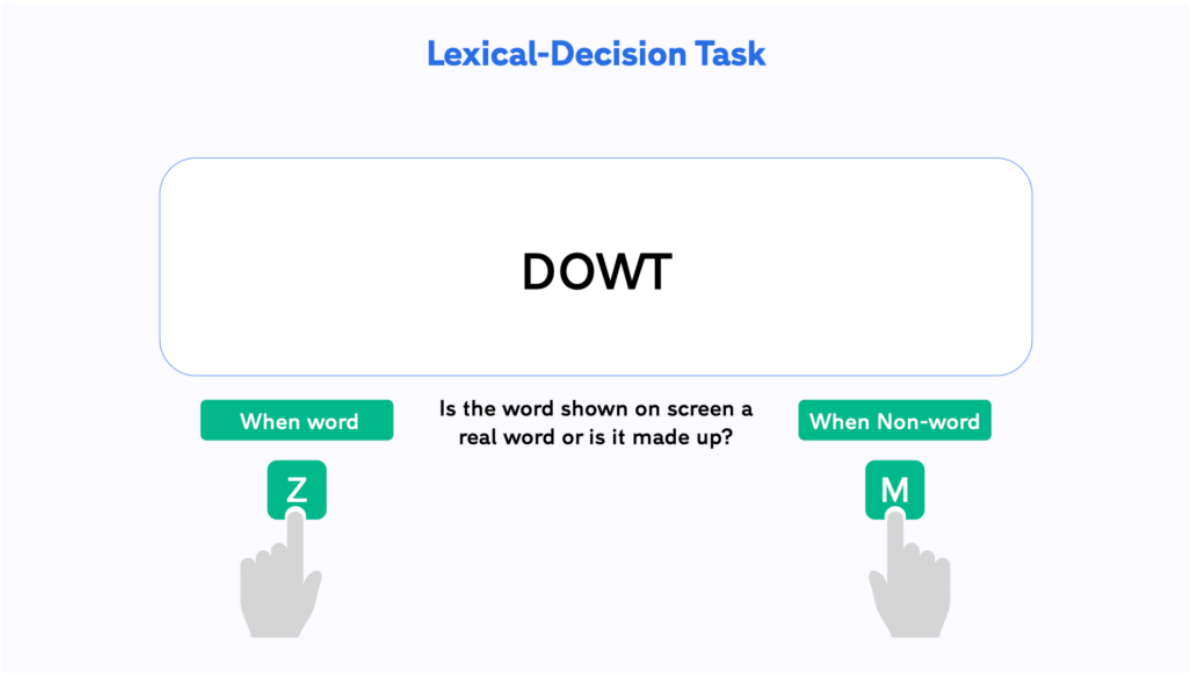


Abbildung 1: Source: https://www.testable.org/wp-content/uploads/2022/11/Lexical_decision_task-1024x576.png

2.1 LDT variables

- common variables collected in a lexical decision task experiment are:
  - reaction time
  - accuracy (correct/incorrect)
  - word category (e.g., real/pseudo, noun/verb)



  - word frequency
- additional variables that might be collected could be:
  - participant demographics (e.g., age, L1/L2, gender)

3 `lexdec` dataset

- `languageR` is a companion package for the textbook (**baayen_analyzing_2008?**)
  - contains linguistic datasets, e.g., `lexdec`
- the `lexdec` dataset contains data for a lexical decision task in English
  - we will be working with variables such as reaction times and accuracy

3.1 `lexdec` variables

- a list of some of the variables is included in Tabelle **??**

Tabelle 1: Data dictionary for `df_lexdec`: Lexical decision latencies elicited from 21 subjects for 79 English concrete nouns, with variables linked to subject or word.

variable	description
Subject	a factor for the subjects
RT	a numeric vector for reaction times in milliseconds
Trial	a numeric vector for the rank of the trial in the experimental list.
Sex	a factor with levels F (female) and M (male).
NativeLanguage	a factor with levels English and Other, distinguishing between native and nonnative speak

3.2 LDT research questions

- before we conduct an experiment, we have research questions that we want to answer with the data
  - today we'll address the following question:
    - do the reaction times differ between native and non-native speakers?

3.3 Load the data

- our data is available in the `lanaugeR` package we've already loaded
  - to print the data, just type the name of the dataset and run it
- below we only see a few variables, but you should see more in your console

```
lexdec
```

```
  Subject       RT Trial Sex NativeLanguage Correct PrevType PrevCorrect
1      A1 6.340359    23   F        English correct     word     correct
2      A1 6.308098    27   F        English correct  nonword     correct
3      A1 6.349139    29   F        English correct  nonword     correct
4      A1 6.186209    30   F        English correct     word     correct
5      A1 6.025866    32   F        English correct  nonword     correct
6      A1 6.180017    33   F        English correct     word     correct
```

- how many variables do we have? observations?

3.3.1 Save data as an object

- to save the data in our Environment, we have to assign it a name
  - let's call it `df_lexdec`, which means "data frame lexical decision"