Unit 2 - Descriptives in R

Flinta* R-Tutorium

WU Wien

18.03.2024

Where are we?

We have done:

1. Get to know R's surface
2. Learned basic codes

What is next?

1. Learn about packages
2. Practice to work with real data
3. Compute descriptive statistics

Packages in R

1. Useful for more complex tasks
2. R is open source, many people contributed with so-called *packages*

How to access packages:

1. Use commands: *install.packages("package_name"))* & *library("package_name")*
2. Help window: Klick on *Installation* and then search for you desired package name

Today's packages

1. **tidyverse**: Especially, *dplyr*, which summarizes, creates and sorts variables (`https://dplyr.tidyverse.org/`
2. **magrittr**: Pipe operator. Connects multiple commands / functions together (`https://magrittr.tidyverse.org/`)

Real data

Checklist:

1. Load data
2. Understand the data (how is it stored in R, how are the variables structured)
3. Modify data to fit your needs (change data, create data)

Example

E.g You have the amount of sleep stored in minutes, but want to know the hours $\rightarrow$ divide by 60!

4. Save data

Descriptives I

- **Mean:** The average of a variable in a sample. Calculated by adding up all values and dividing the sum by the number of observations.

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{1}$$

- **Median:** The middle value of a set of values when they are sorted.

$$1, 3, 3, \boxed{6}, 7, 8, 9 \tag{2}$$

Descriptives II

- **Variance:** A measure of the amount of variation in a variable. How much does individual value differ from the mean of the variable?

$$\text{Var} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \tag{3}$$

- **Standard deviation:** Also a measure of the amount of variation in variable.

$$\sigma = \sqrt{\text{Var}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{4}$$

$\rightarrow$ same units as the variable x, which facilitates interpretation.

Descriptives III

- **Skeweness:** How skewed is a distribution? Strongly skewed distributions result in larger gap between mean and median.