Descriptive Statistics

EDA: exploratory data analysis – Part I

MF9130E – Introductory Course in Statistics
25.04.2023

Chi Zhang
chi.zhang@medisin.uio.no
Oslo Center for Biostatistics and Epidemiology
Department of Biostatistics, UiO

Outline

8:30-9:30	Introduction to R and RStudio
Break	
9:45-10:45	Recap theory: descriptive statistics
	Guided practice (exercise 1, 2)
Break	
	Practice
11:15-11:30	Summary and wrap up

Lab notes for today:
(under *R Lab and Code* tab)

Getting started in RStudio

Introduction to R

EDA I

Link to *R Lab and Code*

https://ocbe-uio.github.io/teaching_mf9130e/lab/overview.html

Descriptive statistics, EDA

EDA: Exploratory Data Analysis

In contrast to Confirmatory analysis (e.g. hypothesis tests)

The goal of EDA is to get a first impression of your data

Descriptive statistics is part of the process of exploration

For example, what is the average of ‘height’ in my data?

In this session, we learn how to explore a dataset with

- Review **descriptive (summary) statistics**
- Some **simple data manipulation** techniques
- **Visualisation** with histogram, boxplot, scatterplot

Descriptive statistics

Central measures

Mean (average)
$(x1 + x2 + .. + xn)/n$

Median
Half values smaller than this value; half greater

Mean is sensitive to extreme values (outliers)

Variation measures

Range

Interquartile range (percentiles, quartiles)

Variance

Standard deviation

Descriptive statistics

Mean

Median

Minimum, maximum

Quantiles (top 5% = 0.95 quantile)

Quartiles (0.25, 0.5, 0.75)

Variance, standard deviation

```r
# x is a continuous variable

mean(x)

median(x)

min(x), max(x)

summary(x)

quantile(x, 0.95)

quantile(x, 0.25)

var(x), sd(x)
```

Simple data manipulation

When you get a dataset, the first thing to do is to get an overview of your dataset:

How many observations?

How many variables are measured?

What data types exist?

```
# df is a data.frame

# first 6 rows
head(df)

# number of observations
nrow(df)

# column names (variables)
colnames(df)

# what data types?
str(df)
class(df$var1)
```

Descriptive statistics with plots

Data visualization is a very effective way to explore, and present your data.

We focus on **base R**
(rather than more complex solutions: ggplot2)

```
# x is a continuous variable

hist(x)

boxplot(x)
```

(Guided practice) explore penguins dataset

Now we are going to practice what we have introduced in R, using **penguins** dataset.

If you didn’t see some commands, you can check the lab notes: **EDA I**

Exercise 1 (weight)

You can open the **exercise (without solution)**, and **lab notes (with solution)** side by side

1a) Generate a variable named weight, with the following measurements

```
50 75 70 74 95 83 65 94 66 65
65 75 84 55 73 68 72 67 53 65
```

Exercise 1 (weight)

1b) Make a simple descriptive analysis of the variable. What are the mean, median, maximum, minimum and quantiles?

Exercise 1 (weight)

1c) Make a histogram of the variable.

Exercise 1 (weight)

1d) Make a boxplot. What do the two dots on the top represent?

Exercise 2 (lung function)

2a) Download and open PEFH98-english data into R

(Use the file **PEFH98-english.csv**)

Exercise 2 (lung function)

2b) How many observations are there? (Number of subjects)

How do you get a list of variables from your dataset?

Exercise 2 (lung function)

2b) Make a histogram of the following variables. Compute means, and interpret the results.

Height, weight, age, pefsitm, pefstam

(Illustrate height)

Exercise 2 (lung function)

2c) Make histograms for the variable **height** and **pefmean** for **men** and **women** separately.

Also make boxplots.

What conclusion can you draw?

(Illustrate height for men)

Exercise 2 (lung function)

2d) Make three scatterplots to compare

Pefmean with **height**

Pefmean with **weight**

Pefmean with **age**

(Illustrate pefmean with height)