

E0410 Fundamentals of Statistics for Scientific Data Using R

by Daria Sapunova, PhD student, RECETOX

daria.sapunova@recetox.muni.cz

Bohunice, D29, room 123

Course description

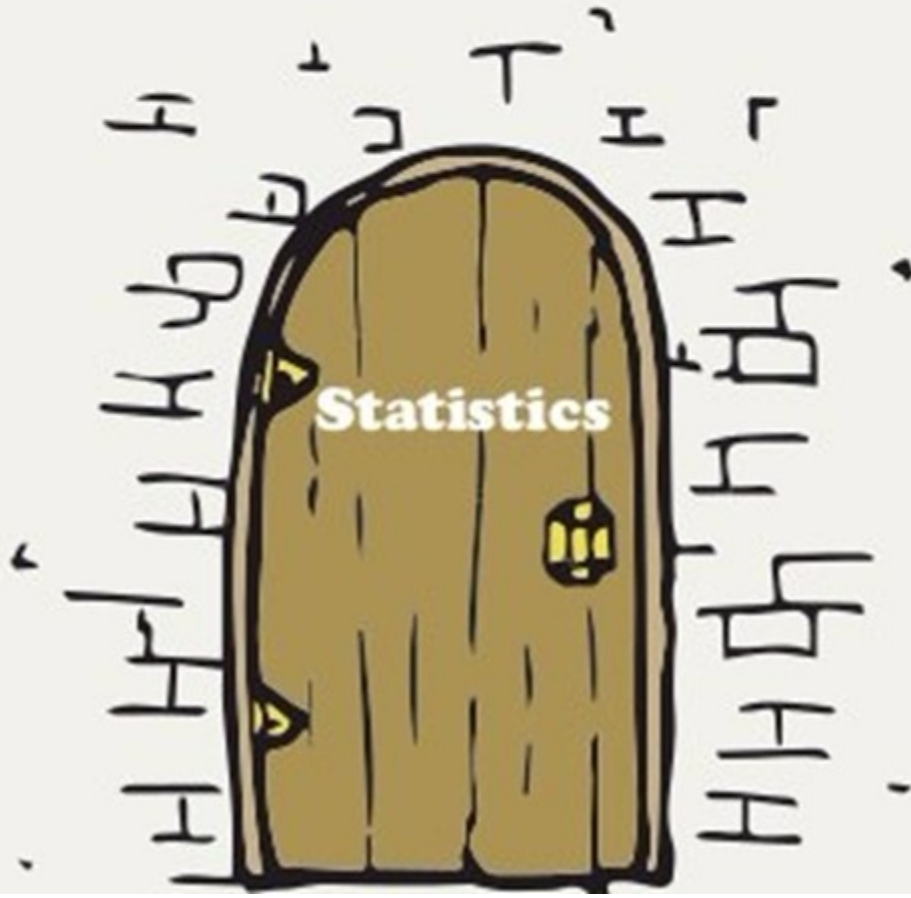
Overview

- **No statistical knowledge or programming skills are required.**
- **Topics to be covered:**
 - **Descriptive statistics**
 - **Visualization**
 - **Data cleaning**
 - **The most common R packages in R (dplyr, tidyverse etc.)**
 - **Hypothesis testing**
 - **Parametric and non-parametric statistics**
 - **Parametric tests (t-test, ANOVA)**
 - **Non-parametric tests (Kruskal-Wallis, Mann-Whitney)**
 - **Pearson and Spearman correlation**
 - **Linear regression (OLS)**
 - **Multiple regression**

Theoretical part

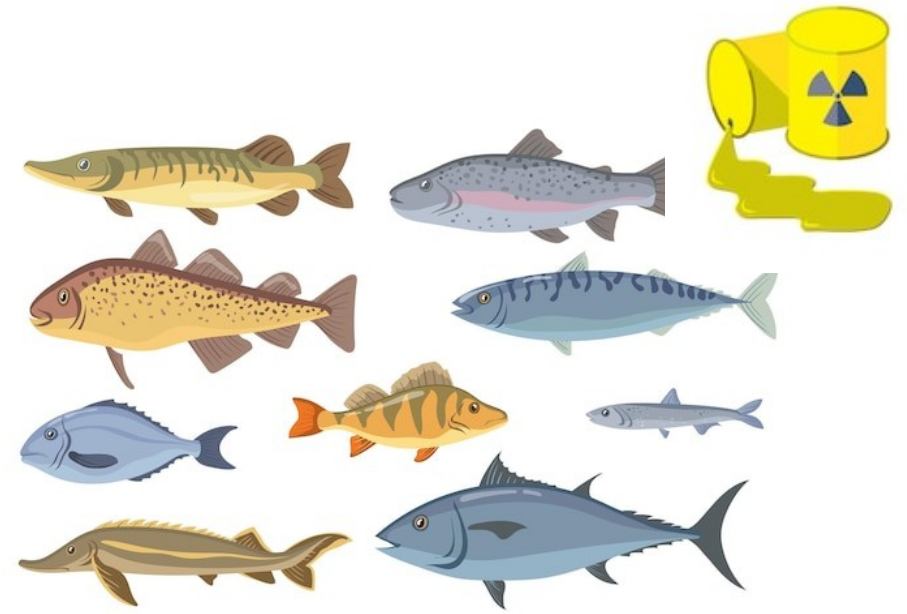
Statistics is scary

Statistics, the science of collecting, analyzing, presenting, and interpreting data.





- Age
- Gender
- Height
- Weight
- Education
- Frequency of fish consumption
- Source of fish
- Chronical disease
- Last medical examination



VARIABLES

characteristics of something or someone

CASES

something or someone

CASES



VARIABLES

Age
Gender
Height
Weight
Education
Frequency of fish consumption
Source of fish
Chronical disease
Last medical examination

CASES



VARIABLES

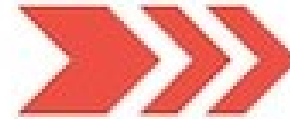
Population

Number of sailors

Weight of caught fish

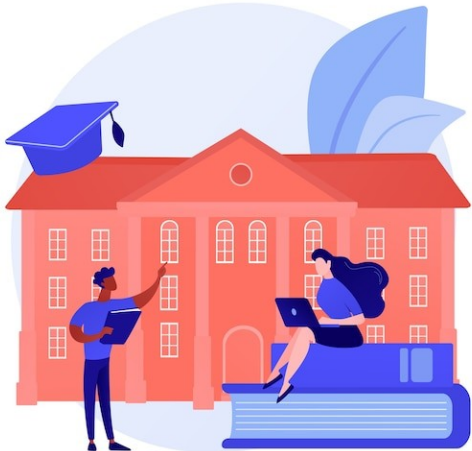
...

CASES



VARIABLES

Need to vary!





Age

6 34 56 12 34 54 47 76 2 21

VARIABLE

No variation!

Village

White Bridge

White Bridge

White Bridge

White Bridge

CONSTANT

VARIABLES

**Categorical
(qualitative)**

**Numerical
(quantitative)**

Nominal

Categories are
unordered

Gender
(male/female)
Marital status
Residence

Ordinal

Categories are
ordered

Disease stage
(mild/moderate/s
evere)
Opinion (strongly
agree...)

Discrete

Integer values

Number of
children
(1/2/3/4/...)
Not 1.222

Continuous

No limitation
on values

Weight in kg
(34.456...)
Height in cm
Speed

VARIABLES

**Categorical
(qualitative)**

**Numerical
(quantitative)**

Nominal

Categories are
unordered

Ordinal

Categories are
ordered

Discrete

Integer values

Continuous

No limitation
on values

Gender
Source of fish
Chronical disease

Education

Frequency of fish
consumption

Last medical
examination

Age
Height
Weight

VALUES

VARIABLES

CASES

DATA MATRIX	Age	Gender	Height	Weight	Education	Freq.of fish consumption (times/month)	Source of fish (majority)	Chronical disease	Last med.exam. (years ago)
Person 1	45	Male	170.4	76.4	PhD	21	Market	Diabetes	10
Person 2	26	Male	163.3	65.3	BS	5	Self-fishing	-	-
Person 3	54	Female	163.6	75.3	MS	23	Grocery store	Bladder infection	10
Person 4	65	Male	156.6	44.2	MS	10	Market	Diabetes	5
Person 5	21	Female	170.1	69.9	HS	8	Market	-	-
Person 6	44	Male	176.4	84.3	MS	43	Self-fishing	-	2
...									
Person 400	6	Male	121.2	21.9	-	15	OBSERVATION		

Frequency Table

shows how the values (observations) are distributed over the cases

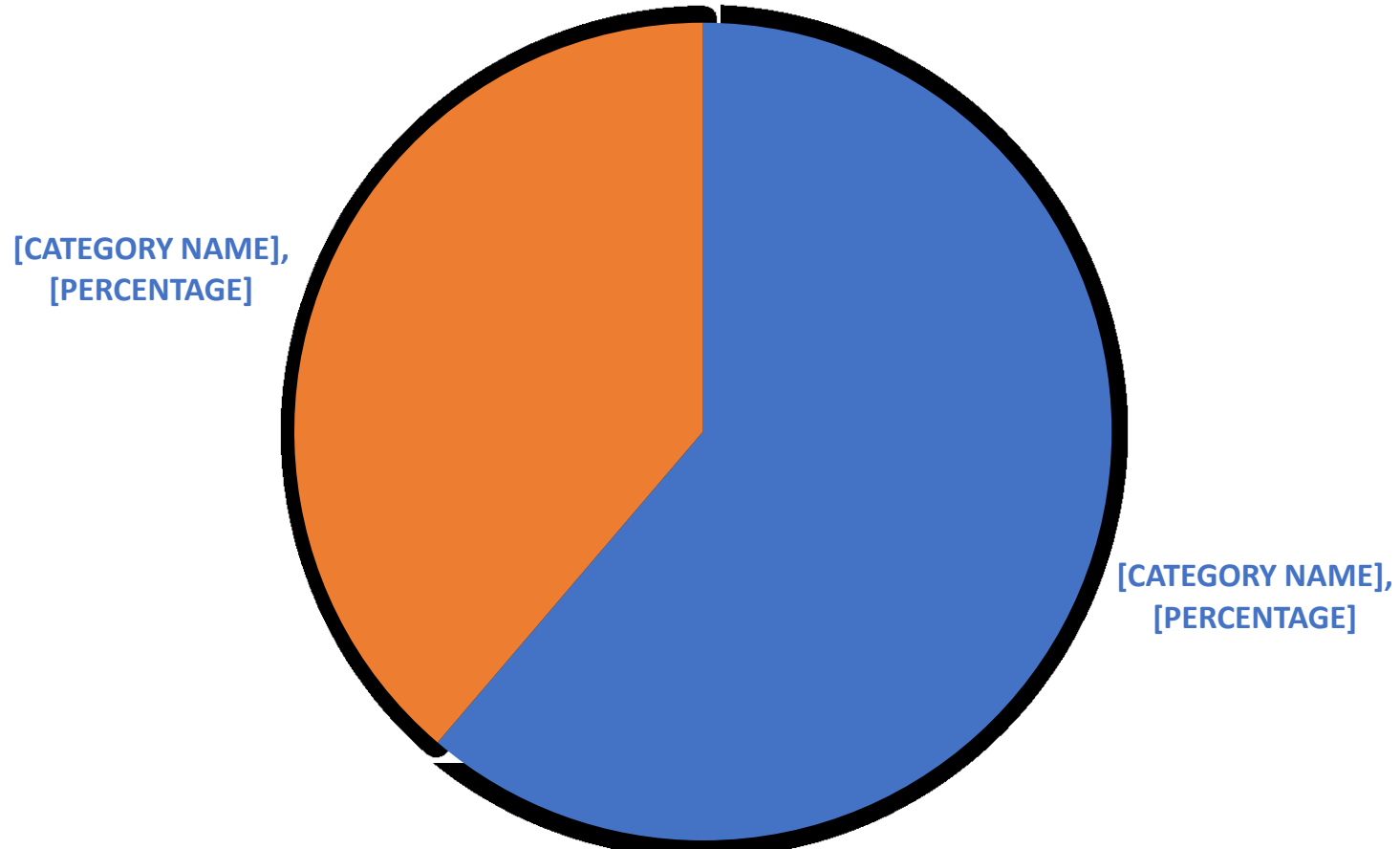
Gender	Frequency	Percentage
Male	245	61
Female	155	39
TOTAL	400	100

Weight	Frequency	Percentage
less than 60	56	14
60-69.9	123	31
70-79.9	169	42
80-89.9	43	11
90 and more	9	2
TOTAL	400	100

Gender	Frequency	Percentage
Male	245	61
Female	155	39
TOTAL	400	100

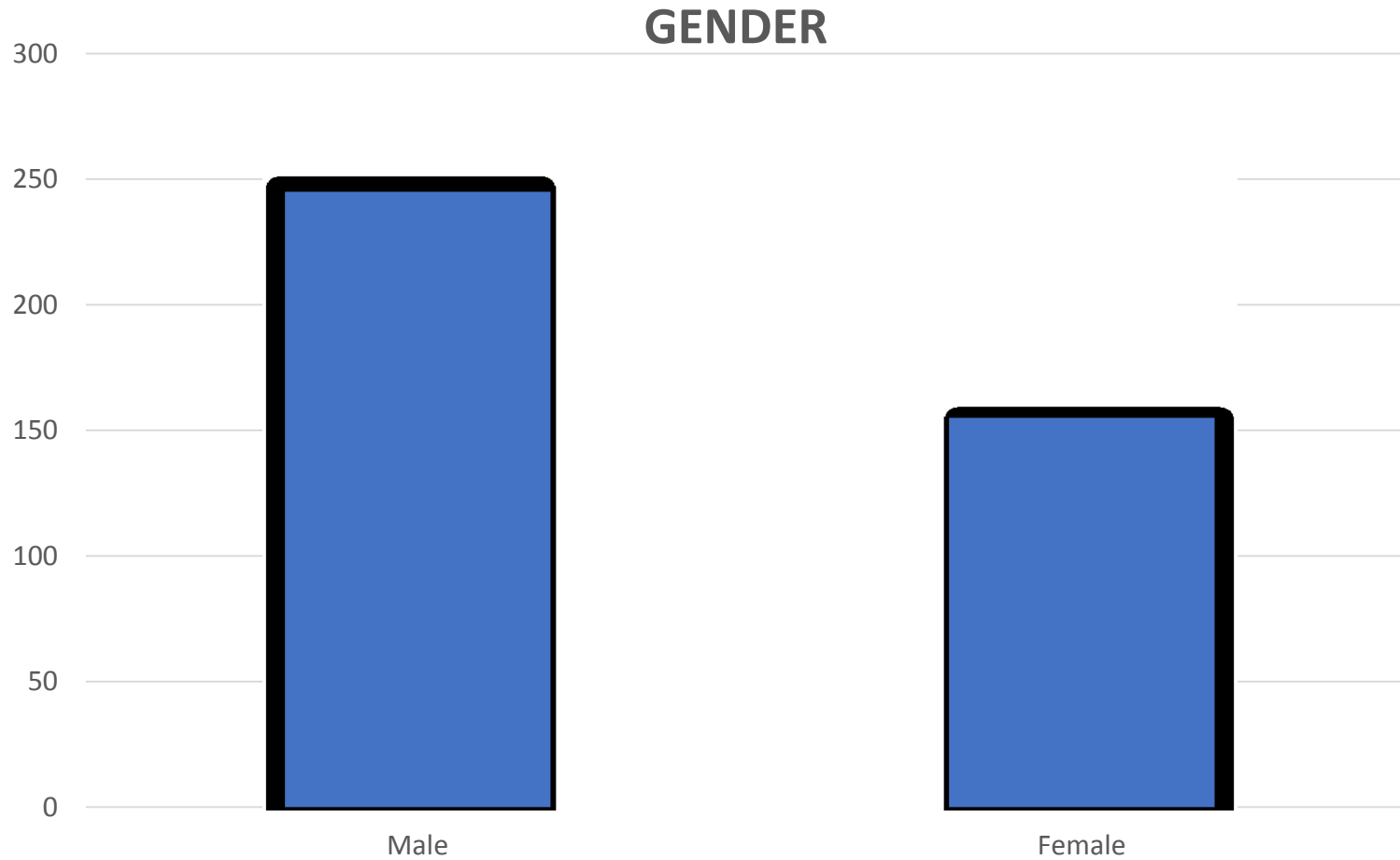
Pie Chart

GENDER

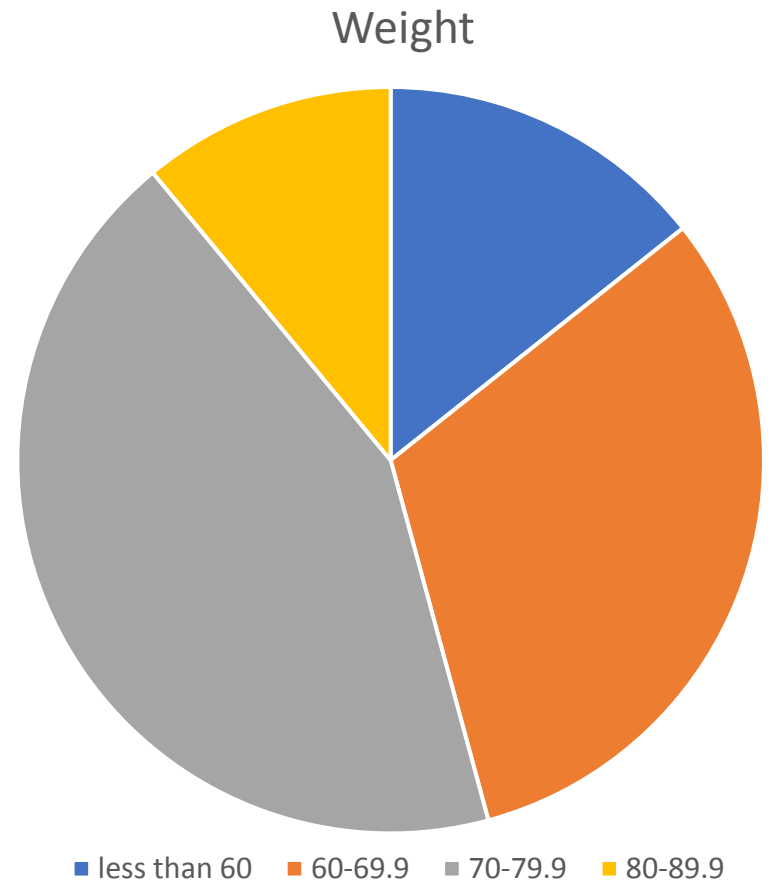
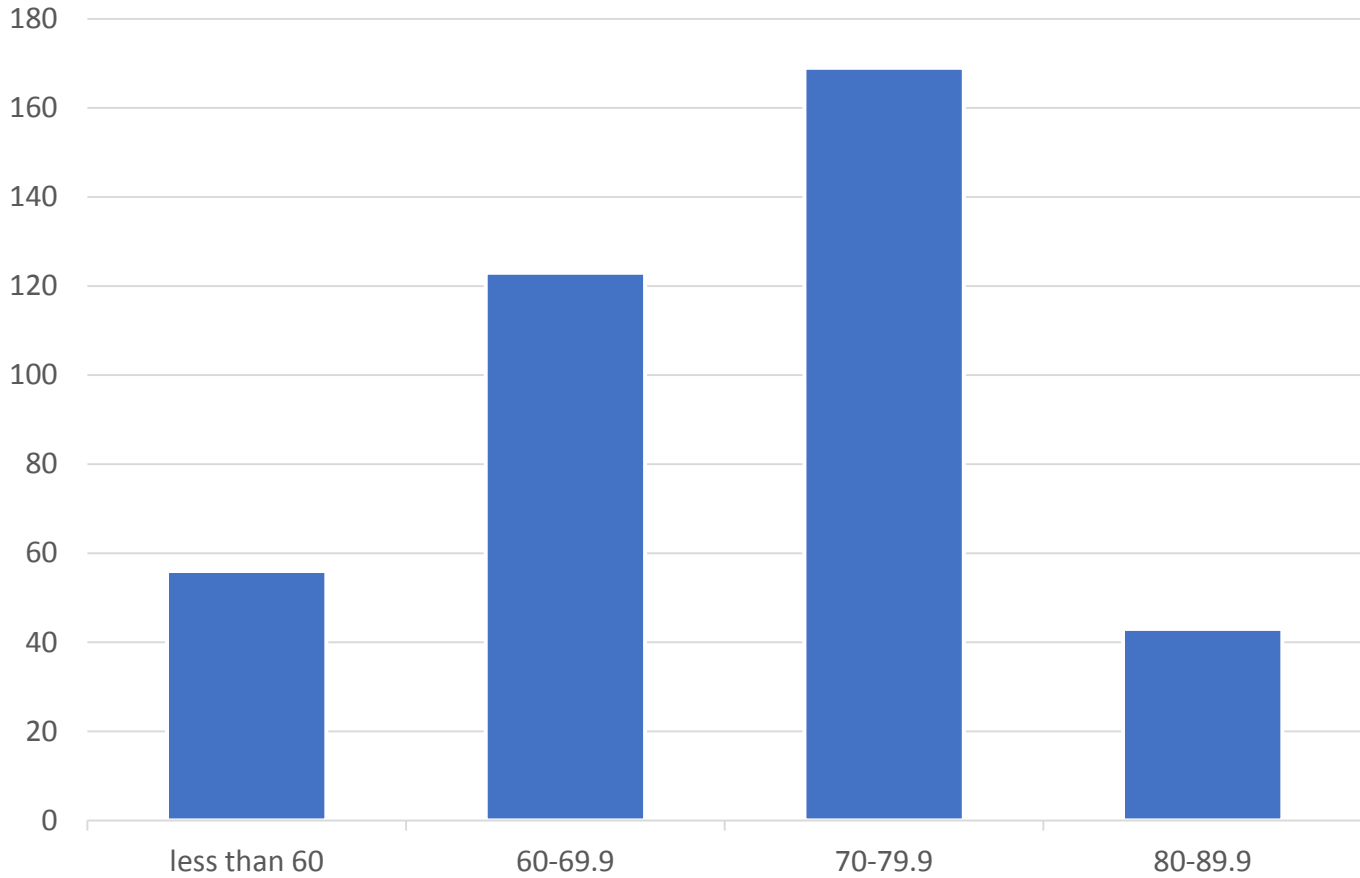


Gender	Frequency	Percentage
Male	245	61
Female	155	39
TOTAL	400	100

Bar Graph

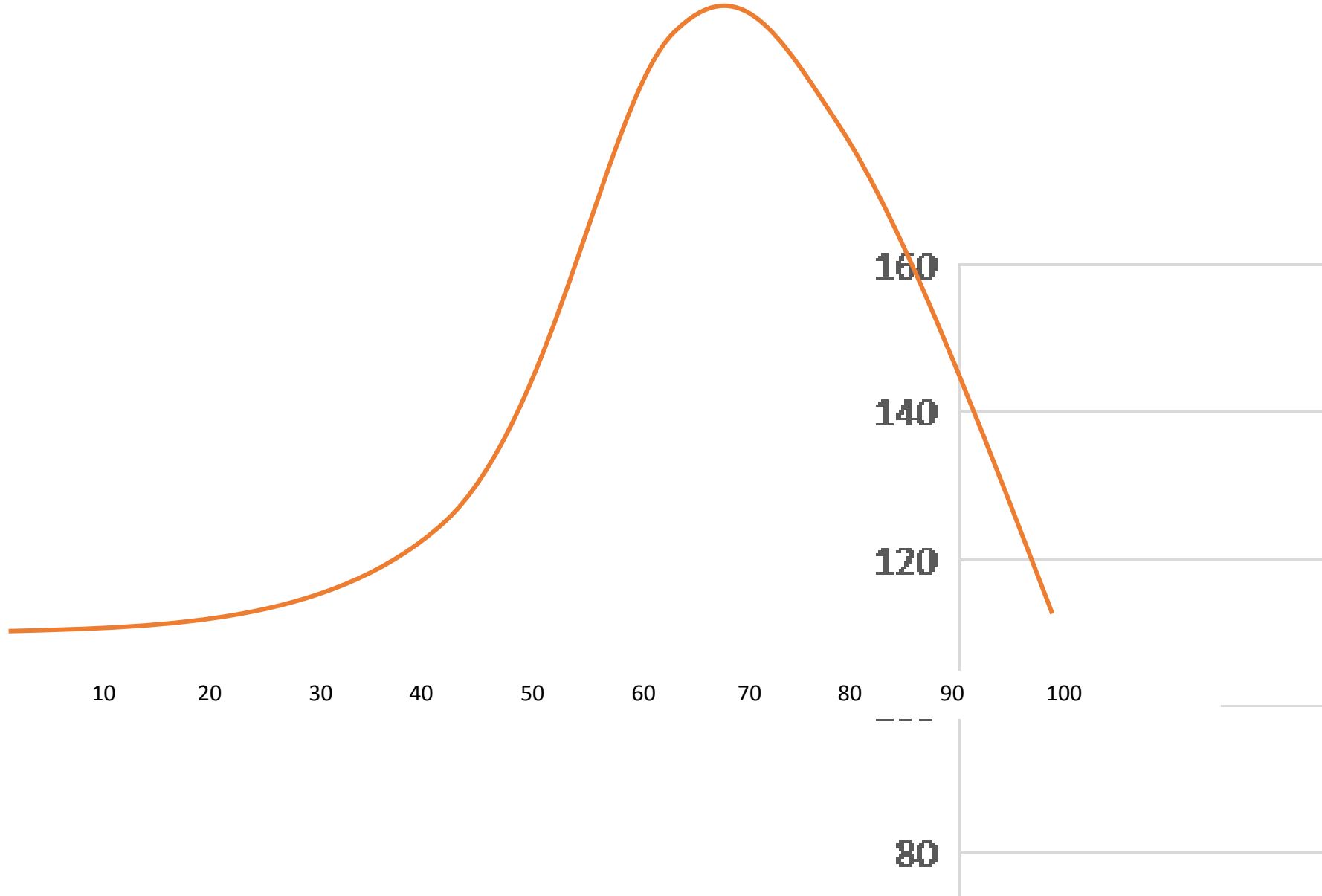


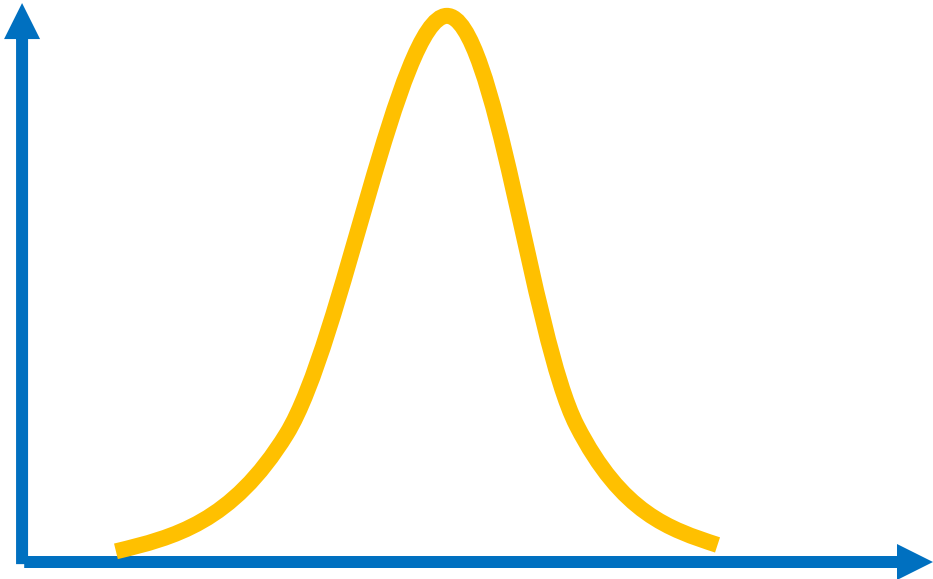
Weight	Frequency	Percentage
less than 60	56	14
60-69.9	123	31
70-79.9	169	42
80-89.9	43	11
90 and more	9	2
TOTAL	400	100



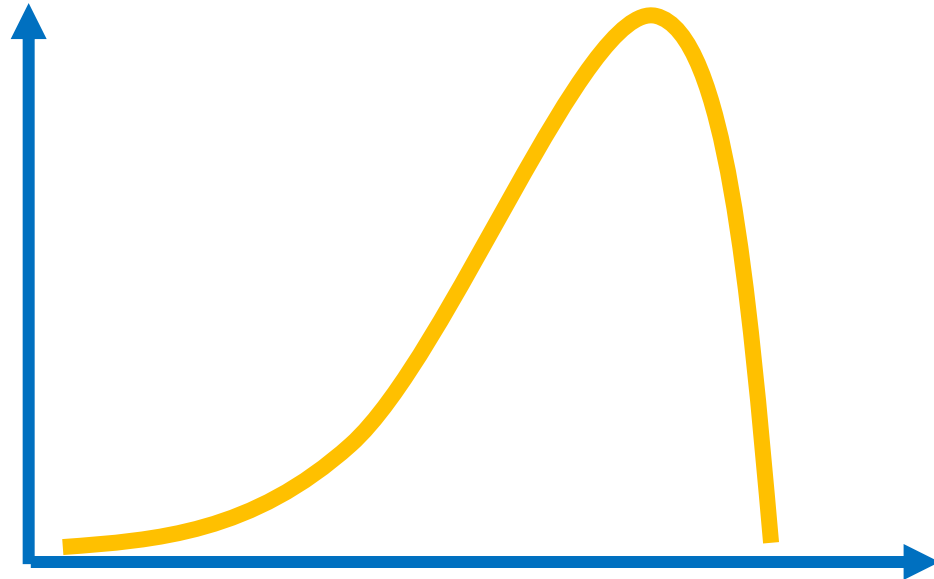
Participants	Weight
Person 1	76.4
Person 2	65.3
Person 3	75.3
Person 4	44.2
Person 5	69.9
Person 6	84.3
...	
Person 400	21.9

Histogram

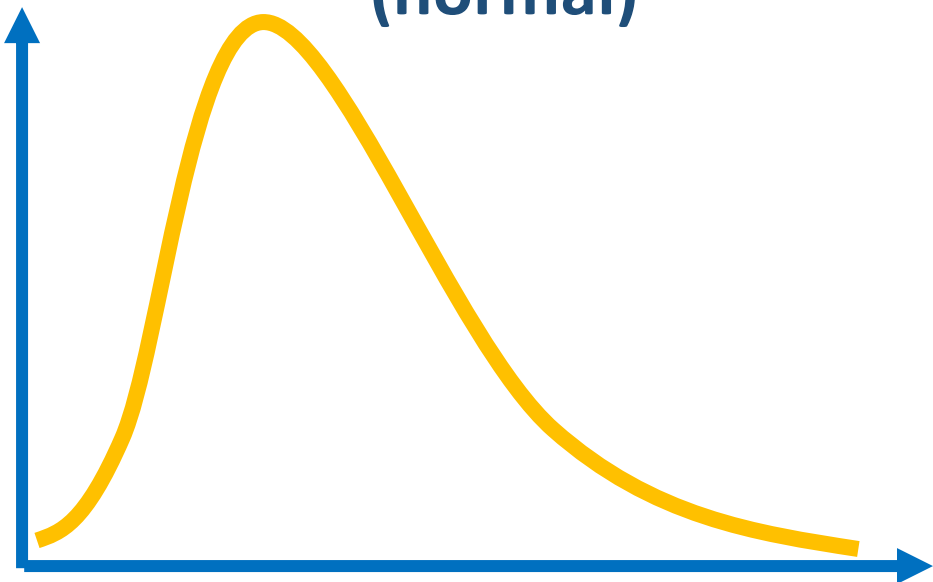




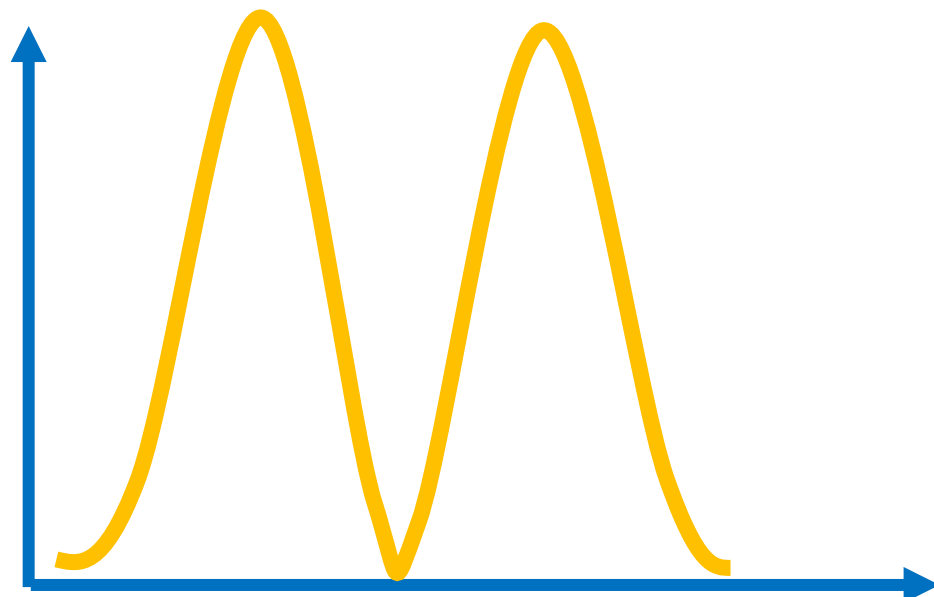
**Bell shaped distribution
(normal)**



Left skewed distribution



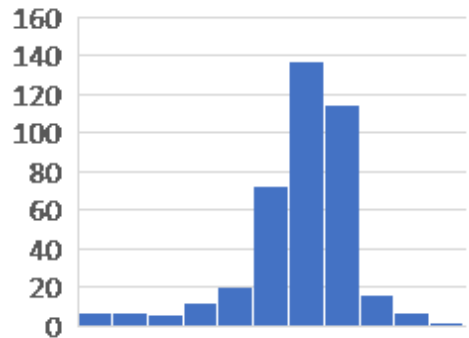
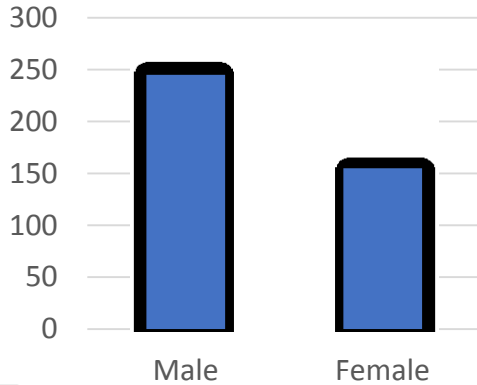
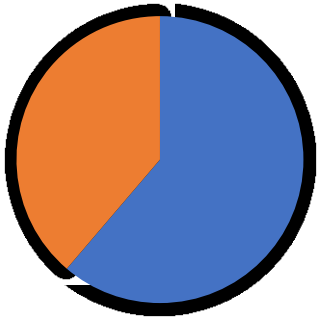
Right skewed distribution



**Two peaks distribution
(bimodal)**

Summarizing a distribution

Graph



Center

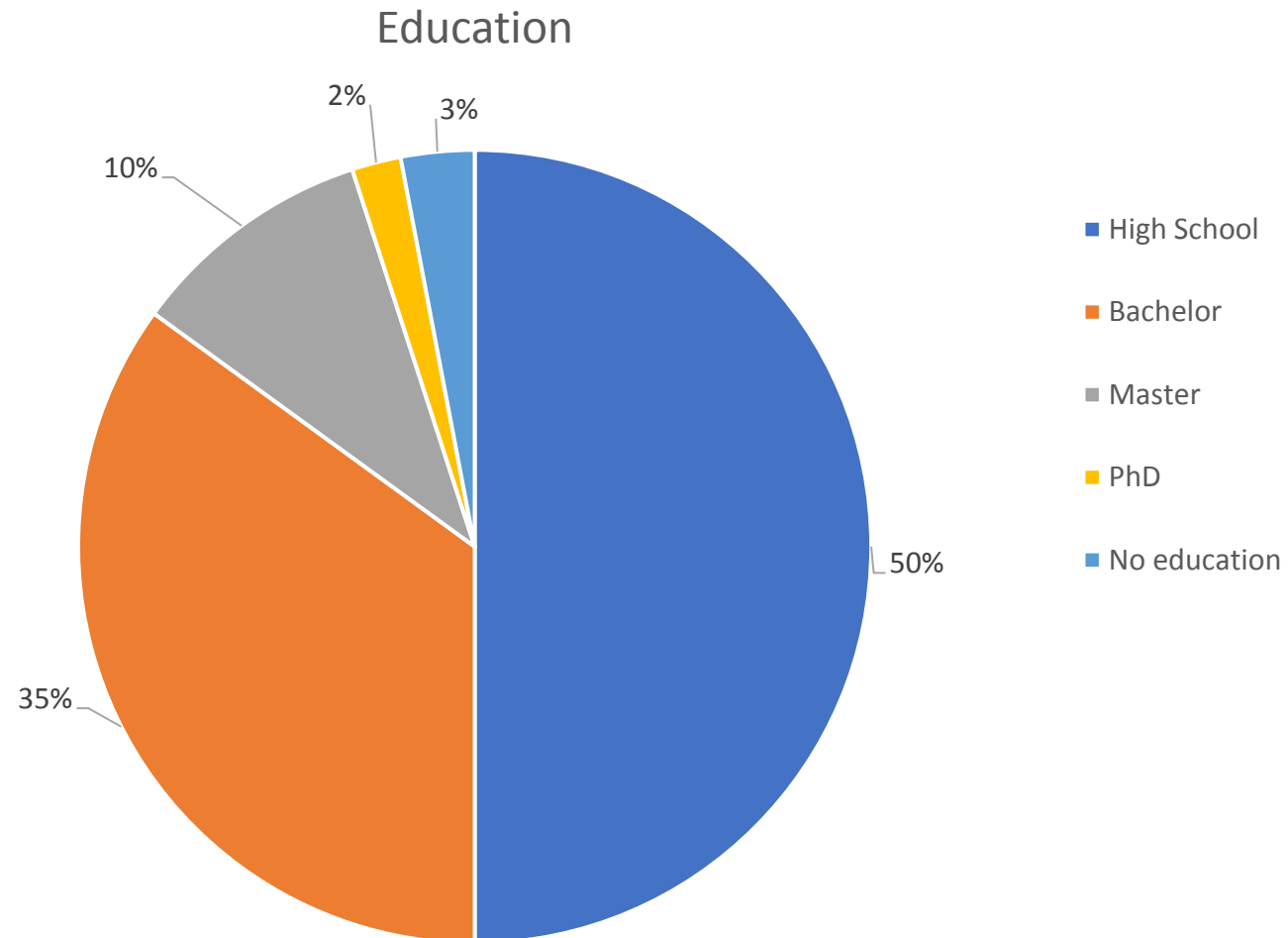
Mode

Median

Mean

Mode

value that occurs most frequently
(the most common outcome)



Participants	Age
Person 1	18
Person 2	56
Person 3	30
Person 4	34
Person 5	31
Person 6	23
...	
Person 55	7

Mode

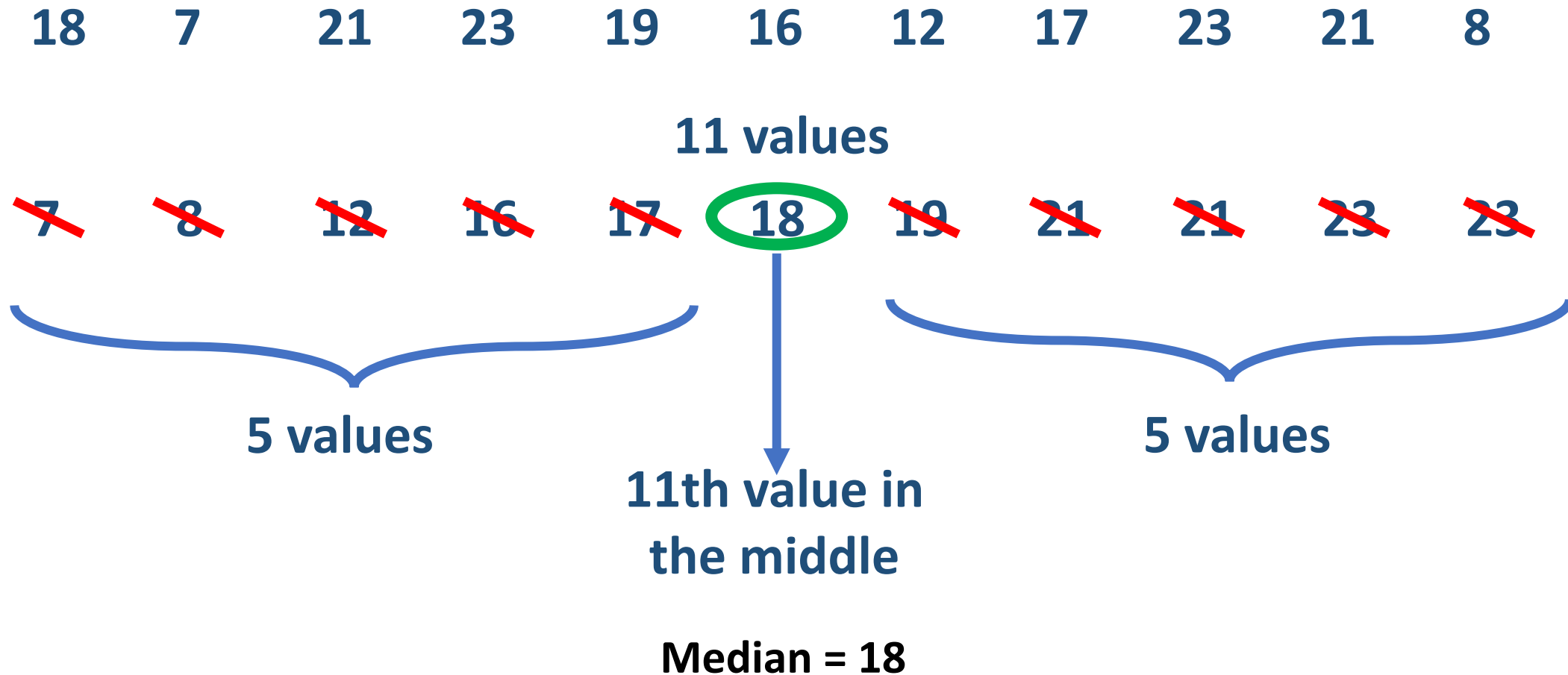
value that occurs most frequently
(the most common outcome)

34	56	34	76	23	21	18	54	32
34	23	48	47	32	21	17	34	45
66	8	25	31	64	29	38	19	66
52	34	51	23	44	42	47	16	43
34	44	12	58	63	35	34	33	34
60	30	34	35	7	23	32	47	54
31								

Mode = 34

Median

the middle value in a set of data when they are ordered from the smallest to the largest



Median

55 values
 $55/2 = 27.5$

55th value in the middle is the median

27 values

7	8	12	16	17	18	19	21	21	23	23	23
23	25	29	30	31	31	32	32	32	33	34	34
34	34	34	34	34	34	34	35	35	38	42	43
44	44	45	47	47	47	48	51	52	54	54	56
58	60	63	64	66	66	76	27 values				

Median

12 values

~~7~~ ~~8~~ ~~12~~ ~~16~~ ~~17~~ ~~18~~ ~~19~~ ~~21~~ ~~21~~ ~~23~~ ~~23~~ ~~23~~



6 values

6 values

~~7~~ ~~8~~ ~~12~~ ~~16~~ ~~17~~ 18 19 ~~21~~ ~~21~~ ~~23~~ ~~23~~ ~~23~~



5 values

5 values

$(18+19)/2=18.5$

Median = 18.5

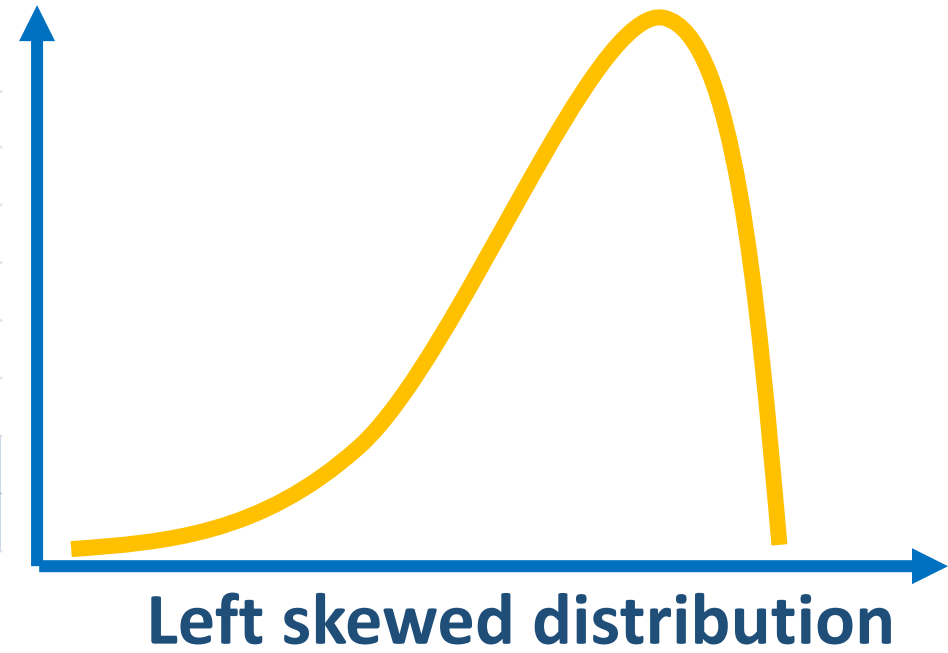
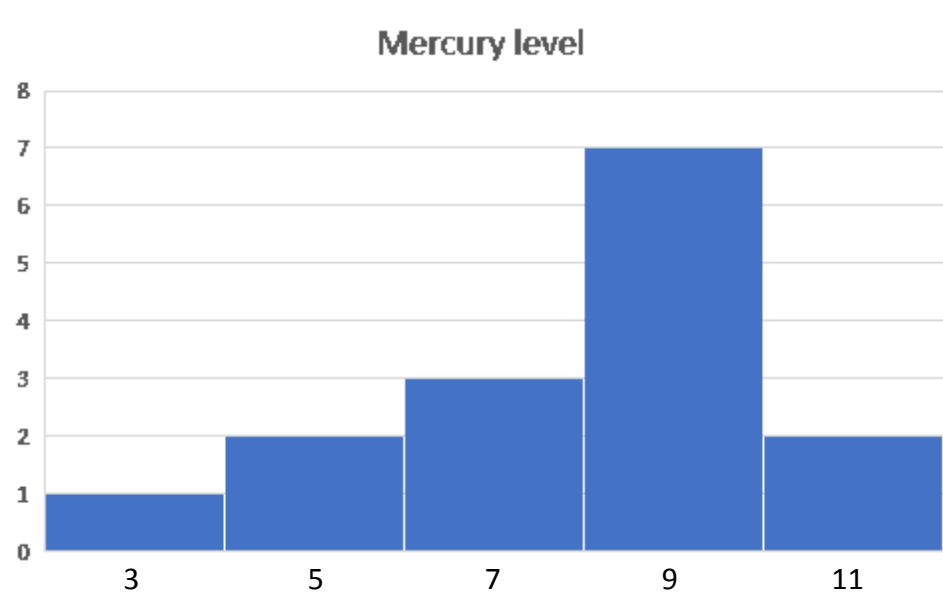
Mean

the sum of all the values divided by the number of values

18 7 21 23 19 16 12 17 23 21 8

$$\text{Mean} = (18+7+21+23+19+16+12+17+23+21+8) / 11 = 16.81$$

Participant	Mercury ($\mu\text{g/L}$)
Person 1	6.454963
Person 2	3.373376
Person 3	10.25793
Person 4	8.003598
Person 5	10.5467
Person 6	9.079663
Person 7	6.5196
Person 8	5.62672
Person 9	8.870464
Person 10	8.57557
Person 11	8.77484
Person 12	8.241835
Person 13	9.019997
Person 14	7.720283
Person 15	6.121072



Median = 8.241835

3.373376 5.62672 6.121072 6.454963 6.5196 7.720283 8.003598 **8.241835**
8.57557 8.77484 8.870464 9.019997 9.079663 10.257931 10.5467

No Mode

Mean = 7.8124408

$(3.373376+5.62672+6.121072+6.454963+6.5196+7.720283+8.003598+8.241835+8.57557+8.77484+8.870464+9.019997+9.079663+10.257931+10.5467)/15$

Participant	Mercury ($\mu\text{g/L}$)
-------------	-----------------------------

Person 1	6.454963
Person 2	3.373376
Person 3	10.25793
Person 4	8.003598
Person 5	10.5467
Person 6	9.079663
Person 7	6.5196
Person 8	5.62672
Person 9	8.870464
Person 10	8.57557
Person 11	8.77484
Person 12	8.241835
Person 13	9.019997
Person 14	7.720283
Person 15	154.121072

Mercury level



Outlier

Median = 8.57557

3.373376 5.62672 6.454963 6.5196 7.720283 8.003598 8.241835 **8.57557** 8.77484
 8.870464 9.019997 9.079663 10.257931 10.5467 154.121072

No Mode

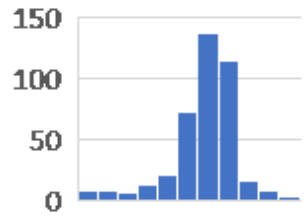
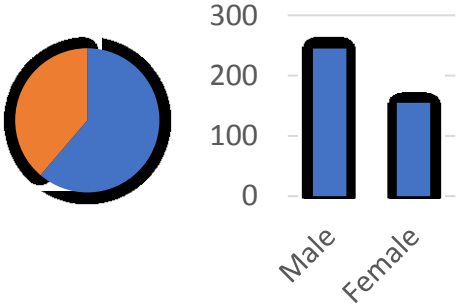
Mean = 17.6791

$(3.373376+5.62672+6.454963+6.5196+7.720283+8.003598+8.241835+8.57557+8.77484+8.870464+9.019997+9.079663+10.257931+10.5467+ 154.121072)/15$

Summarizing a distribution

Descriptive statistics

Graph



Center

Mode

Median

Mean

Spread

Range

Standard deviation

Quartiles

Range

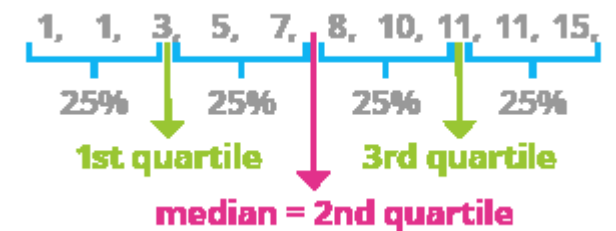
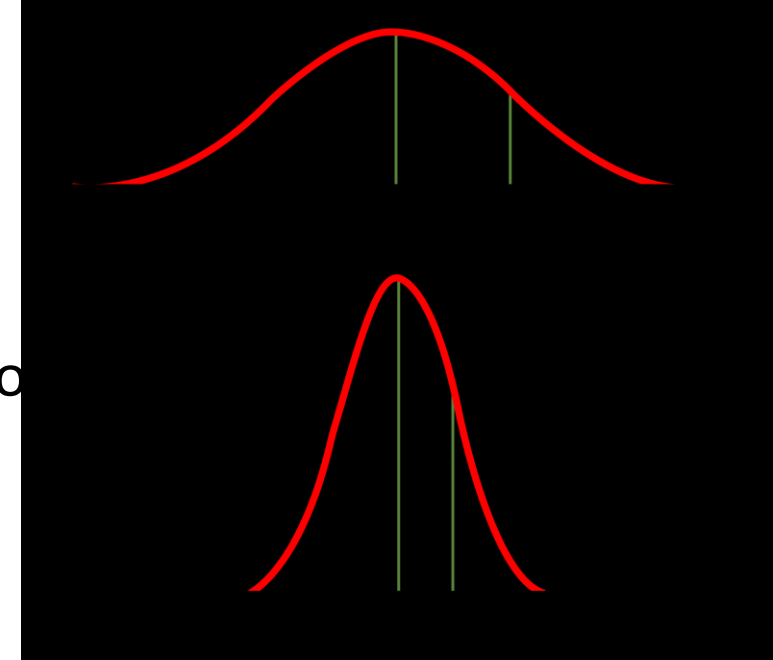
difference between the lowest and highest value

Standard deviation

a summary measure of the differences of each observation from the mean

Quartiles

the first quartile at 25% (Q1), the second quartile at 50% (Q2 or median) and the third quartile at 75% (Q3)



Task!

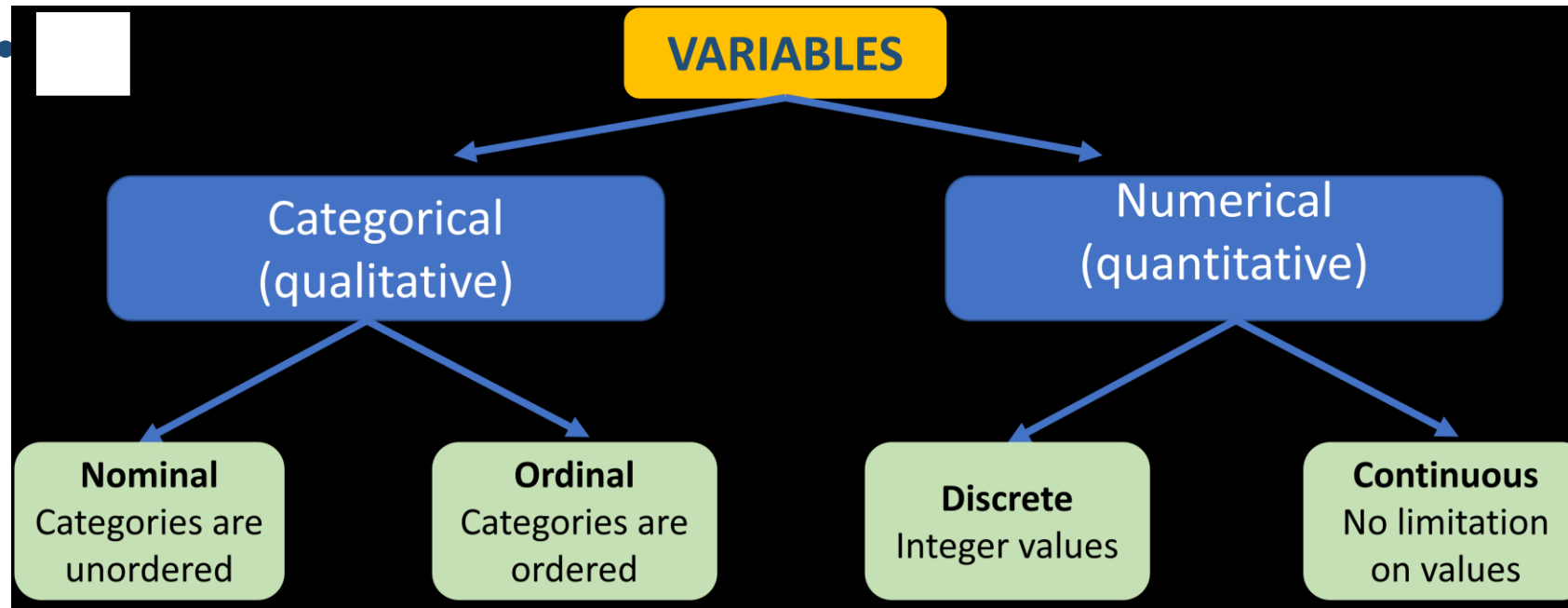


Summary

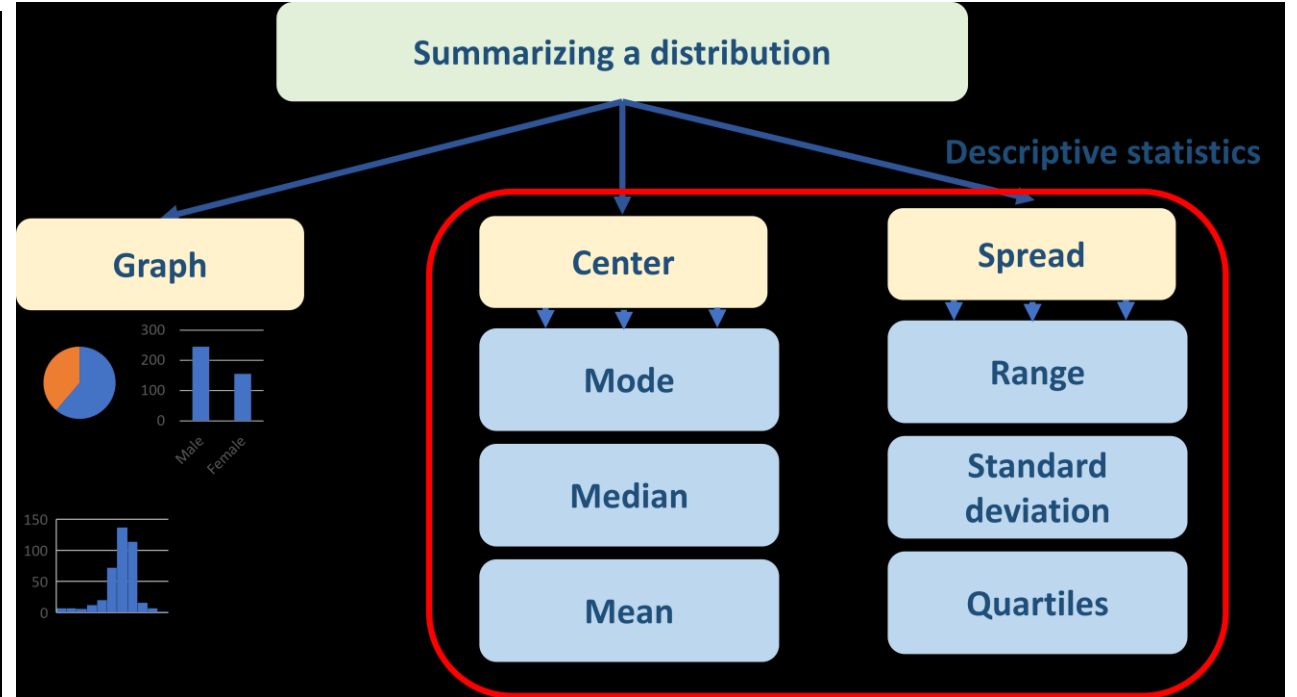
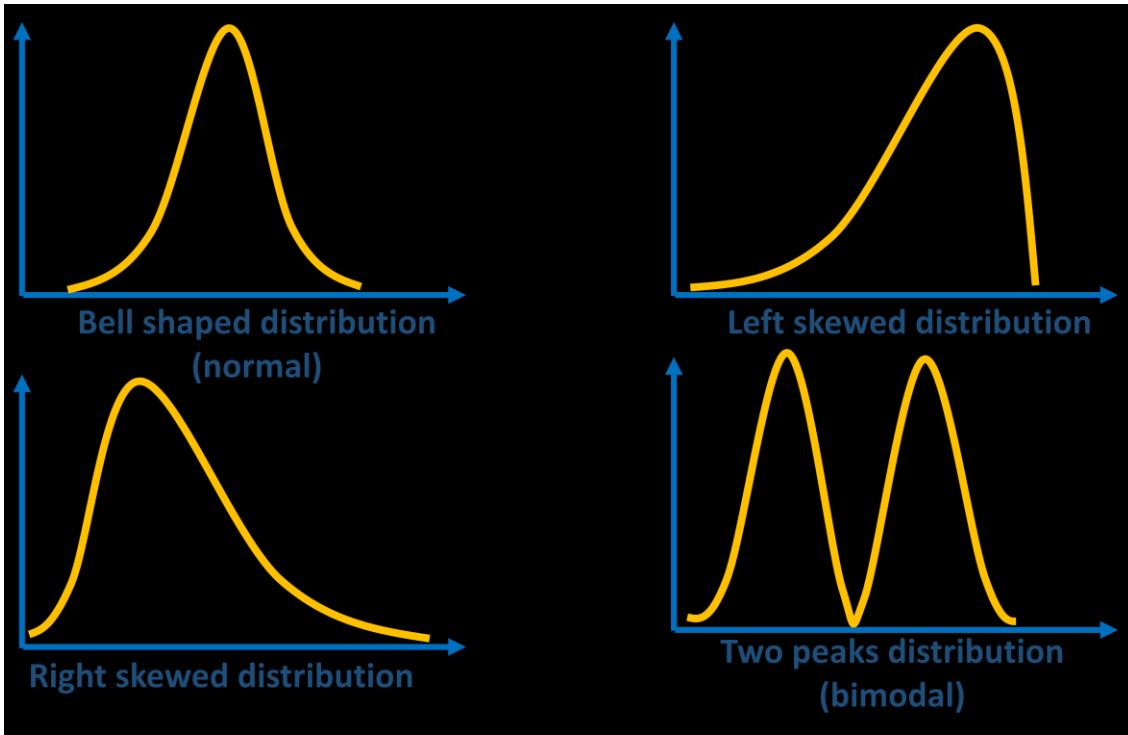
- Statistics isn't scary.
- **VARIABLES** characteristics of something or someone

CASES

something or someone

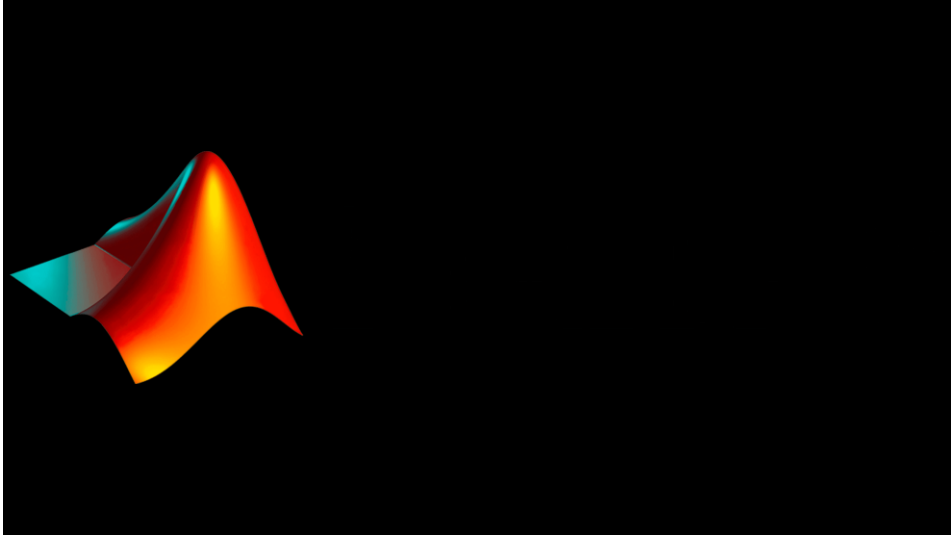


Summary

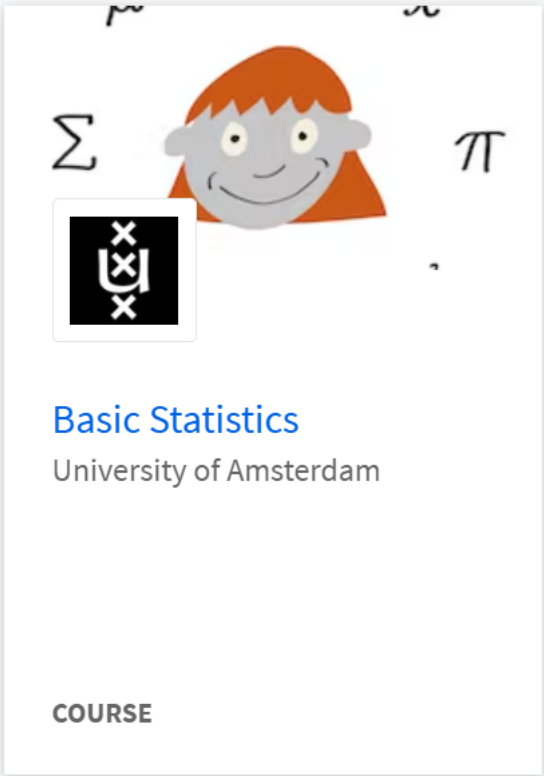


- Pay attention to outliers!
- In many cases the median is more appropriate because it isn't influenced by extremely large values.

Practical part



Recently Viewed Courses

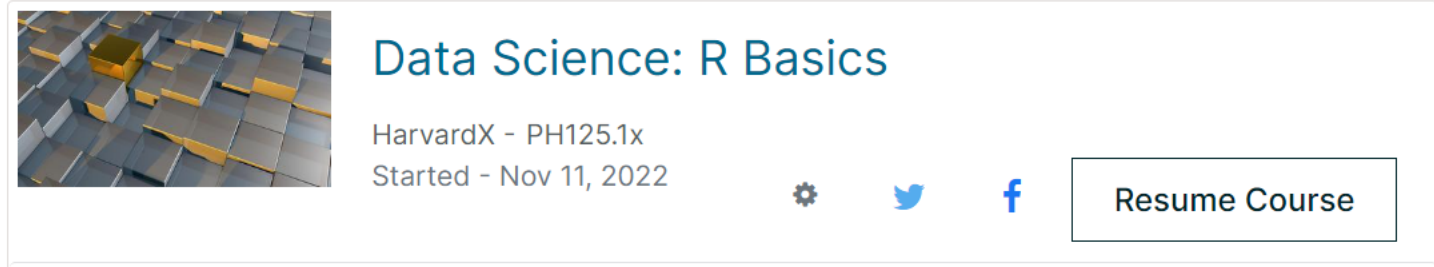


A course card for 'Basic Statistics' by the University of Amsterdam. The card features a stylized orange-haired character, mathematical symbols like Σ and π , and the edX logo. The text 'Basic Statistics' and 'University of Amsterdam' is visible, along with a 'COURSE' label at the bottom.

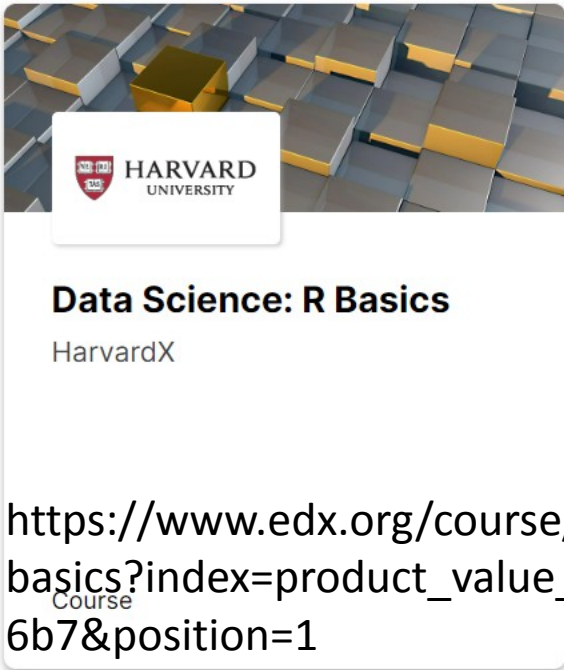
<https://www.coursera.org/learn/basic-statistics>

Add a recovery email to retain access when single-sign on is not available. Go to [your Account Settings](#).

My Courses

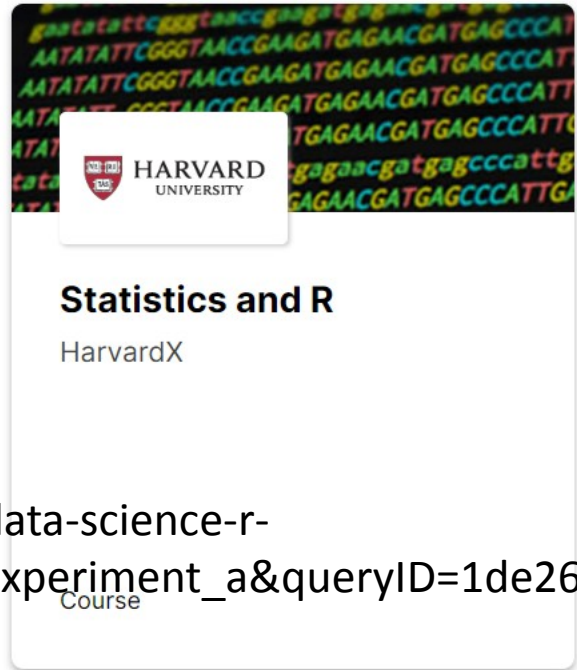


A course card for 'Data Science: R Basics' by HarvardX. It includes a 3D bar chart image, the course title, 'HarvardX - PH125.1x', and 'Started - Nov 11, 2022'. There are icons for settings, Twitter, and Facebook, and a 'Resume Course' button.



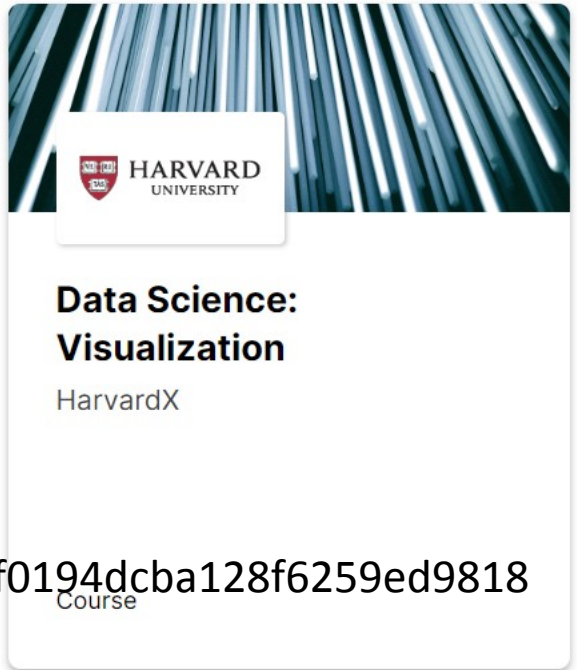
A course card for 'Data Science: R Basics' by HarvardX. It features a 3D bar chart image, the Harvard University logo, the course title, and 'HarvardX'.

https://www.edx.org/course/data-science-r-basics?index=product_value_experiment_a&queryID=1de26ef0194dcba128f6259ed98186b7&position=1



A course card for 'Statistics and R' by HarvardX. It features a DNA sequence image, the Harvard University logo, the course title, and 'HarvardX'.

[https://www.edx.org/course/statistics-and-r](#)



A course card for 'Data Science: Visualization' by HarvardX. It features a blue abstract image, the Harvard University logo, the course title, and 'HarvardX'.

[https://www.edx.org/course/data-science-visualization](#)



WE'RE HIRING

Explore



Sign In

Get Started

Build data skills online

Data drives everything. Get the skills you need for the future of work.

Start Learning For Free

DataCamp For Business

Create Your Free Account

Google

LinkedIn

Facebook

or

Email Address

Email address

Password

Password



Start Learning For Free

By continuing, you accept our Terms of Use, our Privacy Policy and that your data is stored in the USA.

📄 Exercise <

Back to Apples and Oranges

Common sense tells you not to add apples and oranges. The `my_apples` and `my_oranges` variables both contained a number in the previous exercise. The `+` operator works with numeric variables in R. If you really tried to add `"apples"` and `"oranges"`, and assigned a text value to the variable `my_oranges` but not to `my_apples` (see the editor), you would be trying to assign the addition of a numeric and a character variable to the variable `my_fruit`. This is not possible.

🕒 Instructions 100 XP

- Click 'Submit Answer' and read the error message. Make sure to understand why this did not work.
- Adjust the code so that R knows you have 6 oranges and thus a fruit basket with 11 pieces of fruit.

script.R

```

1 # Assign a value to the variable called my_apples
2 my_apples <- 5
3
4 # Print out the value of my_apples
5 my_apples
6
7 # Assign a value to the variable my_oranges and print it out
8 my_oranges <- 6
9 my_oranges
10
11 # New variable that contains the total amount of fruit
12 my_fruit <- my_apples + my_oranges
13 my_fruit
    
```

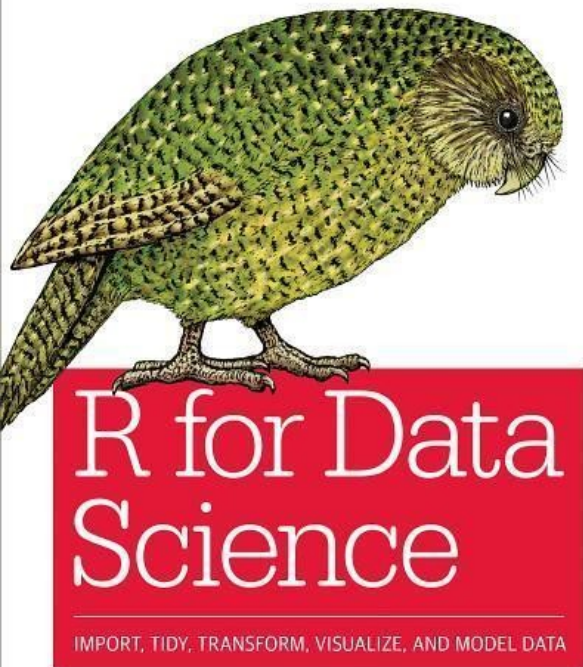


Run Code

Submit Answer

R Console

>



Hadley Wickham &
Garrett Golemund

R for Data Science

Table of contents

Welcome

1 Introduction

Explore

2 Introduction

3 Data visualisation

4 Workflow: basics

5 Data transformation

6 Workflow: scripts

7 Exploratory Data Analysis

8 Workflow: projects

Wrangle

9 Introduction

10 Tibbles

11 Data import

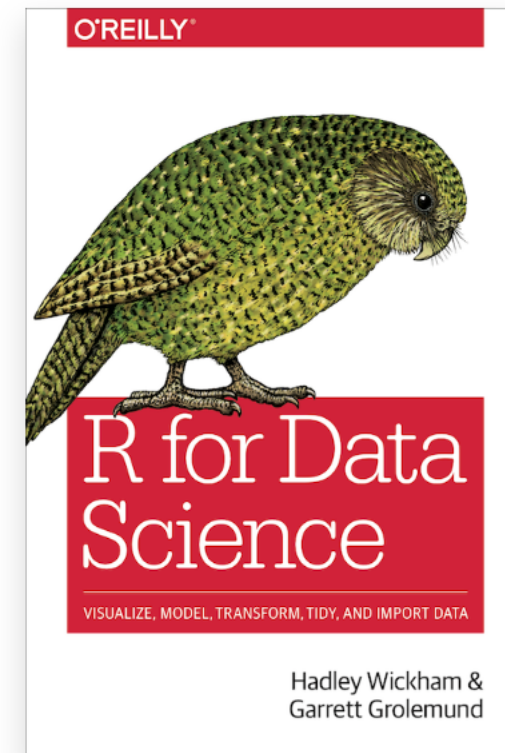
12 Tidy data

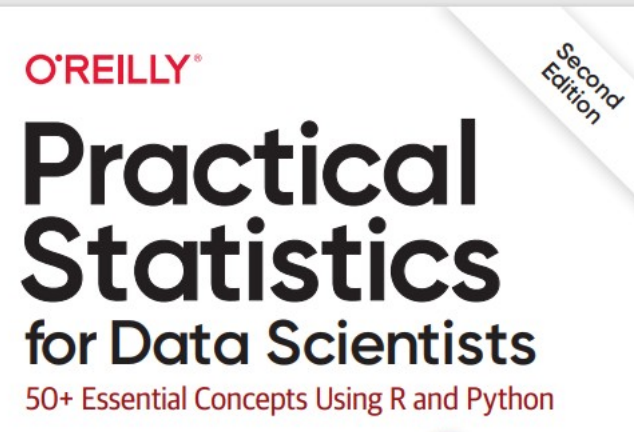
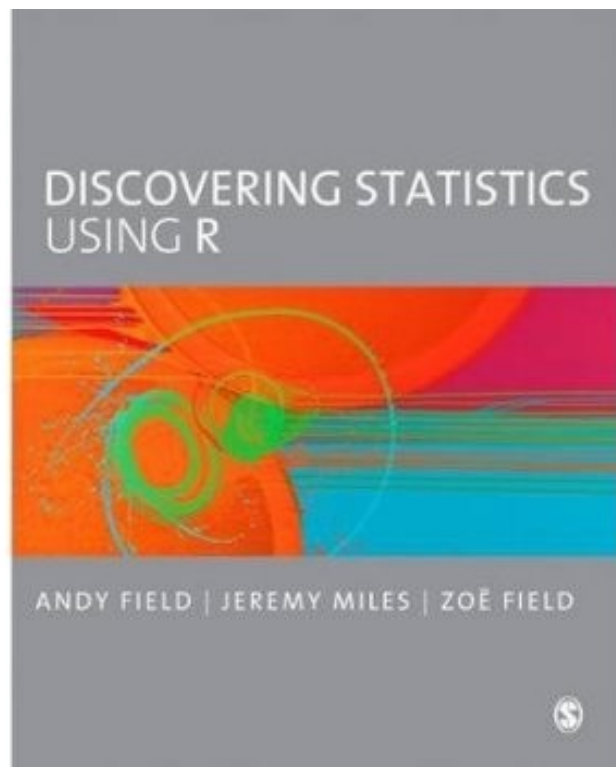
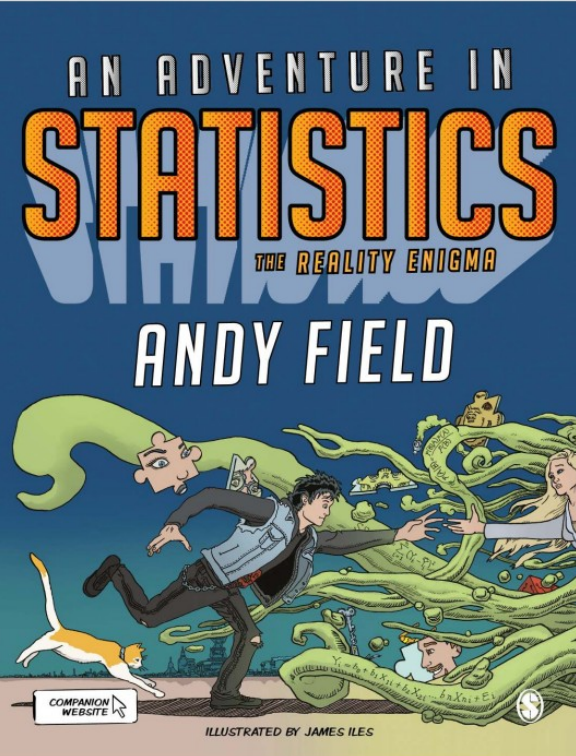
13 Relational data

Welcome

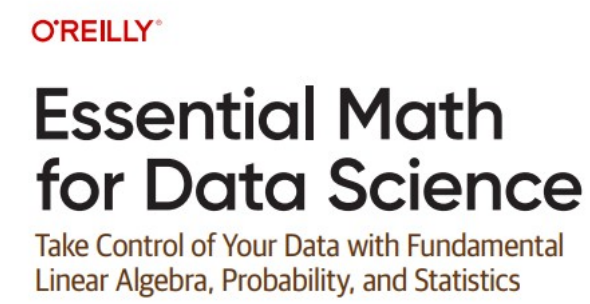
This is the website for “**R for Data Science**”. This book will teach you how to do data science with R: You’ll learn how to get your data into R, get it into the most useful structure, transform it, visualise it and model it. In this book, you will find a practicum of skills for data science. Just as a chemist learns how to clean test tubes and stock a lab, you’ll learn how to clean data and draw plots—and many other things besides. These are the skills that allow data science to happen, and here you will find the best practices for doing each of these things with R. You’ll learn how to use the grammar of graphics, literate programming, and reproducible research to save time. You’ll also learn how to manage cognitive resources to facilitate discoveries when wrangling, visualising, and exploring data.

This website is (and will always be) **free to use**, and is licensed under the Creative





Peter Bruce, Andrew Bruce & Peter Gedeck



Thomas Nield


```
ggplot2.R x
Source on Save
1 library(ggplot2)
2 mpg_plot ← ggplot(mpg, aes(x = displ, y = hwy)) +
3   geom_point(aes(colour = class))
4
5 mpg_plot
6
```

Source

5:9 (Top Level) R Script

```
Console Terminal Background Jobs
R 4.2.0 · ~/rstudio-user-guide/
> library(ggplot2)
> mpg_plot ← ggplot(mpg, aes(x = displ, y = hwy)) +
+   geom_point(aes(colour = class))
>
> mpg_plot
>
```

Console

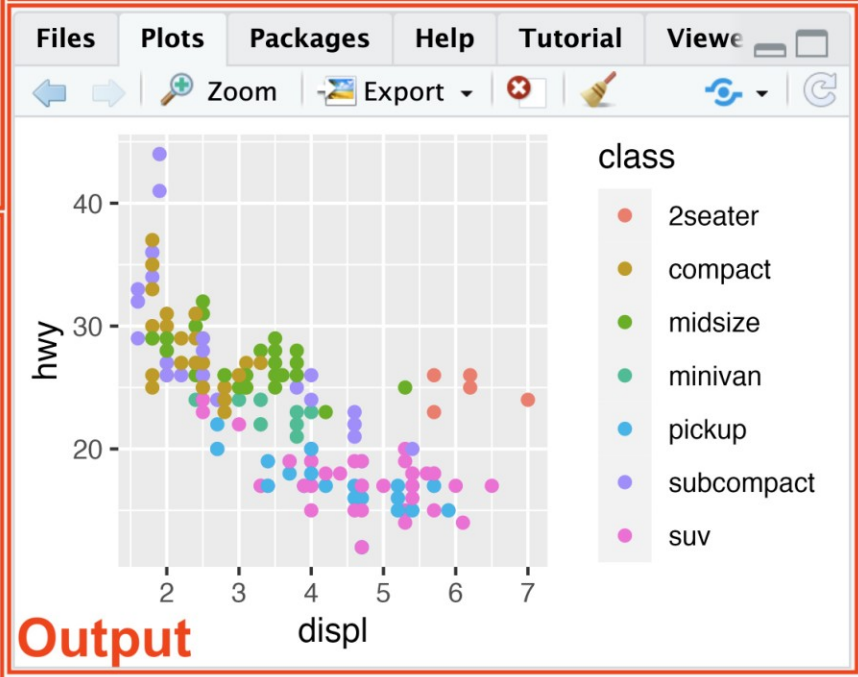
Environment History Connections Build Git

220 MiB Grid

R Global Environment

Name	Type	Len...	Size	Value
mpg_plot	gg	9	29.1...	List of 9

Environments



Output