Data Analytics in R
Session 5

Maria Kunevich

Homework feedback

- Your feedback for the course and course assignments is greatly appreciated!
- *https://miro.com/welcomeonboard/SUw3RHpXWjBpUDF2V0dwTWhkOFVVUnlUTnZ4Qm1oVGtSTVN0SmNCUmJyODhydU9hUzA3VUpNZVZHRnNBenhqVHwzMDc0NDU3MzY2NDE0OTE3Njc4?share_link_id=308470562965*

- Let’s look at our DataCamp leaderboard: https://app.datacamp.com/groups/data-analytics-in-r-db1ae4f4-62a1-4da2-b5d9-94616b38d5d0/leaderboard

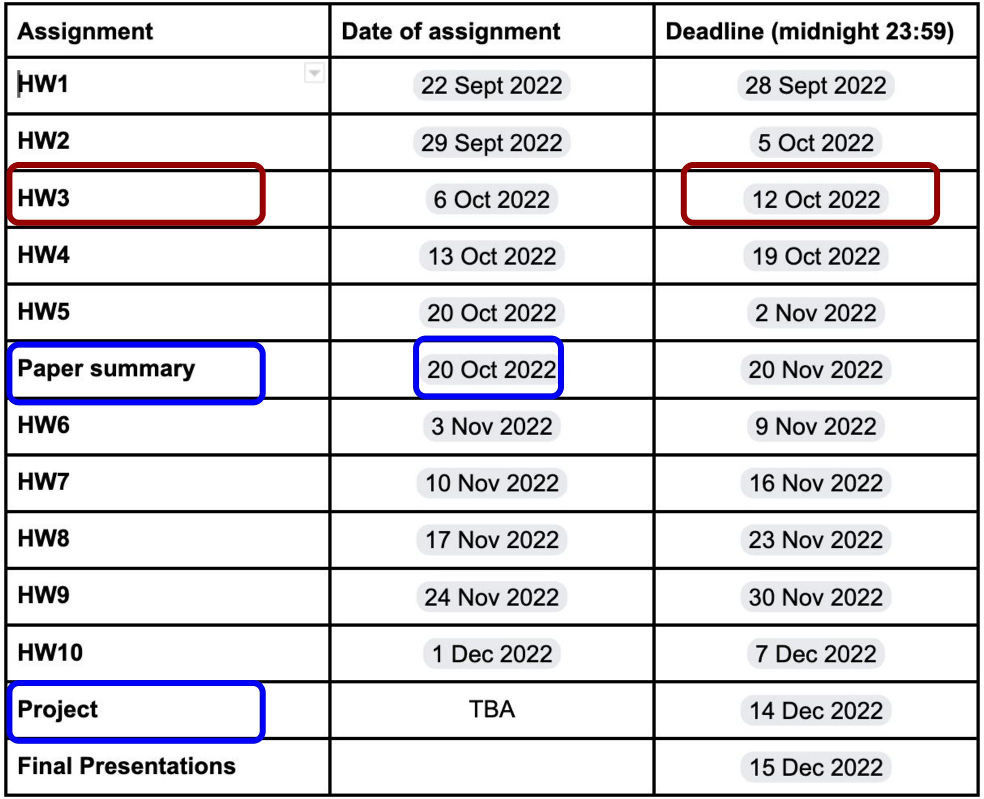
Homework feedback

- Feedback from me is in the comments to your gists: please make sure you **comment** your script thoroughly :)
- Review table: https://docs.google.com/spreadsheets/d/1JyX2fQArbhfkkMf_HTly-FDLDvI-C_-B1DAi8liSfyc/edit?usp=sharing
- Any questions?

Homework assignment 3

- **Part 1:** finish two last chapters on DataCamp - Introduction to R course
- **Part 2:** the same procedure for copy pasting the task (HW3) and creating your own Gist, submit the link on **Wednesday evening** by 23:59
- Please **comment your script** heavily and more carefully, make sure you explain the data set structure and comment on **all the characteristics** of the data set. Provide **descriptive statistics** (with your comments and interpretation!)

Course information: assignments deadlines



Assignment	Date of assignment	Deadline (midnight 23:59)
HW1	22 Sept 2022	28 Sept 2022
HW2	29 Sept 2022	5 Oct 2022
HW3	6 Oct 2022	12 Oct 2022
HW4	13 Oct 2022	19 Oct 2022
HW5	20 Oct 2022	2 Nov 2022
Paper summary	20 Oct 2022	20 Nov 2022
HW6	3 Nov 2022	9 Nov 2022
HW7	10 Nov 2022	16 Nov 2022
HW8	17 Nov 2022	23 Nov 2022
HW9	24 Nov 2022	30 Nov 2022
HW10	1 Dec 2022	7 Dec 2022
Project	TBA	14 Dec 2022
Final Presentations		15 Dec 2022

Plan for today

1. Revision of the basic concepts in statistics: population vs sample, levels of measurement in statistics
2. Basic concepts in descriptive statistical analysis
3. Data import in R
4. Data frames in R

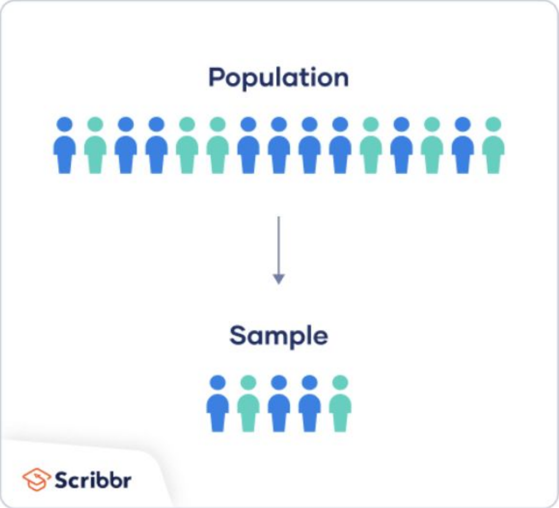
Basic concepts in statistics

❑ Describe the difference between population and sample

A **parameter** is a number that summarises some aspect of the population as a whole. A **statistic** is a number computed from the sample data.

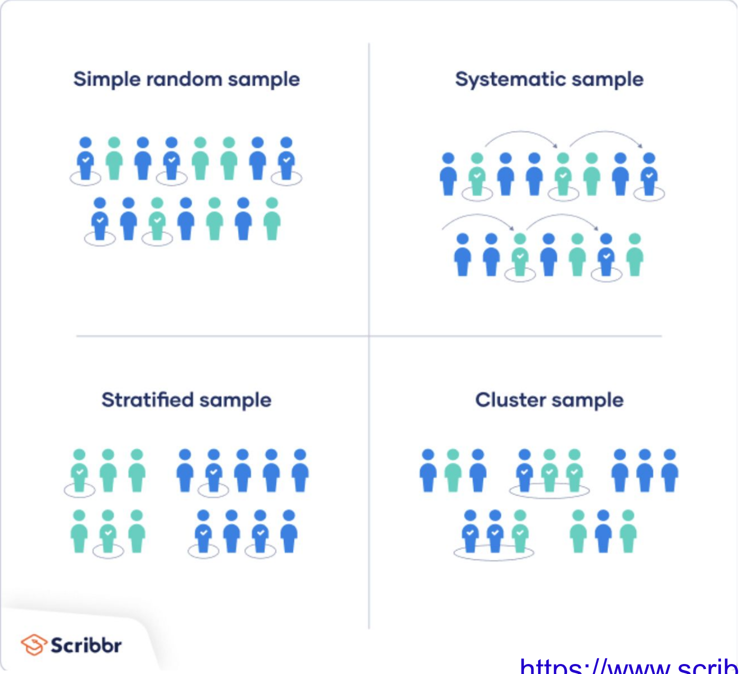
❑ *Example:* how many students know R language and use it for statistical analysis? Is it difficult to learn?

❑ What is the population? What is the sample?



Basic concepts in statistics

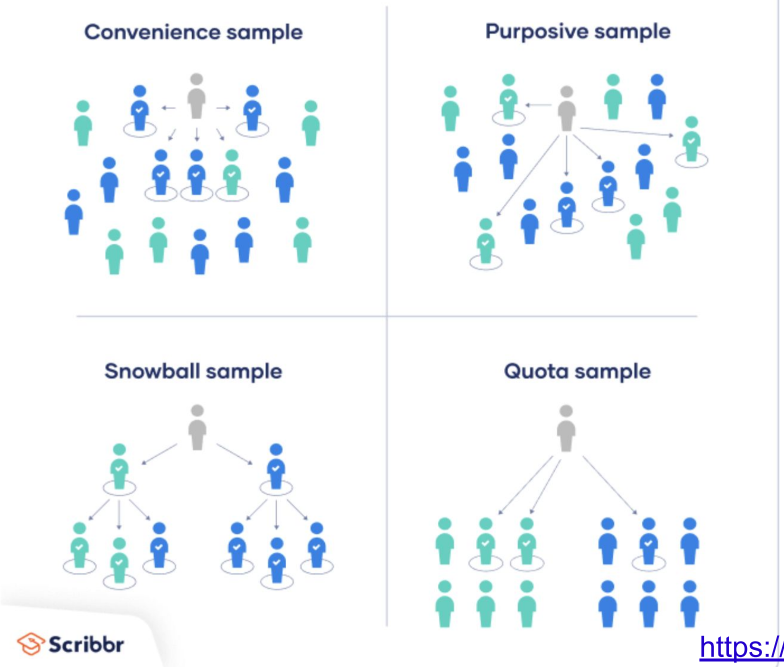
❑ Sampling techniques - **probability** sampling



https://www.scribbr.com/methodology/sampling-methods/

Basic concepts in statistics

❑ Sampling techniques - **non-probability** sampling

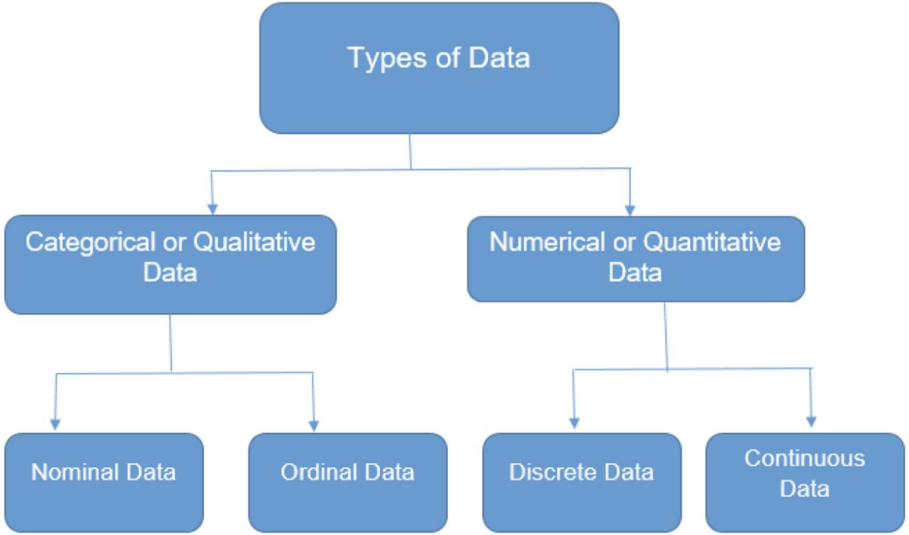


Probability **bias**

https://www.scribbr.com/methodology/sampling-methods/

Basic concepts in statistics

- ❑ Describe the difference and provide examples of **different data types** (levels of measurement) in statistics
- ❑ Gender, age, level of education,
  yes/no answer (0/1),
  level of difficulty (1-7),
  scale (agree-disagree), year
- ❑ In small groups provide your own examples



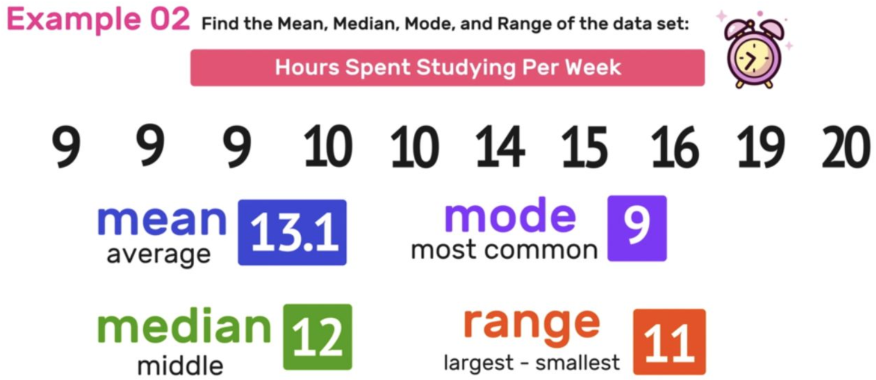
Types of descriptive statistics

- Three basic categories of measures:
- Measures of **central tendency** describe the averages of the values (mean, median, mode)
- Measures of **frequency distribution** describe the frequency of each value (count)
- Measures of **variability** or dispersion describe how spread out the values are (variance, standard deviation)

Measures of central tendency

Measures of central tendency estimate the center, or average, of a data set

The **mean, median and mode** are 3 ways of finding the average



Example 02 Find the Mean, Median, Mode, and Range of the data set:

Hours Spent Studying Per Week

9 9 9 10 10 14 15 16 19 20

mean average 13.1

mode most common 9

median middle 12

range largest - smallest 11

Measures of central tendency

Visualisation: **histogram, density curve, boxplot**

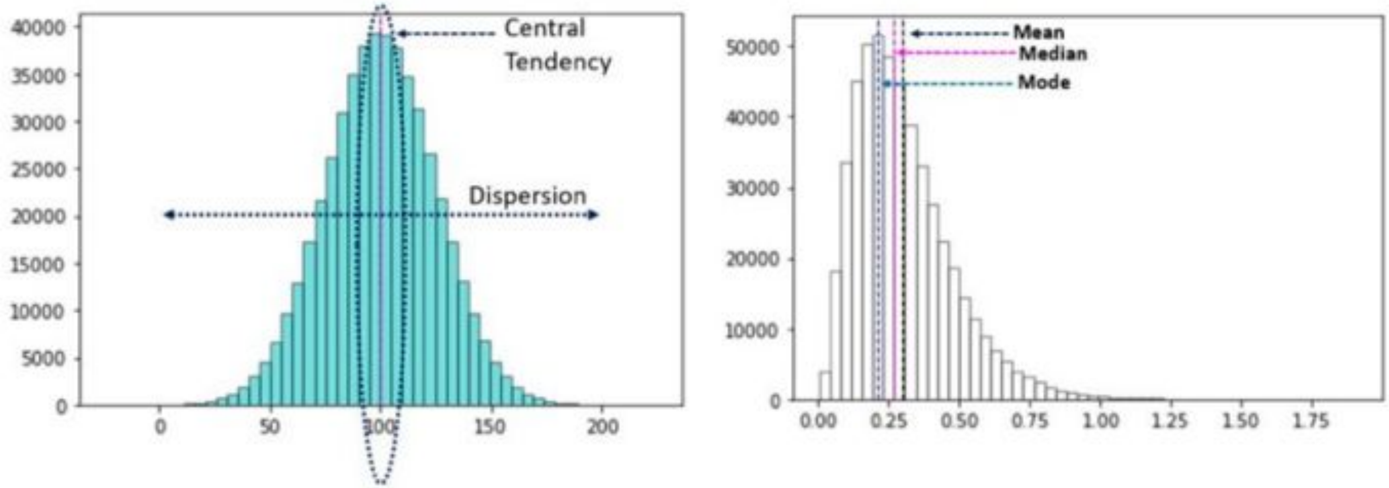
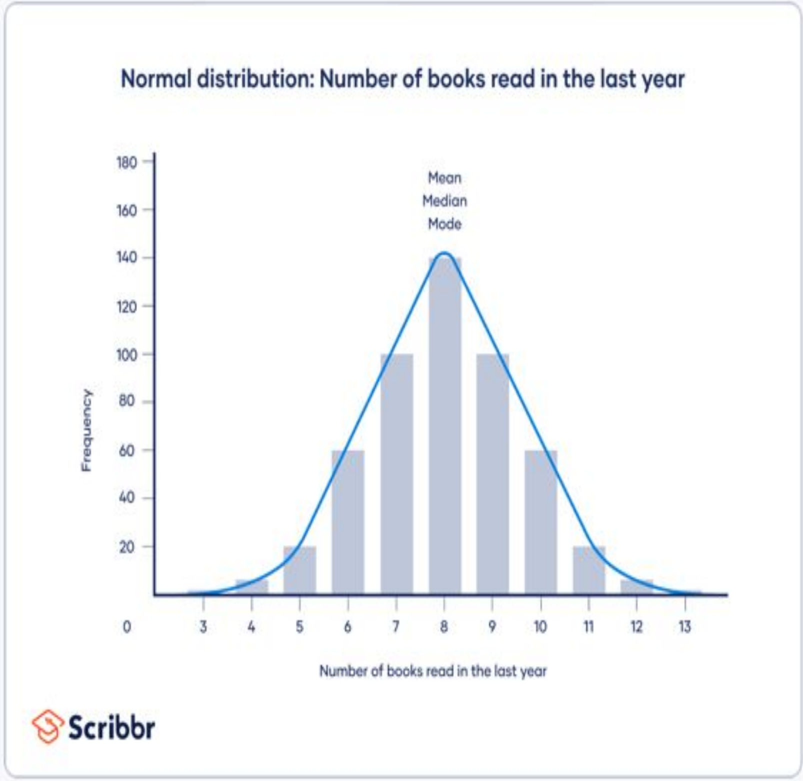


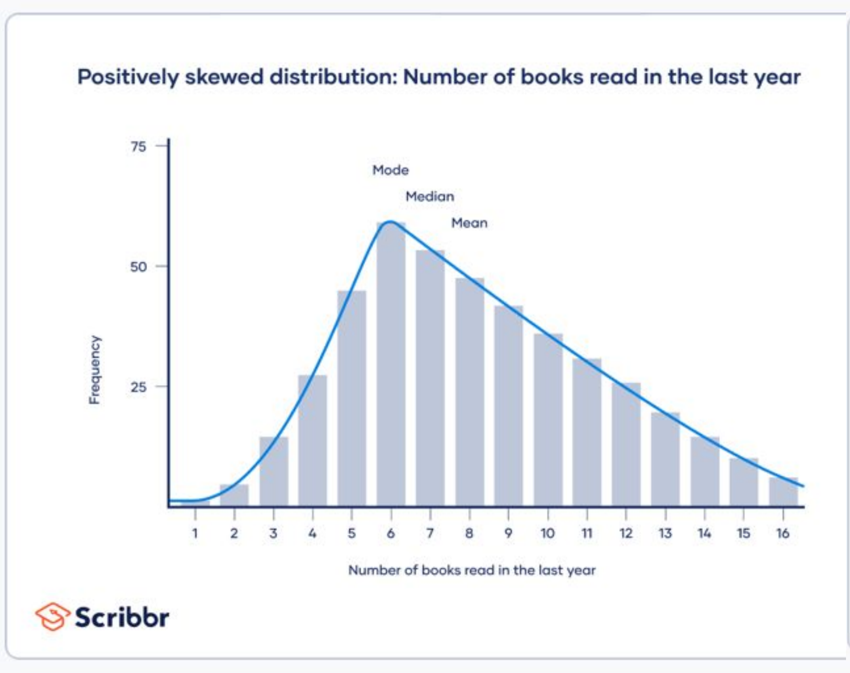
Fig 1: Central Tendency & Dispersion of plotted data in images

Normal distribution

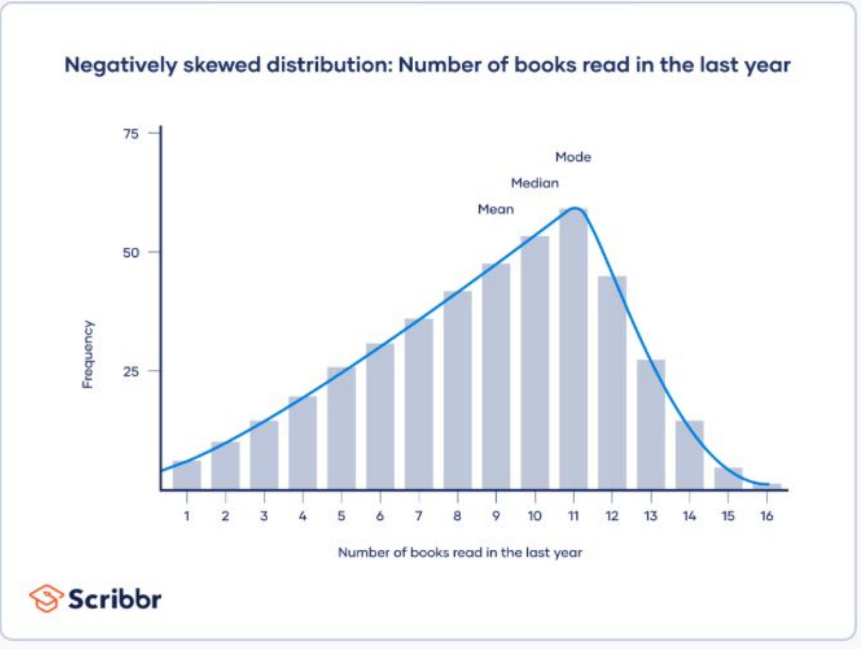
- Data is symmetrically distributed with no skew
- The mean, mode and median are exactly the same in a normal distribution.



Skewed distributions



Positively skewed distribution: Number of books read in the last year



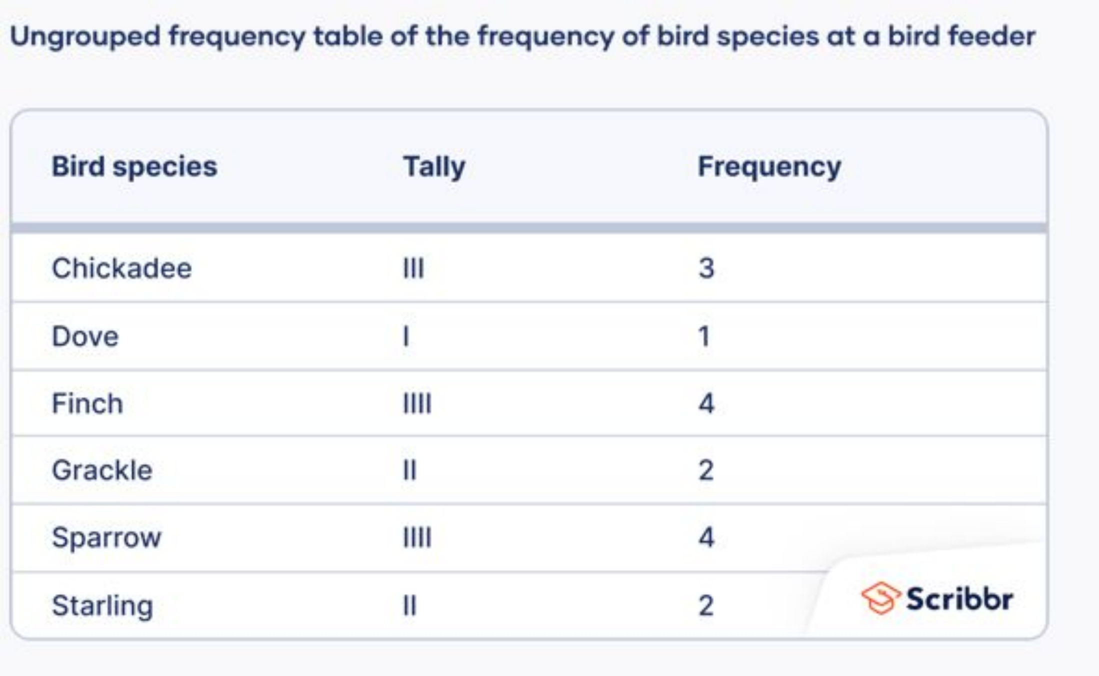
Negatively skewed distribution: Number of books read in the last year

Measures of frequency distribution

- A **frequency distribution** shows the number of times that a particular value occurs in a dataset
- Usually illustrated using **graphs and frequency tables**
- Types: **Ungrouped frequency distributions** (what type of variable?)
- **Grouped frequency distributions**
- **Relative frequency distributions**: The proportion (%) of observations of each value or class interval of a variable
- **Cumulative frequency distributions**: The sum of the frequencies less than or equal to each value or class interval of a variable

Ungrouped frequency distributions

- For categorical variables



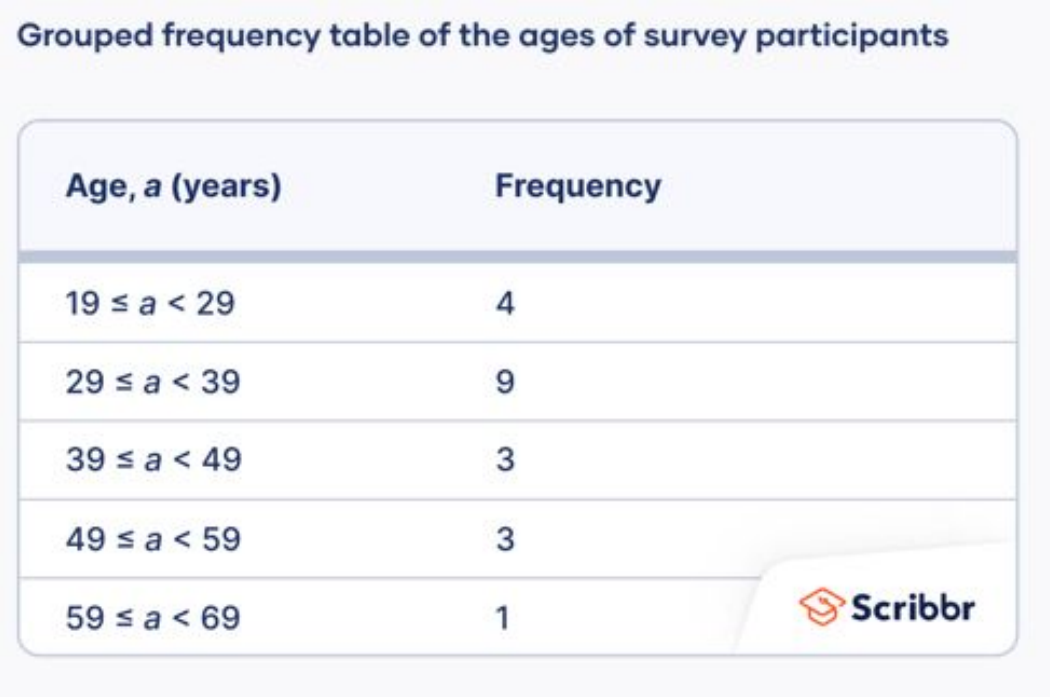
Ungrouped frequency table of the frequency of bird species at a bird feeder

Bird species	Tally	Frequency
Chickadee	III	3
Dove	I	1
Finch	IIII	4
Grackle	II	2
Sparrow	IIII	4
Starling	II	2

Scribbr

Grouped frequency distributions

- Divide the variable into class intervals



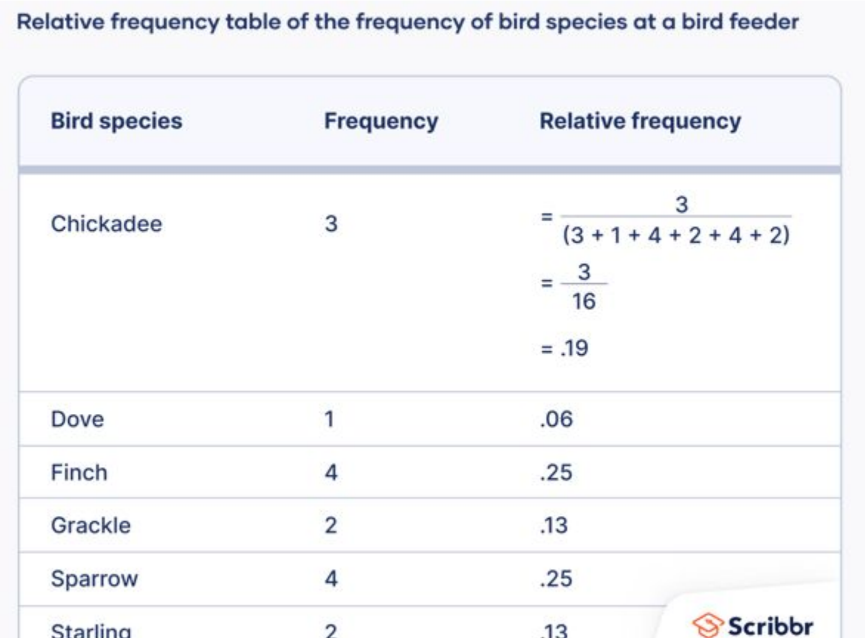
Grouped frequency table of the ages of survey participants

Age, *a* (years)	Frequency
19 ≤ *a* < 29	4
29 ≤ *a* < 39	9
39 ≤ *a* < 49	3
49 ≤ *a* < 59	3
59 ≤ *a* < 69	1

Scribbr

Relative frequency tables

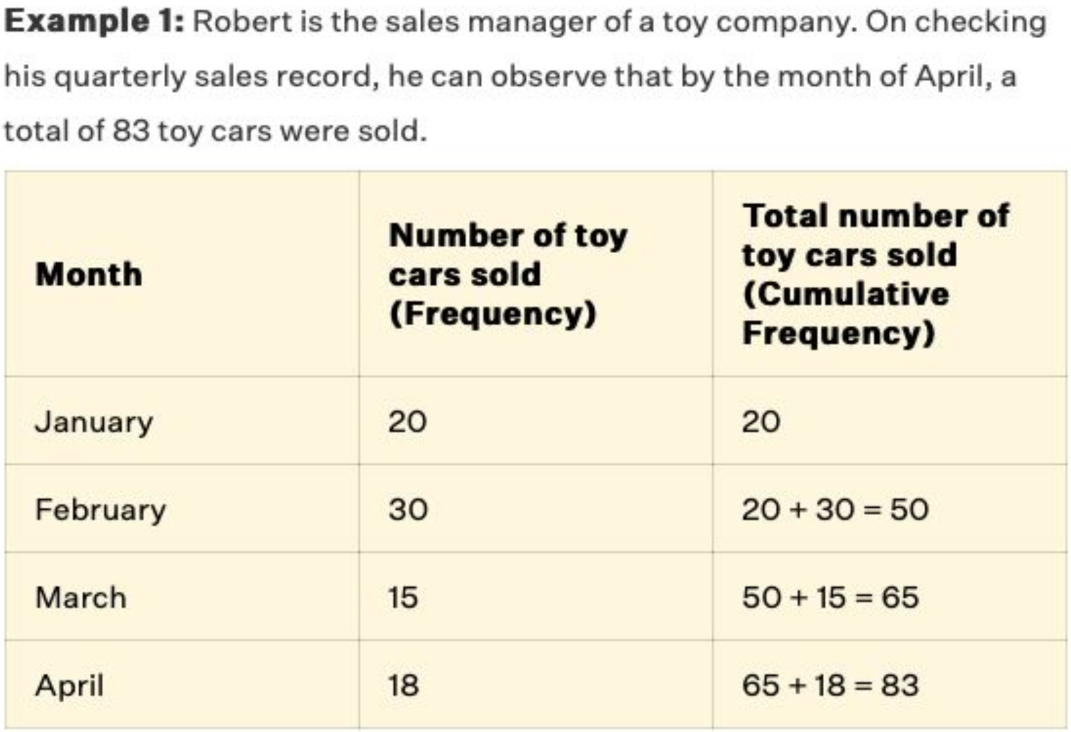
- The sample size -> the sum of the frequencies
- To calculate the relative frequencies, divide each frequency by the sample size



Relative frequency table of the frequency of bird species at a bird feeder

Bird species	Frequency	Relative frequency
Chickadee	3	$= \frac{3}{(3 + 1 + 4 + 2 + 4 + 2)}$ $= \frac{3}{16}$ = .19
Dove	1	.06
Finch	4	.25
Grackle	2	.13
Sparrow	4	.25
Starling	2	.13

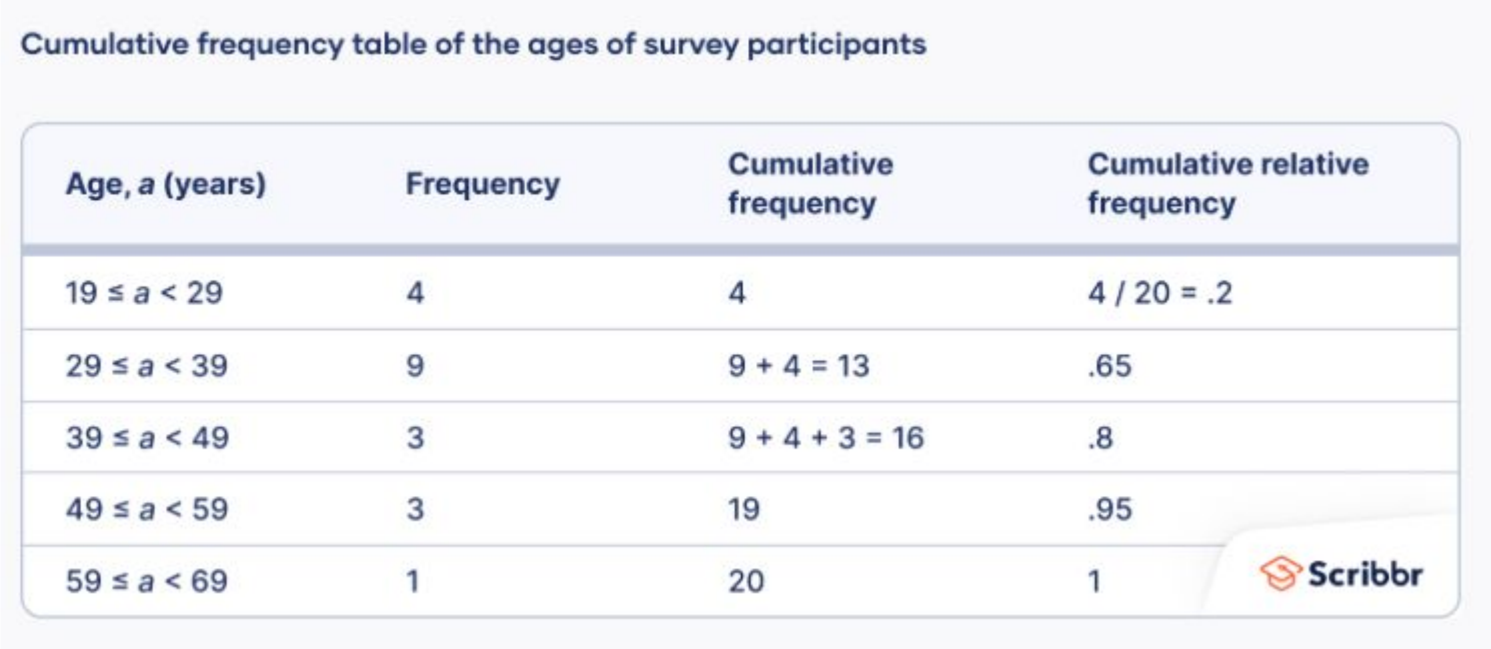
Cumulative frequency tables



Example 1: Robert is the sales manager of a toy company. On checking his quarterly sales record, he can observe that by the month of April, a total of 83 toy cars were sold.

Month	**Number of toy cars sold (Frequency)**	**Total number of toy cars sold (Cumulative Frequency)**
January	20	20
February	30	20 + 30 = 50
March	15	50 + 15 = 65
April	18	65 + 18 = 83

Cumulative relative frequency tables



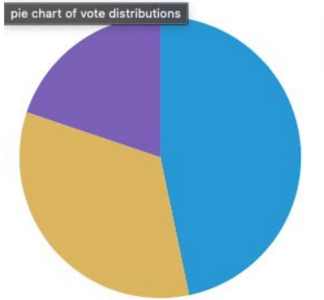
Cumulative frequency table of the ages of survey participants

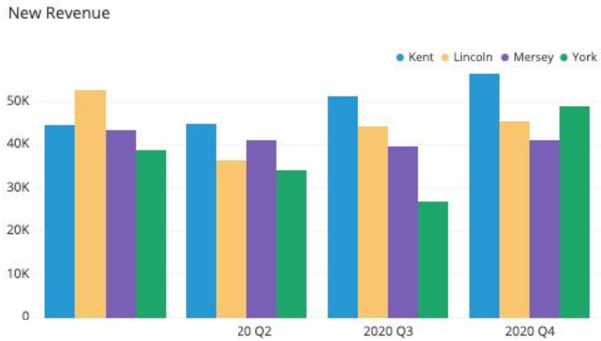
Age, *a* (years)	Frequency	Cumulative frequency	Cumulative relative frequency
$19 \leq a < 29$	4	4	4 / 20 = .2
$29 \leq a < 39$	9	9 + 4 = 13	.65
$39 \leq a < 49$	3	9 + 4 + 3 = 16	.8
$49 \leq a < 59$	3	19	.95
$59 \leq a < 69$	1	20	1

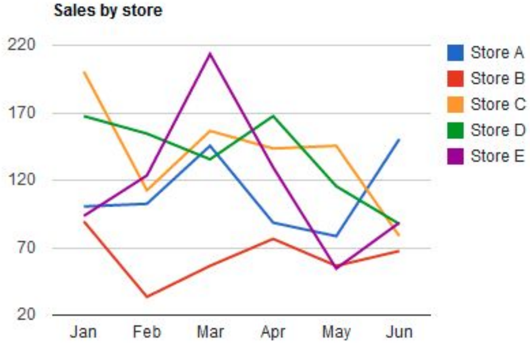
Frequency tables - visualisation

Frequency distribution can be visualised using:

- a **pie chart** (what kind of variable?)
- a **bar chart** (what kind of variable?)
- a **line chart** (what kind of variable?)
- a **histogram** (what kind of variable?)







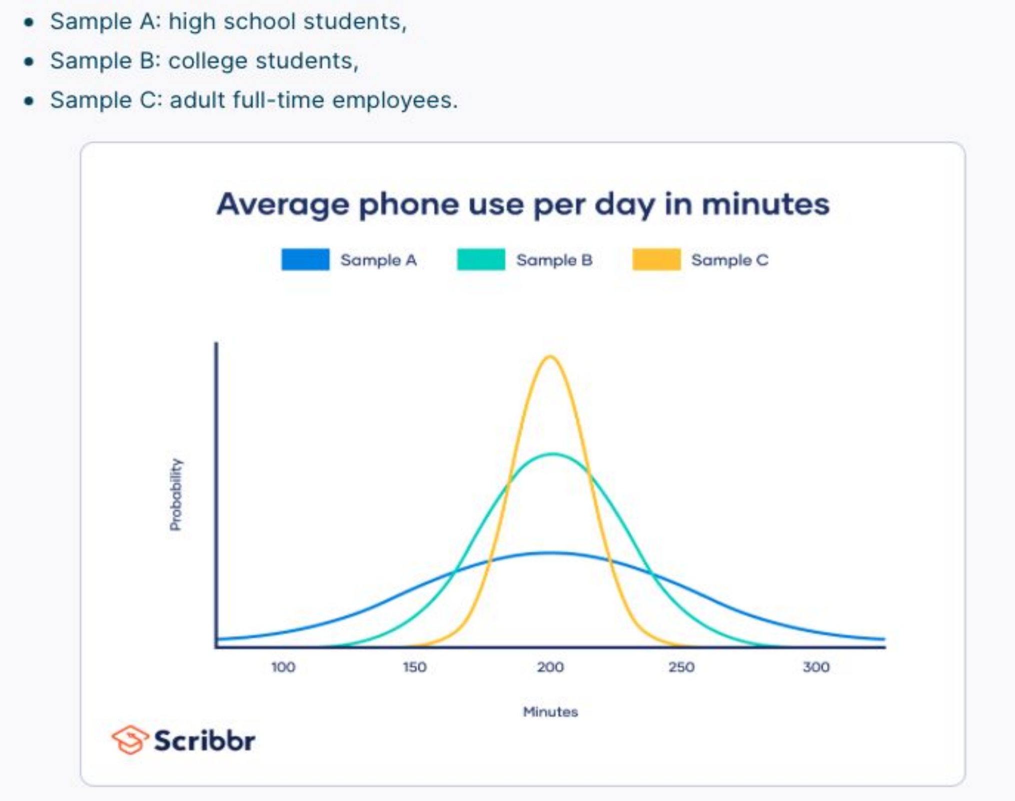
Measures of variability

- **Measures of variability** - to check how spread out the response values are

The range, quantiles, standard deviation and variance each reflect different aspects of spread

- **Range:** the difference between the highest and lowest values
- **Interquartile range:** the range of the middle half of a distribution
- **Standard deviation:** average distance from the mean
- **Variance:** average of squared distances from the mean

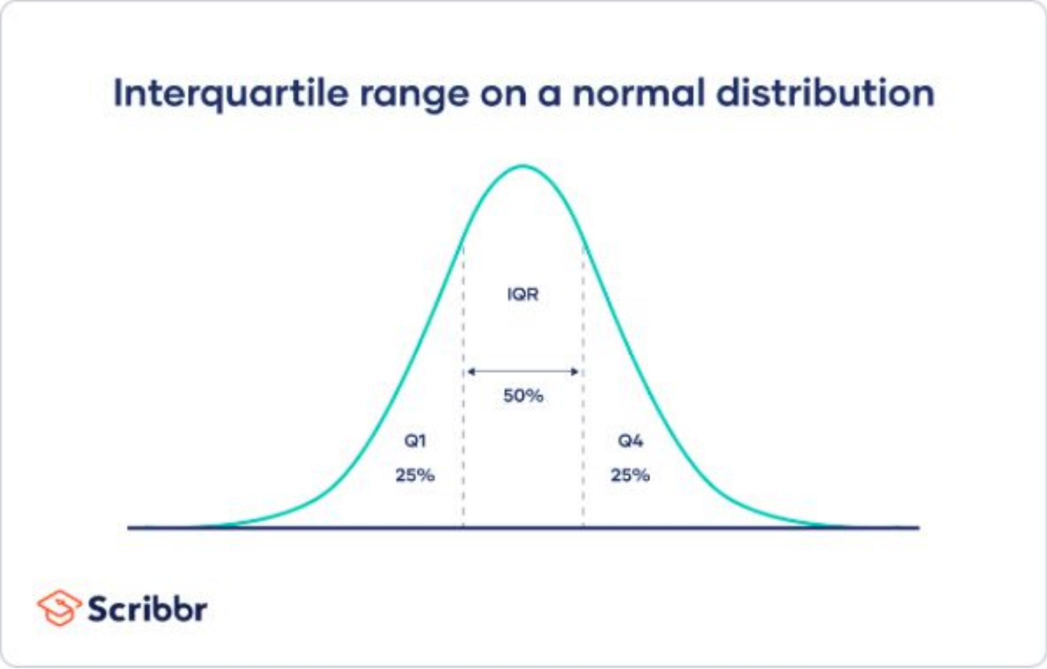
Variability in normal distribution



- Sample A: high school students,
- Sample B: college students,
- Sample C: adult full-time employees.

Interquartile range

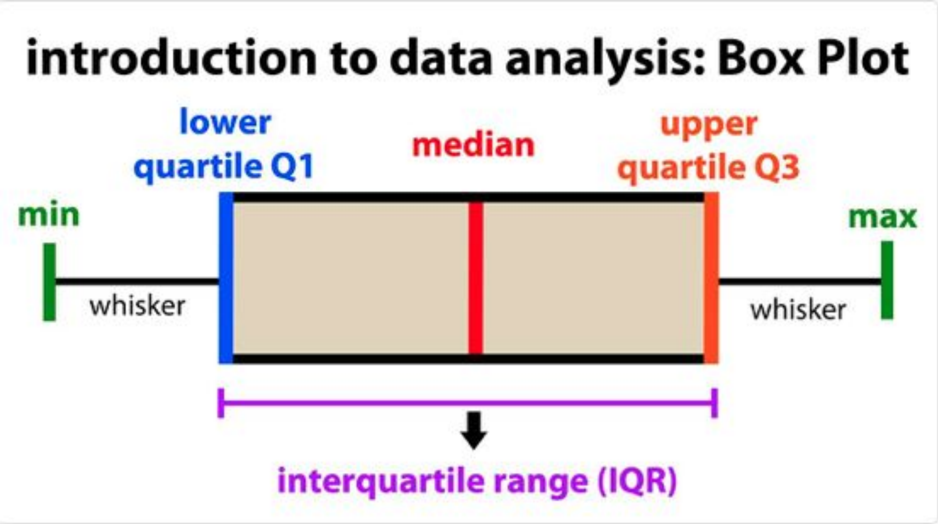
Provides information about the spread of **the middle of** your distribution (50%)



Five-number summary

Every distribution can be organized using **a five-number summary**:

1. Lowest value
2. Q1: 25th percentile
3. Q2: the median
4. Q3: 75th percentile
5. Highest value (Q4)


introduction to data analysis: Box Plot
lower quartile Q1
median
upper quartile Q3
min
max
whisker
whisker
interquartile range (IQR)

Comparison of three distributions

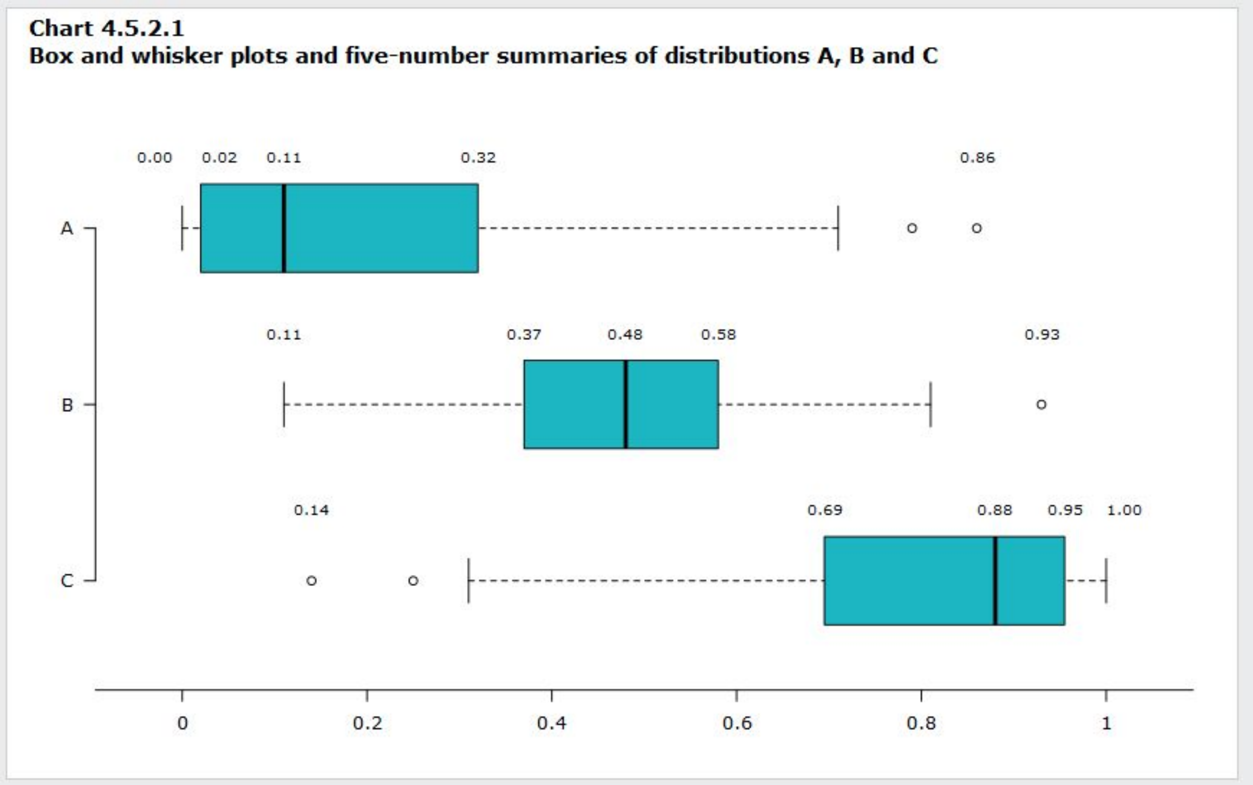


Chart 4.5.2.1
Box and whisker plots and five-number summaries of distributions A, B and C

Standard deviation

- **The standard deviation** is the average amount of variability in your dataset
- Information about **how far** each score lies from **the mean**
- Interpretation - **the larger** the standard deviation, **the more variable** the data set is

There are six steps for finding the standard deviation:

1. List each score and find their mean
2. Subtract the mean from each score to get the deviation from the mean
3. Square each of these deviations
4. Add up all of the squared deviations
5. Divide the sum of the squared deviations by N – 1
6. Find the square root of the number you found

Source: https://www.scribbr.com/statistics/variability/

Variance

- The **variance** is the average of squared deviations from the mea
- Variance reflects **the degree of spread** in the data set
- **The more** spread the data, **the large**r the variance is in relation **to the mean**
- To find the variance, you need to square the standard deviation

Source: https://www.scribbr.com/statistics/variability/

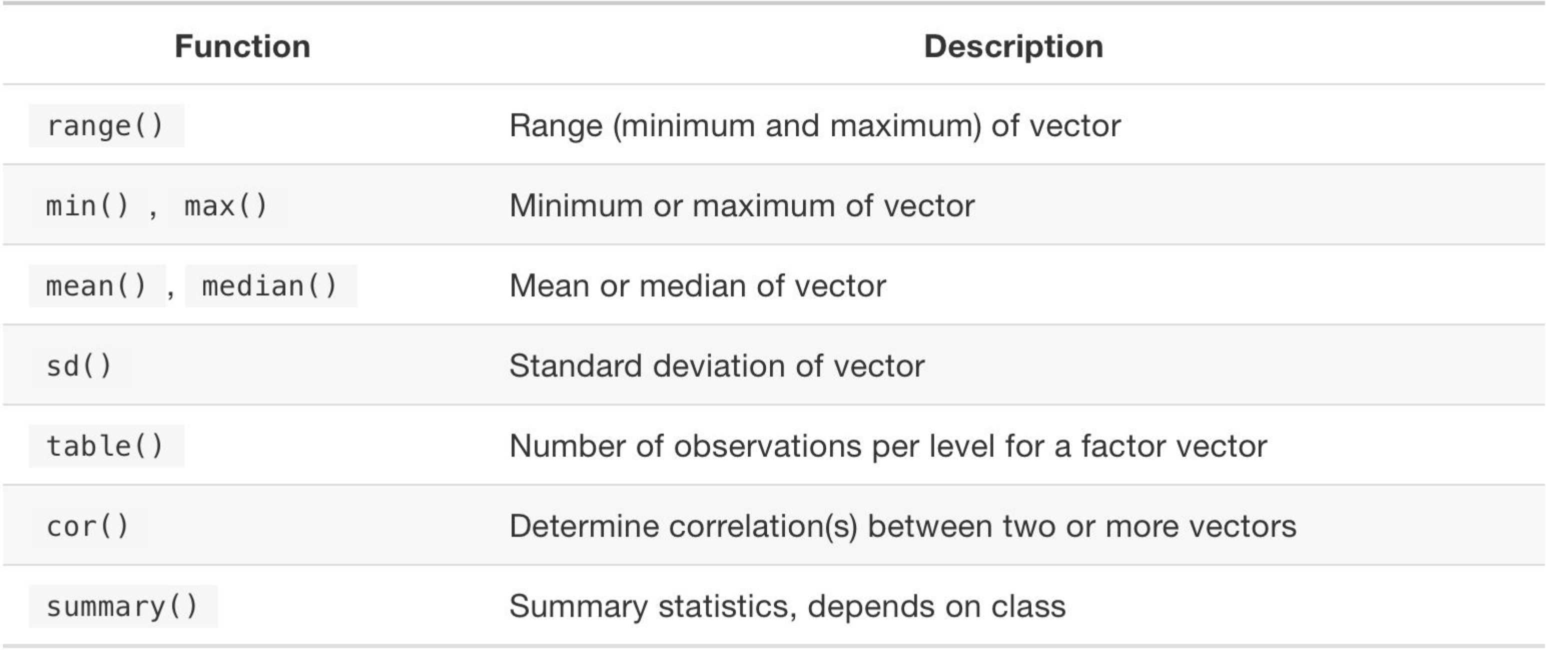
Data Exploration (EDA)

Exploratory Data Analysis - introduce the data

Questions to ask yourself when you explore the dataset:

- ❏ What metadata is available for this data set? Are the descriptions of variables provided? What do we know about the population / sampling?
- ❏ What are the observed population, the observation unit and the reference period?
- ❏ What are the data types of the variables? Do we need to change them?
- ❏ What are the frequency distributions of these variables? What are the measures of central tendency and dispersion? Anything that surprises you?
- ❏ Are there any outliers? Are there any values that look like errors?
- ❏ What is the mean for each variable?
- ❏ Are there any Null / NA values?

Data import in R



Function	Description
`range()`	Range (minimum and maximum) of vector
`min()` , `max()`	Minimum or maximum of vector
`mean()` , `median()`	Mean or median of vector
`sd()`	Standard deviation of vector
`table()`	Number of observations per level for a factor vector
`cor()`	Determine correlation(s) between two or more vectors
`summary()`	Summary statistics, depends on class

Data import in R

- **data()** function to load the datasets in R
- If you run the data function without an argument, R will display a list of the available datasets
- **data()** - the list of preloaded datasets from the datasets package
- Homework assignment - dataset “women”
- How to you import it to R?
-

Data frames in R

- A data frame is **a table** or a two-dimensional array-like structure: each column contains values of one variable and each row contains a set of values from each column
- **Create** a data frame - > *data.frame()* function
- Check if a variable is a data frame or not, if not **change it** to a data frame
- **Summary** of data : a list of functions to explore the data set (check the R script)
- **Extracting** data from a data frame -> **$** operator and column name or **square brackets** and index, i.e. **subsetting data**
- **Modifying** data frames: expanding the data frame by adding columns or rows
- **Deleting** components