impr-is

An Introduction to Statistical Analysis of Big Data in R

Hannah Götsch

hannah.goetsch@uni-tuebingen.de

2025 Boot Camp - 25th September

Tutorial Objectives

By the end of this session, you will be able to:

- Compute basic descriptive statistics
- Visualize data with histograms and boxplots
- Understand hypothesis testing concepts
- Run and interpret simple statistical tests

Course materials (updated afterwards)

github.com/Ha-nn-ah/intro_stats_R

Why Statistics?

Aim

developing and applying methods to collect, summarize, analyze, and interpret data

Main Tasks:

1. Estimating (point and interval estimates)
2. Hypothesis Testing
3. Predicting & Classifying

Data = Information + Noise

- Extract information ⇒ parameters
- Quantify uncertainty ⇒ random fluctuations

Types of Data

- Qualitative variables: things you put into categories
- Quantitative variables
  - Discrete: things you count
  - Continuous: things you measure

Get an Initial Overview of Data by ...

- Computing summaries (descriptive statistics)
- Generating graphs (EDA)

Helps to choose an appropriate statistical model for data!

R and RStudio

- language & environment for statistical computing & graphics
- highly extensible
- open source and free

- integrated development environment (IDE)
- provides a user-friendly interface for R
- free version is enough

Loading Data in R

```r
# Load a CSV file
bigdata <- read.csv("path_to_file/file.csv")

# Load built-in dataset
data("iris")
head(iris) # Inspect the dataset
```

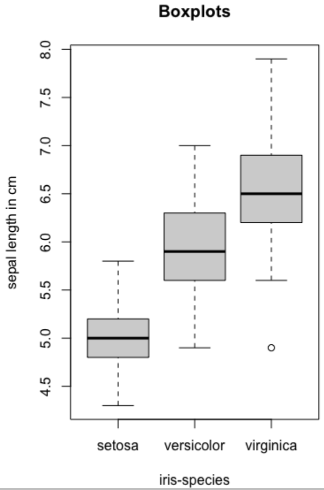
What are Descriptive Statistics?

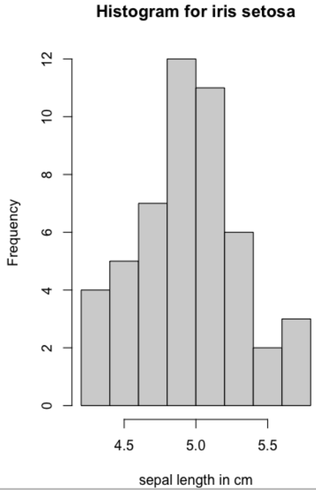
- Summarize and describe data $x_1, ..., x_n$ $(x_{[1]} \leq ... \leq x_{[n]})$
- Measures of location:
  - (arithmetic) mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$,
  - median: $\tilde{x} = x_{(n+1)/2}$ for n odd, $\tilde{x} = x_{n/2} x_{(n+2)/2}$ for n even
- Measures of spread:
  - variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$,
  - standard deviation $s = \sqrt{s^2}$,
  - inter-quartilerange $\hat{Q}(0.75) - \hat{Q}(0.25)$
- $100\alpha\%$-quantile $\hat{Q}(\alpha)$:
  at least $100\alpha\%$ of data below, and at least $100(1-\alpha)\%$ above

Summary Statistics in R

```
mean(iris$Sepal.Length)
median(iris$Sepal.Length)
sd(iris$Sepal.Length)
var(iris$Sepal.Length)
summary(iris$Sepal.Length)
```

Exploratory Data Analysis (EDA)





Visual Summaries in R

```
hist(iris$Sepal.Length)

boxplot(iris$Sepal.Length)

# Library for nicer and more complicated graphs:
  library(ggplot2)
```

Discussion

- What insights do the boxplot and histogram provide?
- When would you prefer the median over the mean?
- What new challenges arise when visualizing or summarizing very large datasets (big data)?

Efficient Summaries in R with dplyr

```r
library(dplyr)

bigdata %>%
  summarise(
    mean_val = mean(variable, na.rm = TRUE),
    median_val = median(variable, na.rm = TRUE),
    sd_val = sd(variable, na.rm = TRUE)
  )
```

Working with Big Data (in R)

```r
# Take a sample for quick exploration
set.seed(123)
sampled <- bigdata %>% sample_n(1000)

summary(sampled)
```

Point Estimator

- aim: estimate unknown parameter
- estimator = method/algorithm to obtain estimate (= actual number) of parameter
- quality of a specific estimator depends on underlying model
- constructing estimator: "Maximum Likelihood" principle, plug in principle (e.g. sample median for population median, method of moments), least squares principle (regression), etc.
- sometimes different principles lead to same estimator

"First Principle of Statistics:
Do not (blindly) trust any Principle"
L. LeCam

(further reading: Confidence Intervals)

Basics of Hypothesis Testing

Two competing hypotheses:

- Null hypothesis (H_0): no effect
- Alternative hypothesis (H_a): effect exists

Goal: confirm alternative hypothesis (interesting finding).
When in doubt, we decide for null hypthesis.
Usually we cannot prove null hypothesis, but only rule it out.

Is there sufficient evidence against H_0?

(further reading: Multiple Hypothesis Testing)

Classical way test is carried out

Consider H_0 ($\Theta = \Theta_0$) vs. H_a (estimator $\hat{\Theta}$ of Θ)

- Compute a test statistic T from data, often $T = \frac{\hat{\Theta} - \Theta_0}{\hat{SE}(\hat{\Theta})}$
- Decide for alternative, if T exceeds some cut off value.
- Alternative: Compute a p-value and reject if $< \alpha$ (usually $\alpha = 0.05$)

One-Sample t-test in R

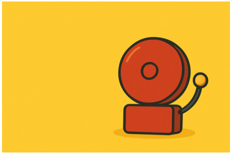
```
# Is the mean sepal length for iris setosa (or
   iris versicolor) different from the overall
   mean?

t.test(iris[iris$Species == "setosa",]$Sepal.
   Length, mu=mean(iris$Sepal.Length))

t.test(iris[iris$Species == "versicolor",]$Sepal
   .Length, mu=mean(iris$Sepal.Length))
```

Significance Level

p-value $\leq$	evidence against H_0
0.1	weak
0.05	moderate
0.01	strong
0.001	very strong

Level α of a test: Maximum probability of type I error.

Types of Errors

	H_0 true	H_a true
decision for H_0	correct 	type II error (false negative) 
rejecting H_0 (decision for H_a)	type I error (false positive) 	correct 

How to Choose a Statistical Test

1. What is your research question?
   (comparing means, testing relationships ...)
2. What type of data do you have?
   (numeric, categorical ...)
3. What other properties does your data have?
   (independent, pairwise, normal distribution ...)

Always check the requirements for your statistical test!
(most of the time pre-tests are needed)

Two-Sample t-test in R

```
# Compare sepal length for iris setosa and iris
   versicolor:
t.test(Sepal.Length ~ Species, data=iris[iris$
   Species %in% c("setosa","versicolor"),])

# Test for normal distribution:
shapiro.test(iris[iris$Species == "setosa",]$
   Sepal.Length)
shapiro.test(iris[iris$Species == "versicolor",]
   $Sepal.Length)
```

Discussion

- What does a p-value mean in plain language?
- How does sample size affect statistical power?

Interpretation of p-Values

- Smaller a p-value $\Rightarrow$ more evidence against H_0
- Gives probability under H_0 to get a deviation from expected value under H_0 that is larger or equal the observed one
- **Not** the probability that H_0 is true

Significance $\neq$ Importance

Big Data Considerations in Hypothesis Testing

- Very large samples often yield very small p-values
- Practical vs statistical significance: both matter
- Always consider effect sizes and confidence intervals

Mini-Exercise

- Choose another variable from `iris`
- Compute descriptive statistics
- Run a t-test (or statistical test of your choice)

Try on data of your choice (maybe your own data?)!

Key Takeaways

- Descriptive statistics summarize data (foundation for exploring big data)
- Hypothesis testing helps answer research questions
- R provides tools for computation and visualization
- Interpretation is as important as significance