Queen Mary School Hainan
Queen Mary University of London

QHP5701 Exploratory Data Analysis

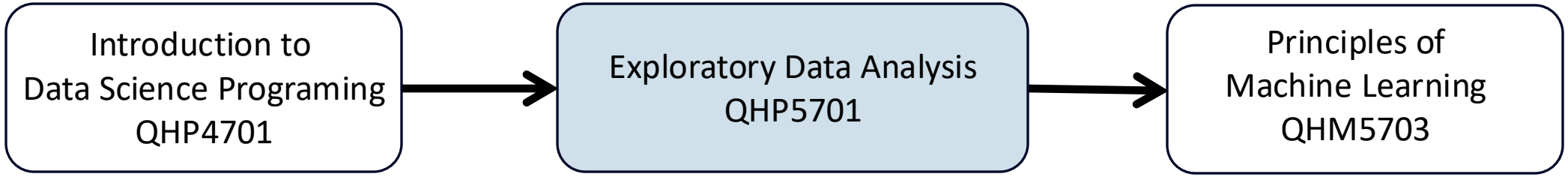
Nikesh Bajaj, PhD
Lecturer in Data Science,
Queen Mary University of London
nikesh.bajaj@qmul.ac.uk
https://nikeshbajaj.in

Introduction of QHP5701: EDA

What this course is about?

This course is:

- An intermediate of Introduction to Data Science Programming (QHP4701) and Principles of Machine Leaning (QHM5703)
- ***About exploring and analysing data before we give an attempt to build prediction and forecasting models***
- Focused on analysis different kinds of data, specifically time-series (time dependent data)



Introduction of QHP5701: EDA

Aims and Objective of QHP5701

This module aims to provide introductory skills to perform exploratory analysis of data and:

- build foundations of statistics and signals and systems,
- underpin analysis of time dependent sources and signals (time-series)

That will allow you to explore and understand the given data and will help you to go further building prediction models, which you will learn in next module (PML)

To build the strong foundations, the large part of this module includes extensive theoretical components, which include solving and deriving mathematical problems and theorems.

Module Assessments

- Coursework: 40%

  - 10% Quiz-1 (focused on theory)
  - 10% Quiz-2 (focused on theory)
  - 20% Assignment (focused on programming)

- Examination: 60%

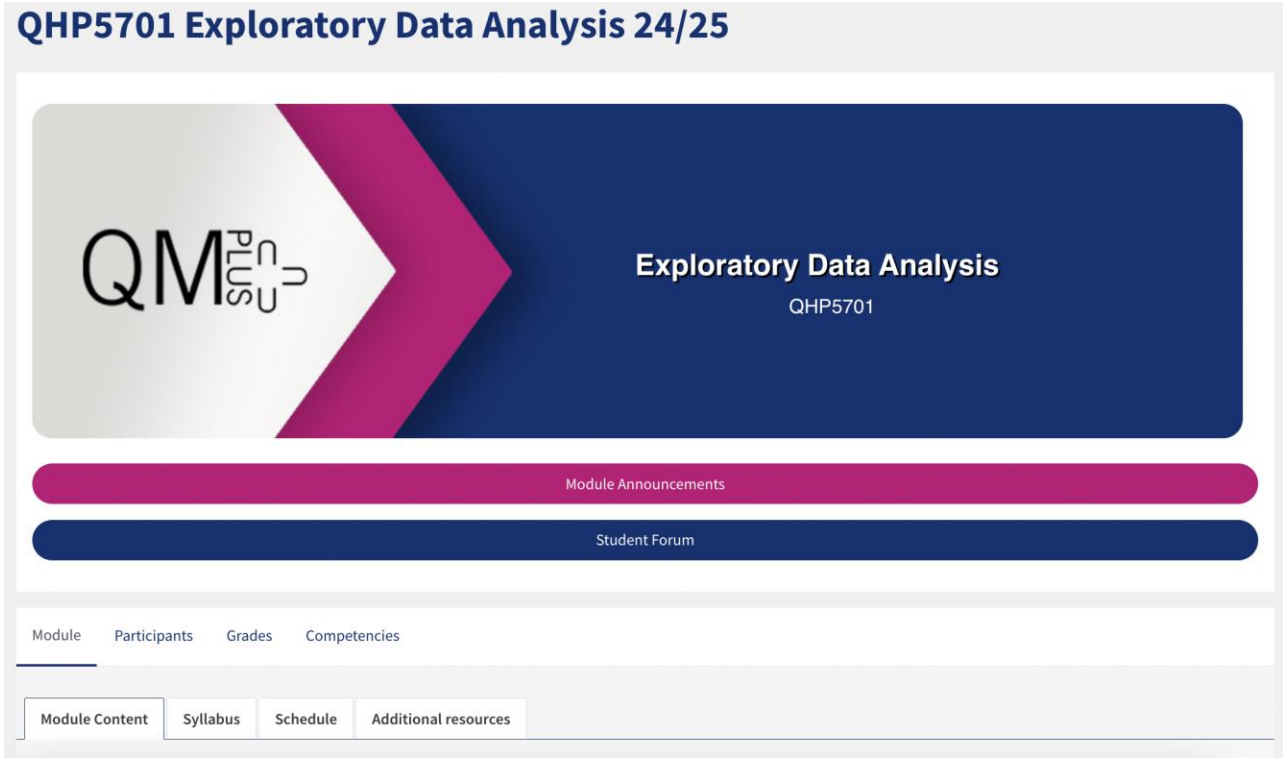
  - Exam (focused on theory)

Team

- Module Lecturer
  - Nikesh Bajaj (***Nik***)
- Teaching Fellow
  - Jiayu Men

Module Information

All the relevant information about the Module - **QHP5701**, can be found on QM+ Page



Communication

Any question, query and doubt can be asked in following ways

- During Lectures (**highly recommended**)
- On campus or remotely (via MS teams)
- Email: Please make sure its subject is formatted as follows: "[QHP5701] <DESCRIPTIVE SUBJECT HERE>"

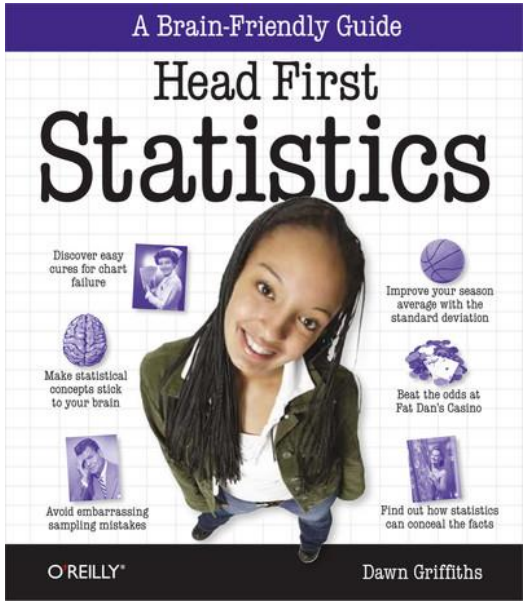
  *e.g "**[QHP5701] Question about Coursework 1**"*

- Forum on QM+: Primary means, questions might have been answered already and answers might be useful to others.

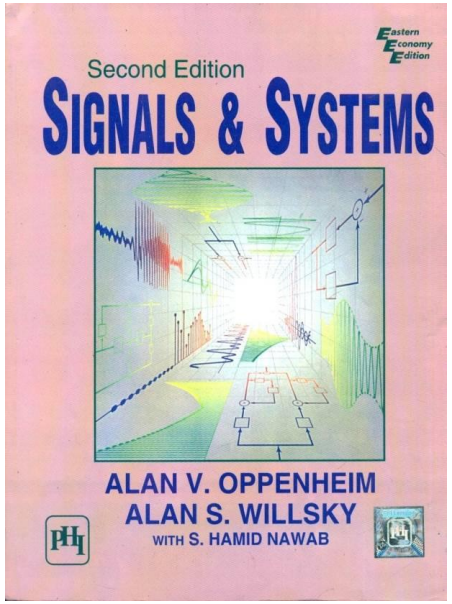
Learning Resources: Books

Books

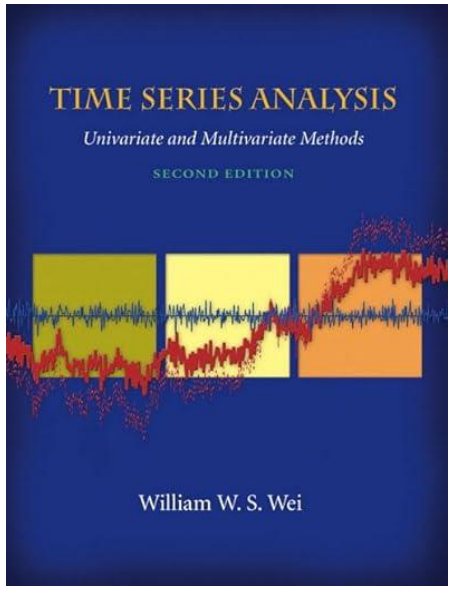
- Statistics



- Signal & Systems



- Time Series Analysis



Overview

Module Contents

Statistics

- Describe your data : Summarising and Visualisation
- Confidence Interval, Standard Error, Correlation

Types of Data

- Tabular, Grid Data, Graph
- Time-series data

Signals

- Elementary & Properties
- Sampling, Fourier

Sources

- Stochastics processes
- AFC, AFV, Noise,, Visualisation

Systems: Processing

- Properties, Convolution, Filtering
- MA, ARMA, ARIMA
- Stationary & Non-stationary
- Seasonality

Queen Mary University of London

QHP5701 Exploratory Data Analysis

Statistics

Nikesh Bajaj, PhD
Lecturer in Data Science,
Queen Mary University of London
nikesh.bajaj@qmul.ac.uk
https://nikeshbajaj.in

Statistics

- Describe your data
  - Descriptive statistics - summarising the data
  - Visualisation (plots and figures)
- Inferential Statistics
  - Confidence Interval, Standard Error, Correlation

QHP5701 Exploratory Data Analysis

Statistics - Part 1: Describe your data

Nikesh Bajaj, PhD
Lecturer in Data Science,
Queen Mary University of London
nikesh.bajaj@qmul.ac.uk
https://nikeshbajaj.in

Describe your data

- Types of Variables
- Descriptive Statistics:
  - Average: Mean, Mode, Median, Frequency distribution
  - Spread/variability: Range, Percentile, Standard deviation
  - Skewness, Outliers
- Visualization: Plots

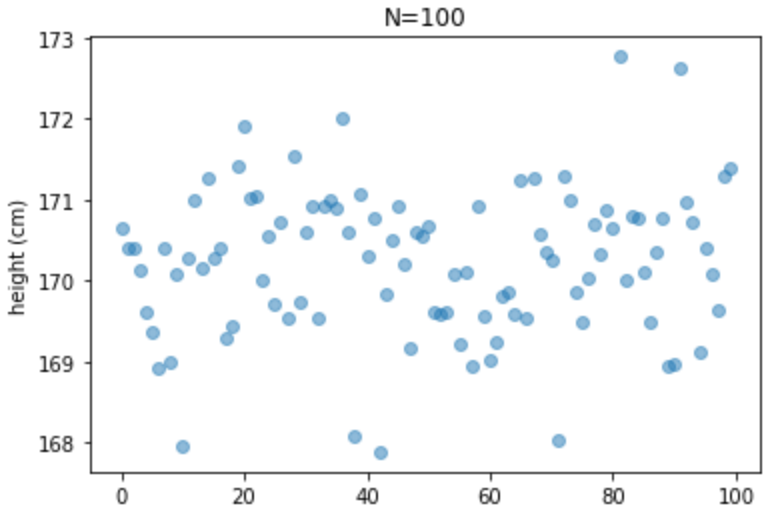
How would you describe it?

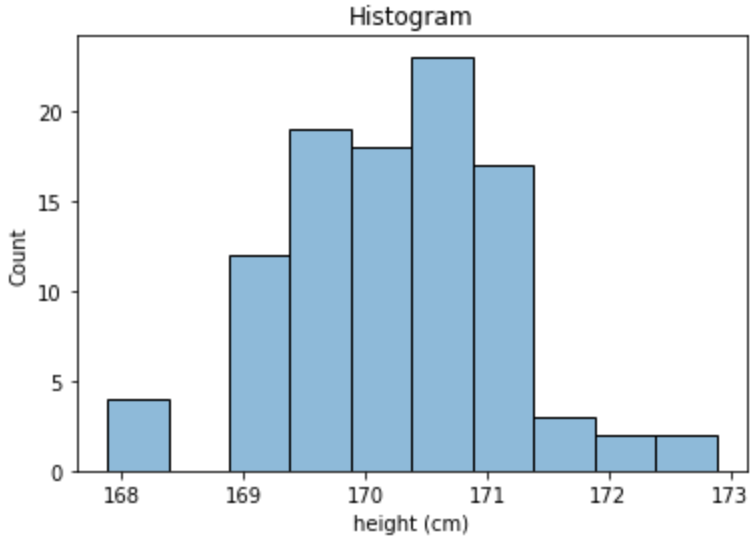
N=9

Height (cm)
167
168
160
170
171
160
162
165
167

N=100

Height (cm)
167
168
160
170
171
.
.
.
.
.
.
.





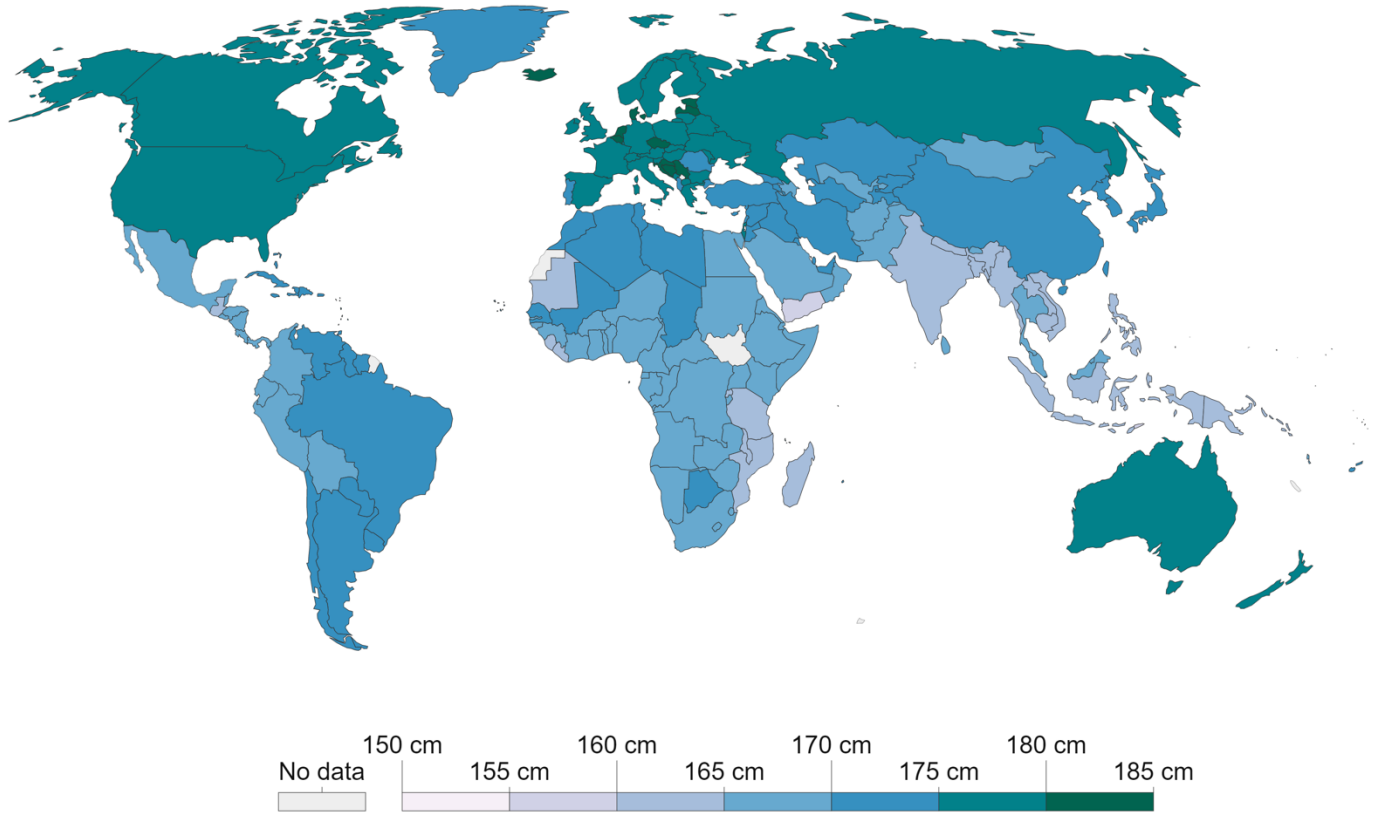
Average – centre tendency of data
Variability – spread of data
Nice plots!

Example:

Average height of men by year of birth, 1996

Mean height of adult men by year of birth. Data for the latest cohort (the year 1996) is therefore the mean height of men aged 18 in 2014.

Our World in Data



Source: NCD RisC, Human Height (2017)

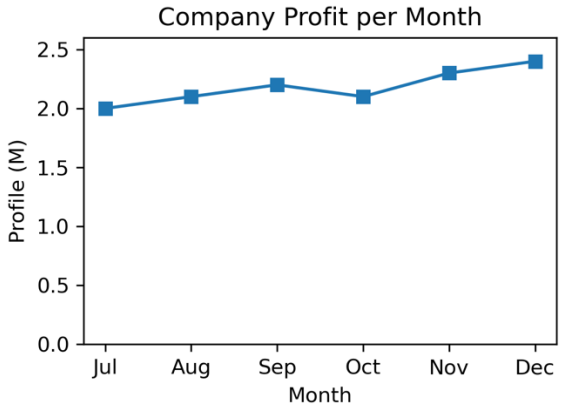
OurWorldInData.org/human-height • CC BY

Visualisation – just plotting the data

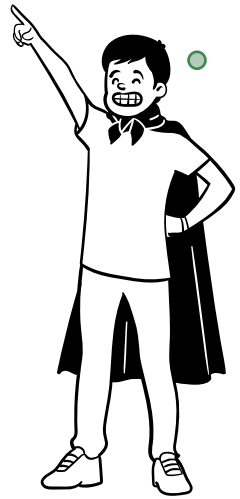
Impression of visualization: Profit of company

Months	Jul	Aug	Sep	Oct	Nov	Dec
Profit (M)	2.0	2.1	2.2	2.1	2.3	2.4

The profit is holding steady, it's nothing special.



This stock is so hot, it's smoking!



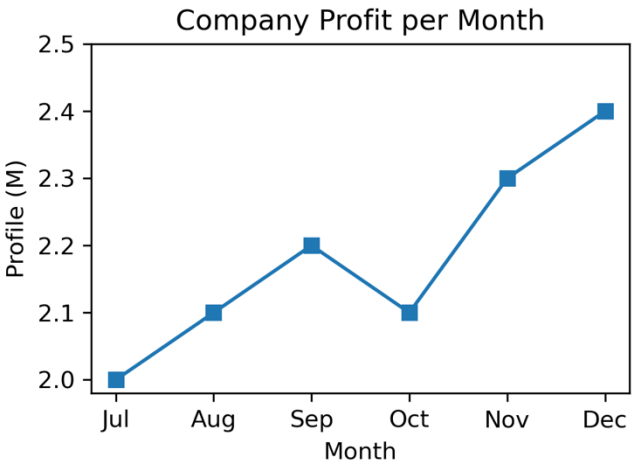


Image Ref: Book: First Head "Statistics" O'Relley

Plot & type of variable

Numerical

- Quantitative: (height)

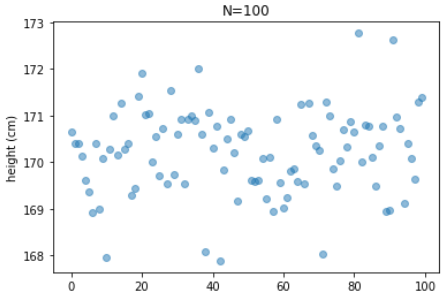
Categorical

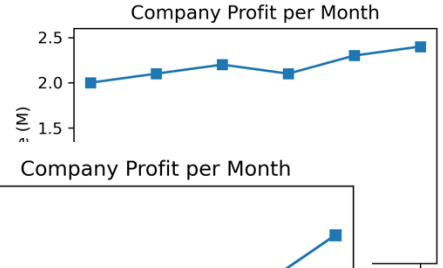
- Qualitative:

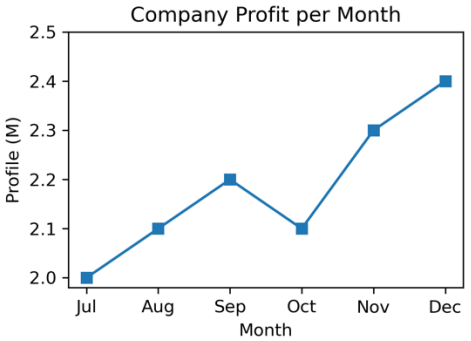
Genre	Unit Sold
Sports	27,500
Strategy	11,500
Action	6,000
Shooter	3,500
Other	1,500

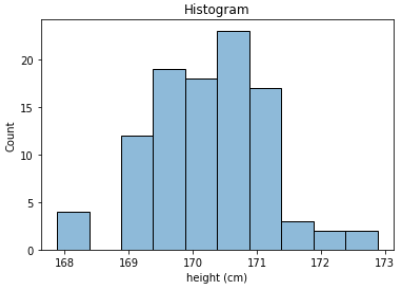
- Ordinal:

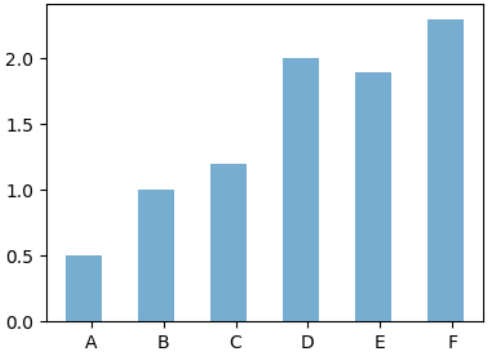
Hours	Frequency
0 --1	4,300
1--3	6,900
3--5	2,000
5--10	1,000
10--24	3,000

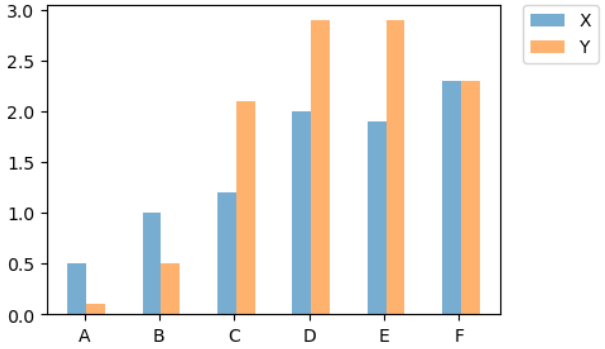












Ref: Book: First Head “Statistics” O’Relley

Types of Variable

Numerical

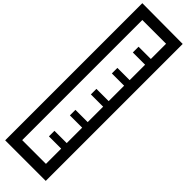
- Quantitative:
  - blood pressure, sugar level, no of cells, height, BMI
  - Continues, Discrete

Categorical

- Qualitative:
- Nominal
  - ethnicity, disease or not? , sex?
  - nominal (>2 cat.), binary (2 categories),
- Ordinal:
  - satisfaction-rating, age-group

Numerical or Quantitative Data

time

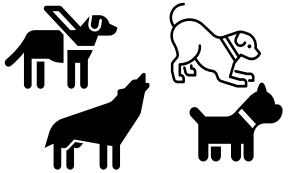


distance

weight

Categorical or Qualitative Data

breed of dog



type of dessert

rating

age-group

Descriptive Statistics

Summarizing the data

- Average: Mean, Mode, Median
- Frequency distribution
- Spread/variability: Range, Percentile, Standard deviation
- Skewness, Outliers

- What?, When?, Which?

Types of Variable

Numerical

- Quantitative:
  - blood pressure, sugar level, no of cells, height, BMI
  - Continues, Discrete

Categorical

- Qualitative:
- Nominal
  - ethnicity, disease or not? , sex?
  - nominal (>2 cat.), binary (2 categories),
- Ordinal:
  - satisfaction-rating, age-group

Operators (where we can use?)

+. -, x

>, <

=, ≠

Average: mean, mode, median

Most representative value of data

$$\tilde{x} = \frac{1}{N}\sum x_i$$

- Height in class

[4, 4, 5, 4, 4, 4, 5, 4, 3, 5.5, 6, 4.5, 4.2, 5.2, 5,5, 6,1]

- Preference of drink

[tea, tea, coffee, coffee, tea, tea, milk, tea, coffee, coffee]

- Age-group

[10-15, 10-15, 10-15, 15-20, 20-25, 20-25, 20-25]

Mean: sum of all values/ number of values

Median: middle value of sorted sequence

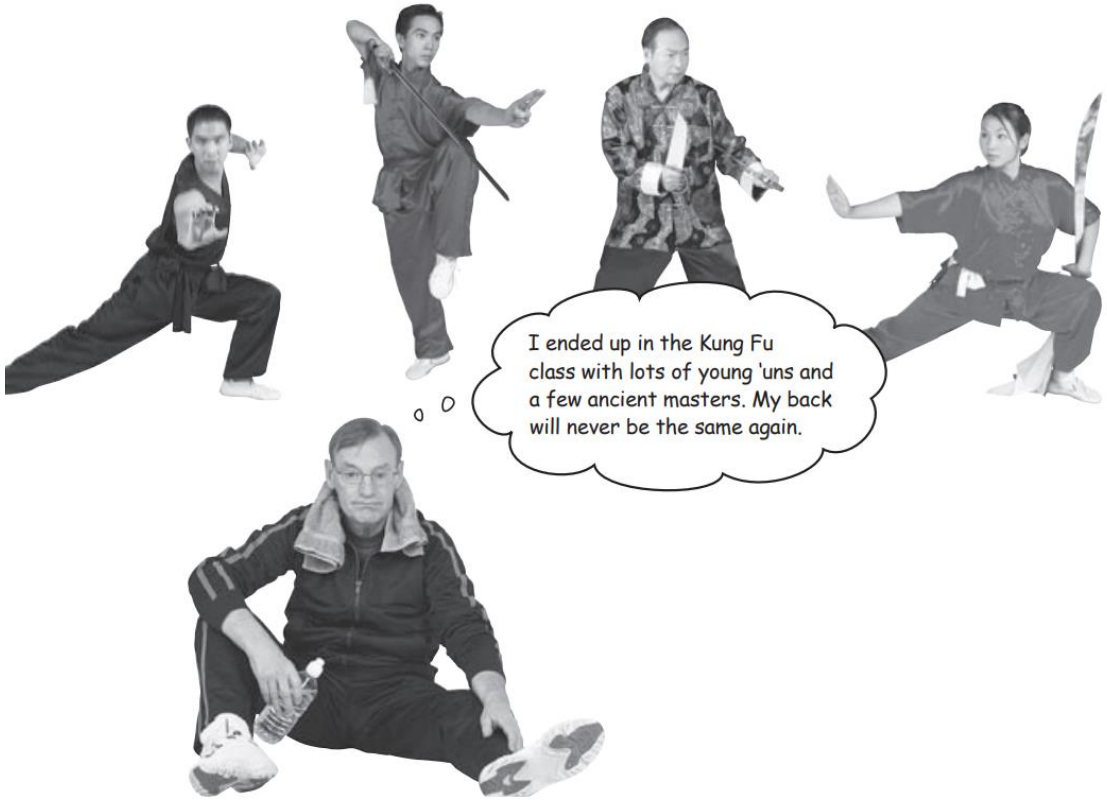
Mode : most frequent value

Let's see a case: A Health Club

The London Health Club
A perfect match for you

"London Health club proud to have class for everybody"

New customer in 50s



- Tuesday Evening

Class	Mean age
Class 1 :	17
Class 2 :	25
Class 3 :	38

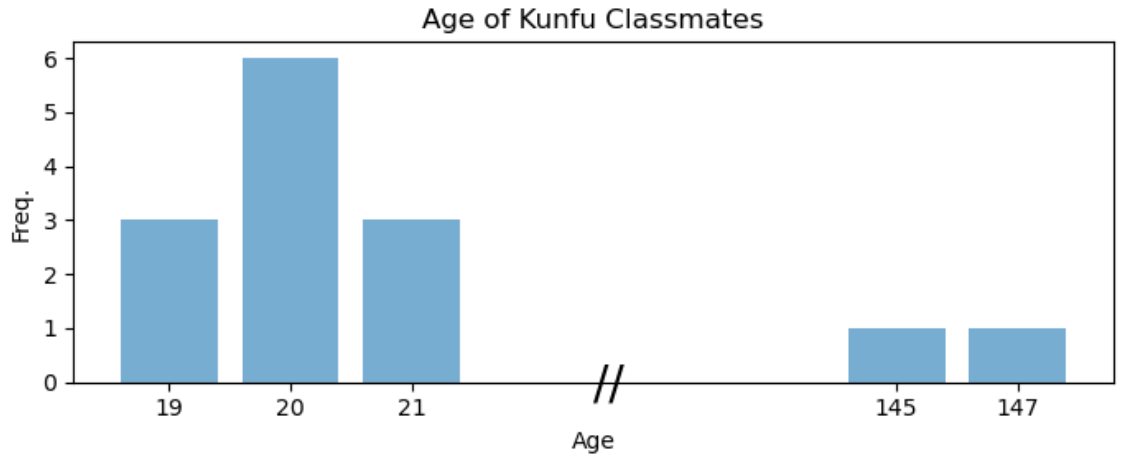
Which class new customer should attend??

Ref: Book: First Head "Statistics" O'Relley

Health club

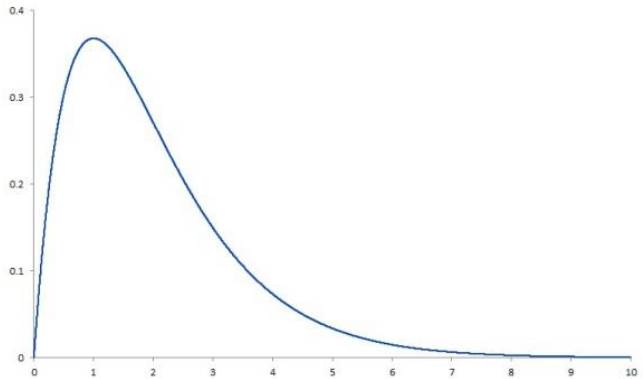
The London Health Club
A perfect match for you

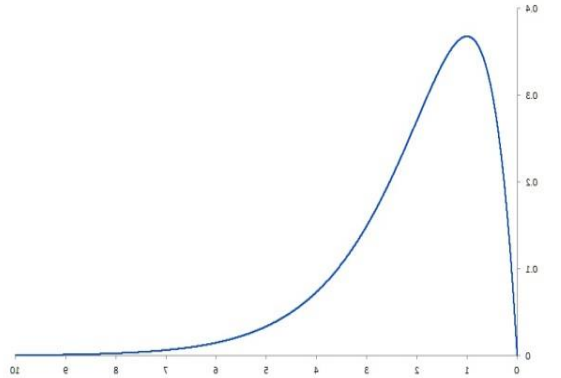
Age	19	20	21	145	147
Frequency	3	6	3	1	1



Median saved the day

Skewed distribution





Median:

19 19 20 20 20 21 21 100 102

Here's the number in the middle. This is the median, 20.

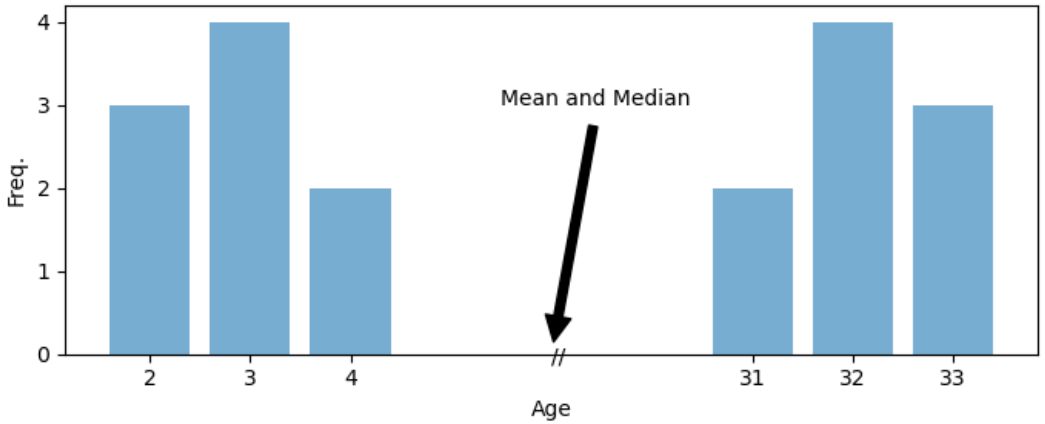
Ref: Book: First Head "Statistics" O'Relley

Health club

Sounds Cool.. Sign me up right now!!

teenager

Can anything go wrong?



Class was for parents who bring their children to teach swimming?

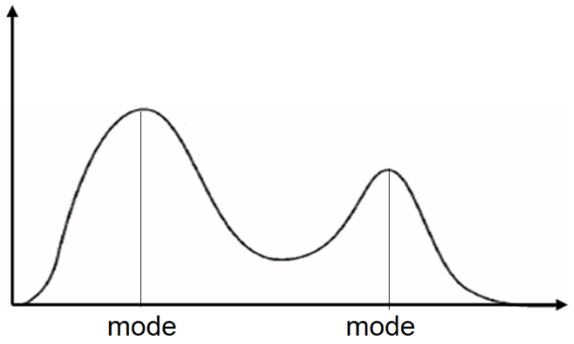
Mode is our solution

The London Health Club
A perfect match for you

Swimming Class
Median Age: 17

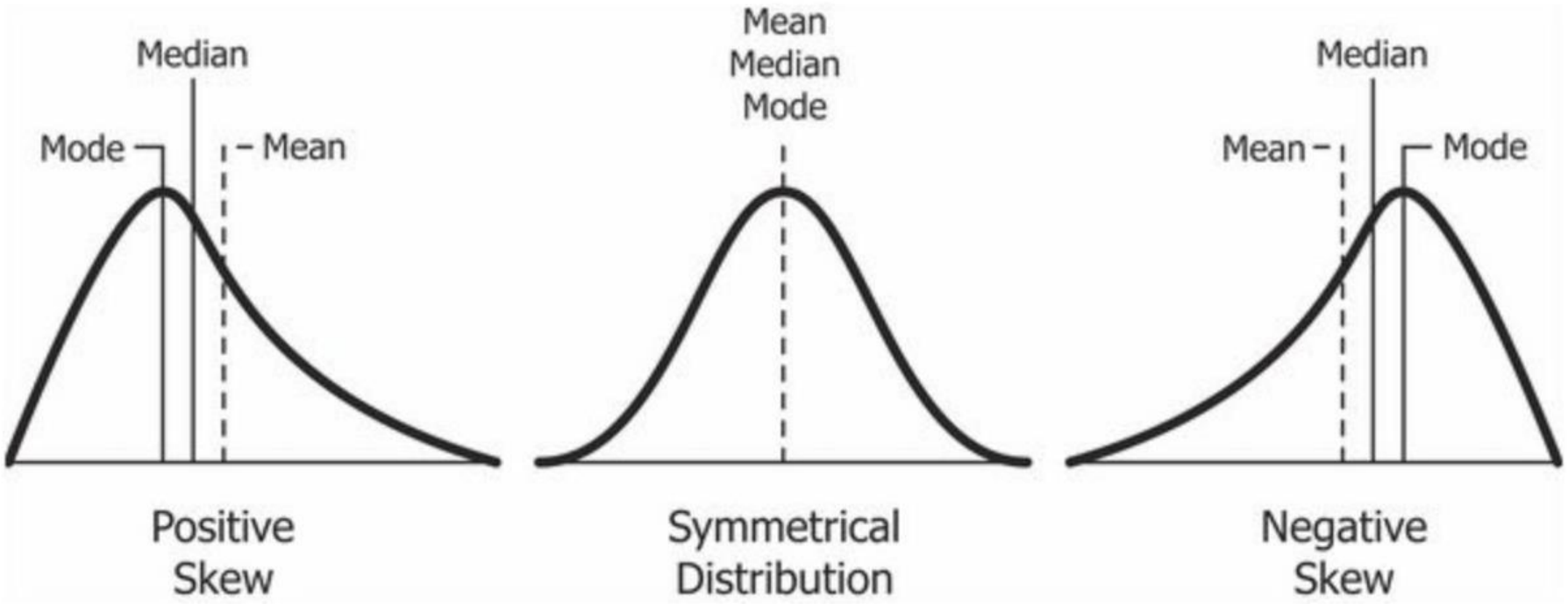
Mean = 17

Median = 17



Ref: Book: First Head “Statistics” O’Relley

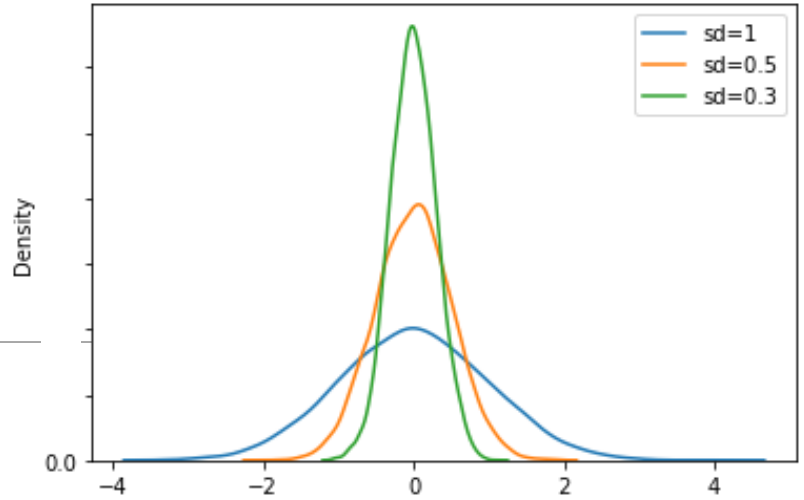
Skewness

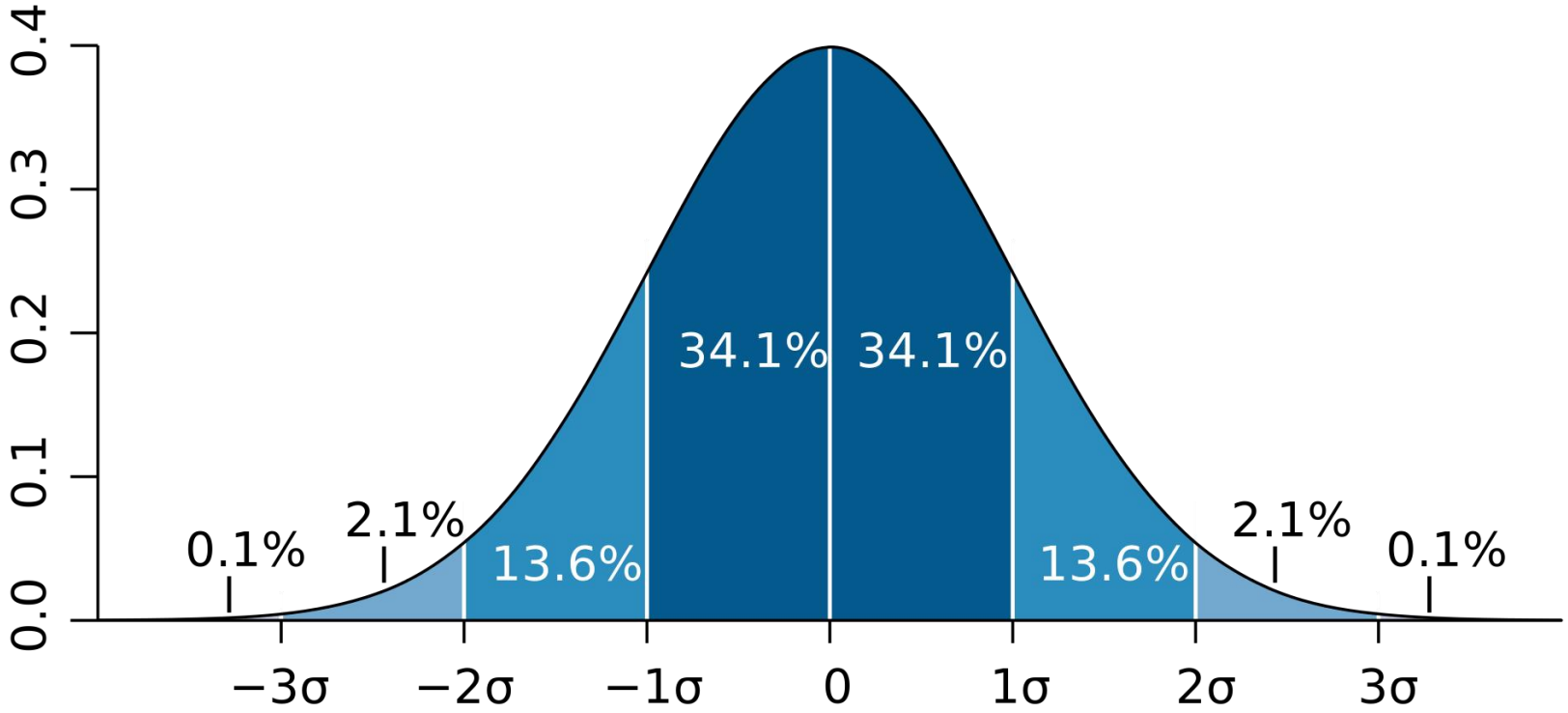


https://en.wikipedia.org/wiki/Skewness

Spread/Variability

Normal distribution





- Standard deviation σ *(sd)*
- Variance σ^2 *(var)*
- Range
- Interquartile Range (IQR)

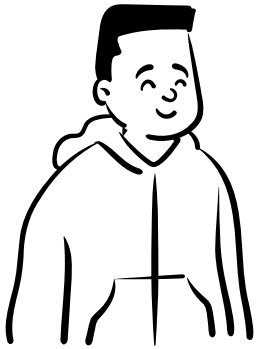
https://en.wikipedia.org/wiki/Normal_distribution

Question

You are a coach for a cricket (or football) team, need to hire a new player

- Player 1, **mean score** of run rate (or goal rate) is 8, with **standard deviation** of 4
- Player 2, mean score of run rate (or goal rate) is 8, with **standard deviation** of 2

Who would you hire?

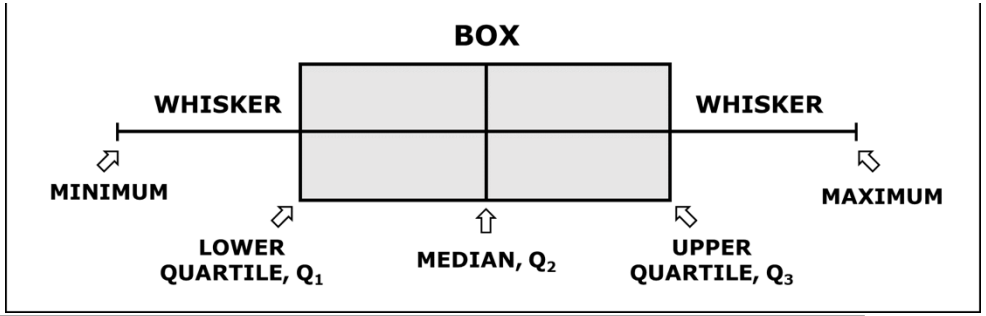


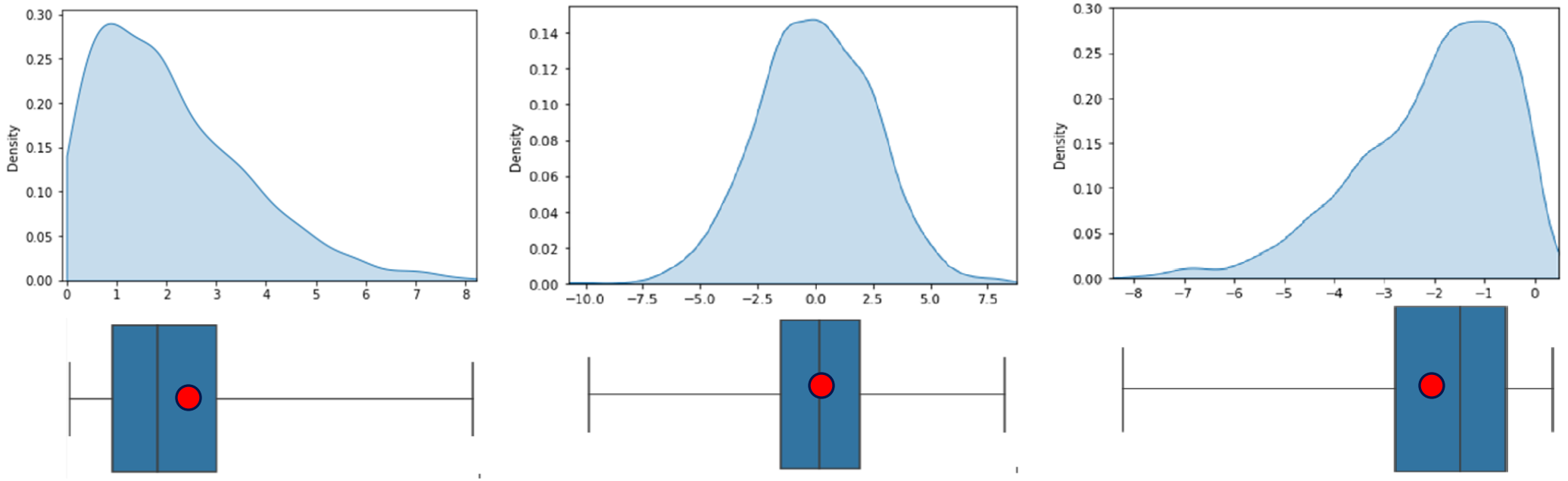
Player 1

Player 2

Go to **www.menti.com** and use code 1435 4444

Box-whisker plot

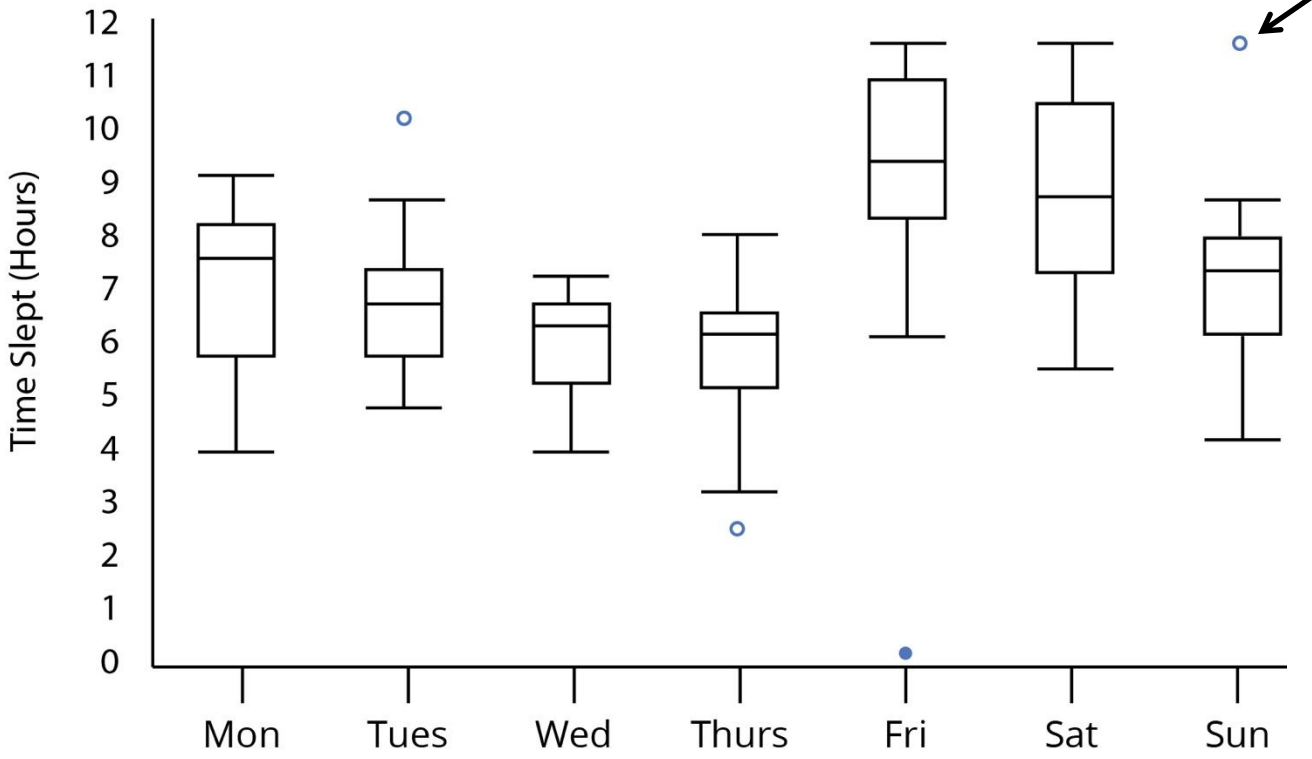




Visual comparison with boxplot

Outlier:
> Q3 + 1.5xIQR
< Q1 - 1.5xIQR

For a year, number of hours person P sleeps on different days



Questions:

1. On which day, P sleeps very less/ a lot?
2. Day with most (in)consistent hours of sleep

Can we categorise cont. data?

- To improve the interpretation
- Example : BMI into 2 or 3 categories - high or low BMI
- Implications?
  - Loss of information – loss of statistical power to detect the differences
  - Impact of choosing –where to cut
  - Splitting at median (dichotomising) - reduces statistical power
- Worst for binary than 4 or more categories

Formulas

Mean: $\bar{x} = \frac{1}{N}\sum x_i$

Median: sort x and choose a middle value*

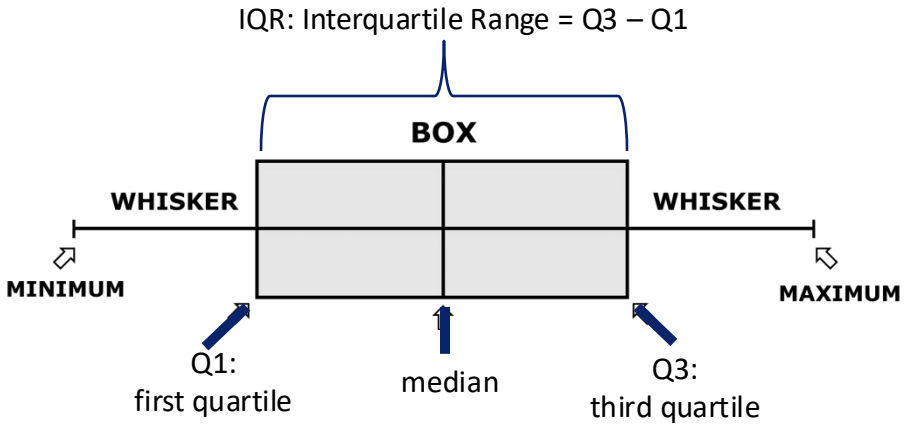
Mode: most frequent values

Variance: $\sigma^2 = \frac{1}{N}\sum(x_i - \bar{x})^2$

Standard Deviation: σ

Skewness: $g_1 = \frac{m_3}{m_2^{3/2}}$ $m_k = \frac{1}{N}\sum(x_i - \bar{x})^k$

Skewness with corrected bias: $G_1 = \frac{\sqrt{N(N-1)}}{N-2} g_1$



Outlier:
> Q3 + 1.5 IQR
< Q1 - 1.5 IQR

Range = maximum – minimum
Range excluding outliers = maximum* - minimum*

Queen Mary University of London

QHP5701 Exploratory Data Analysis

Statistics - Part 2: Inferential Statistics

Nikesh Bajaj, PhD
Lecturer in Data Science,
Queen Mary University of London
nikesh.bajaj@qmul.ac.uk
https://nikeshbajaj.in

Inferential Statistics

- Sample and Population
- Estimate population parameters from sample, and its accuracy (standard error)
- Standard error and standard deviation
- Confidence interval
- Size of data

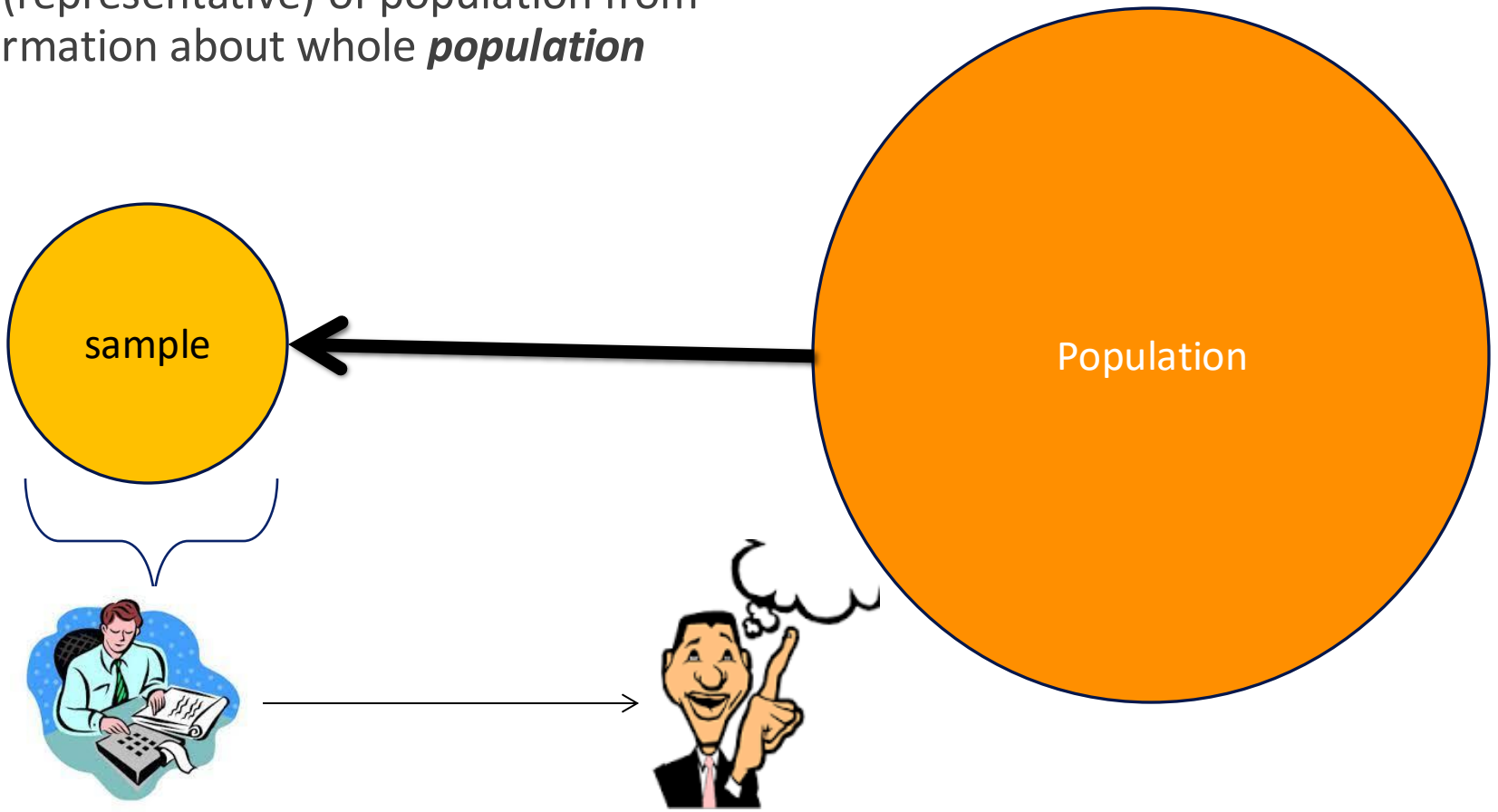
Population: A complete set of individual (objects)

Sample aka sample-data
One point in sample-data aka item, point, ~~sample~~

Size of sample : N

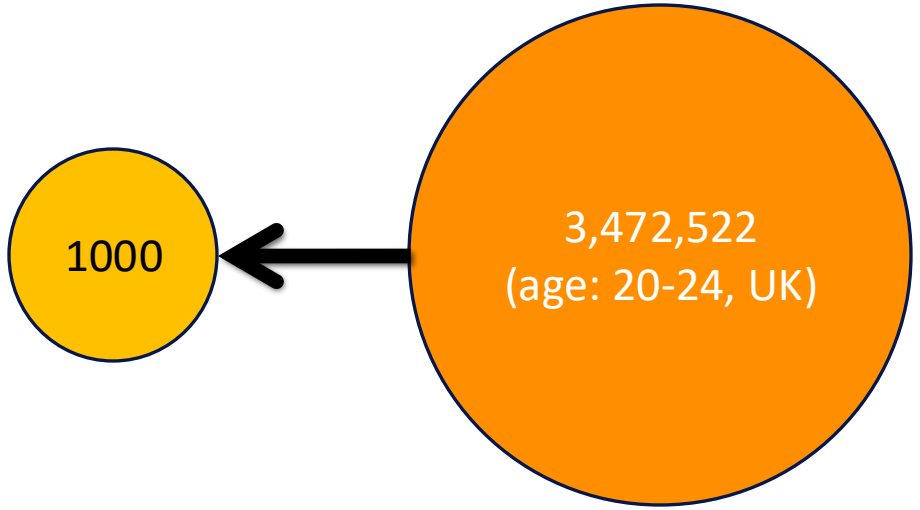
Population & Sample

Sample is a subset (representative) of population from which we infer information about whole ***population***



Examples:

1. Average weight of individual in UK of age 20-24.
2. Average marks of med. student in first year at Imperial.
3. Average Heart Rate of patients with High Blood Pressure in UK
4. Average height of Italian men

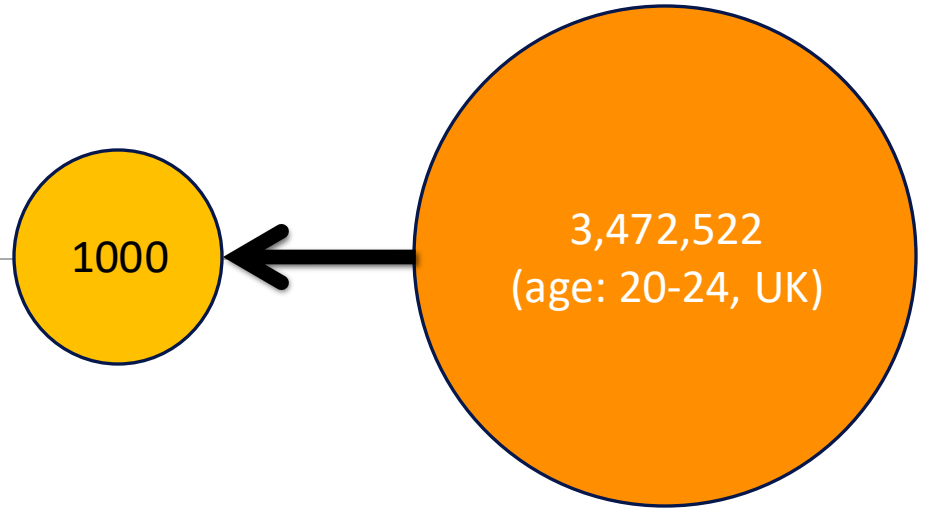


Q: Why take sample, why not entire population?

Q: How to sample (extract points/items from population)

Q: How many points (samples, items) are required in a sample

Estimation of parameter



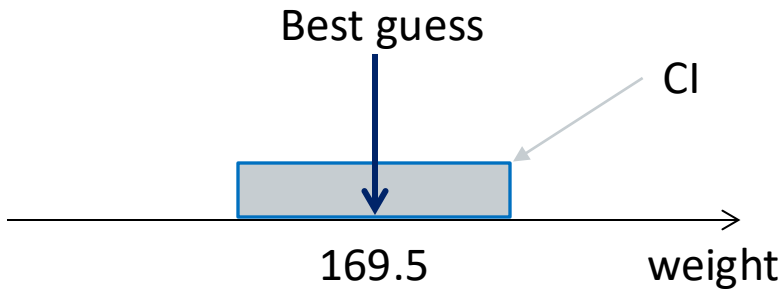
Mean weight of population

Best guess:

- Sample mean is estimation of population mean
- Uncertainty of this estimation (not exact value)

Accuracy of best guess

- Standard Error (SE)

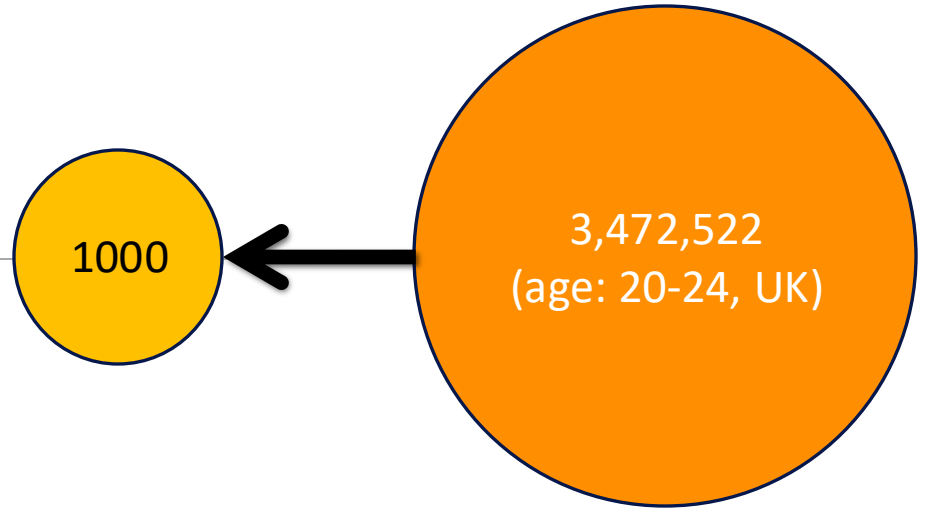


CI for the mean weight	
Mean	95%CI
169.5	163.7 – 174.2

Plausible values for true unknown

- Confidence Interval (CI)

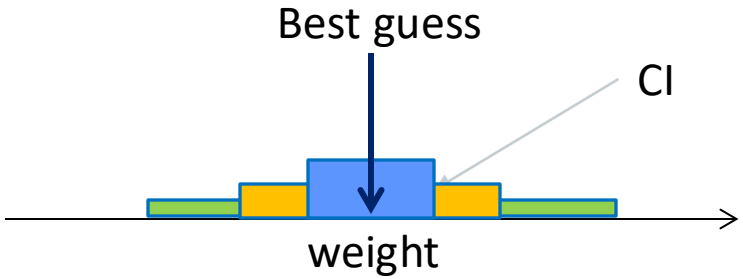
Estimation of parameter



Which colour-bar is the **90% CI**?

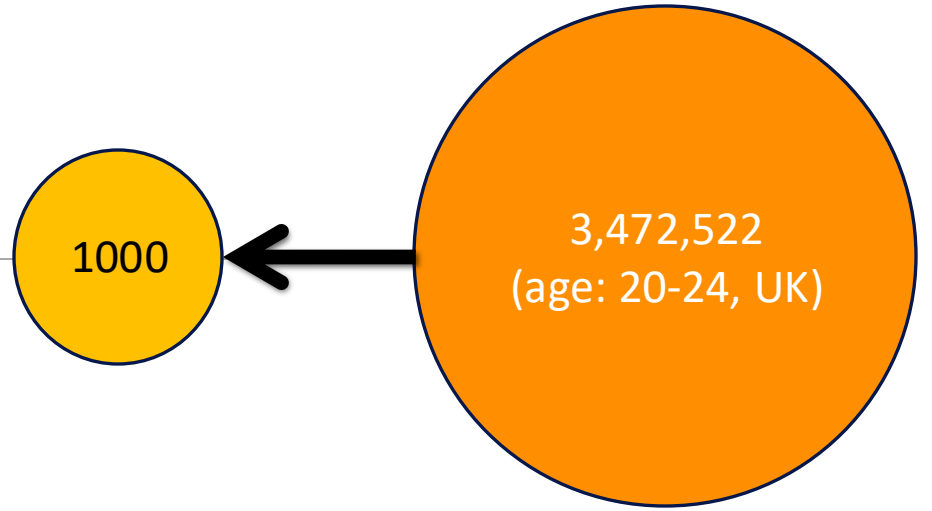
Which colour-bar is the **95% CI**?

Which colour-bar is the **99% CI**?



CI for the mean weight			
Mean	90%CI	95%CI	99%CI
169.5	165.5-173.4	163.7 – 174.2	163.0-175.9

Estimation of parameter



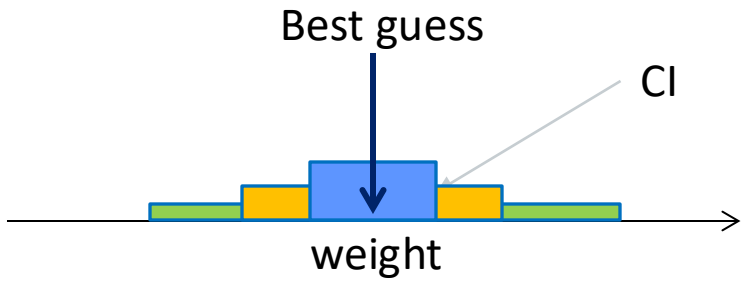
Mean weight of population

Best guess:

- Sample mean weight of is estimation of population mean weight
- Uncertainty of this estimation (not exact value)

Accuracy of best guess

-Standard Error (SE)



Plausible values for true unknown

- Confidence Interval (CI)

CI for the mean weight			
Mean	90%CI	95%CI	99%CI
169.5	165.5-173.4	163.7 – 174.2	163.0-175.9

Standard Error and CI

Standard Error:

$$SE = \frac{SD\ of\ population}{\sqrt{sample\ size}} \approx \frac{SD\ of\ sample}{\sqrt{sample\ size}}$$

95% Confidence Interval

$$95\%\ CI = sample\ mean \pm 1.96 \times SE$$

Means of all (hypothetical) samples follow normal distribution and 95% of them lie within mean ± 1.96xSE

Standard Deviation Vs Standard Error

SD:

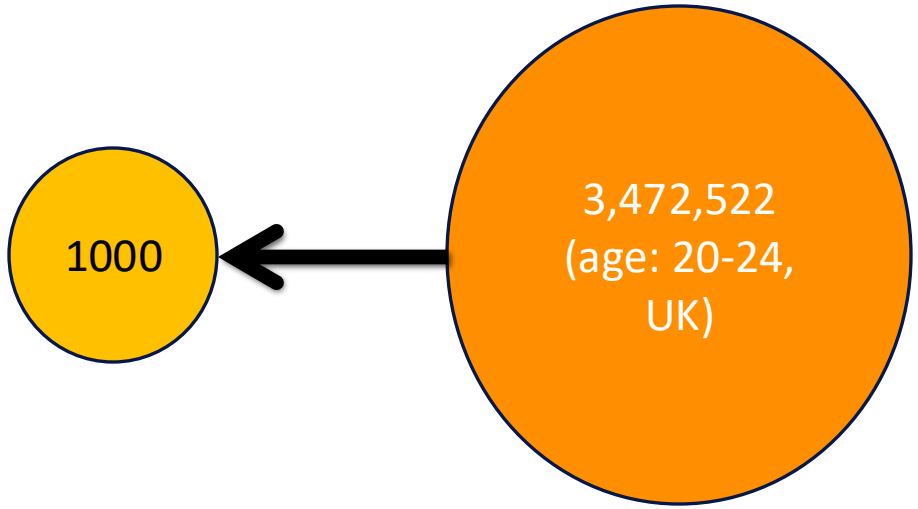
- measure of spread/variability of data
- descriptive statistics
- for normally distributed data, 2SD includes 95% of observed values

SE:

- accuracy of estimation of population
- inferential statistics
- for 95% CI, hypothesis testing etc
- range of values likely to include true population parameter

When it can be misleading?

- Sample is not **representative** of population (validity)
- **Not large enough** data (accuracy)



Lab

Let’s compute – Lab session

Data

Sample mean

Standard deviation

Standard error

Confidence interval

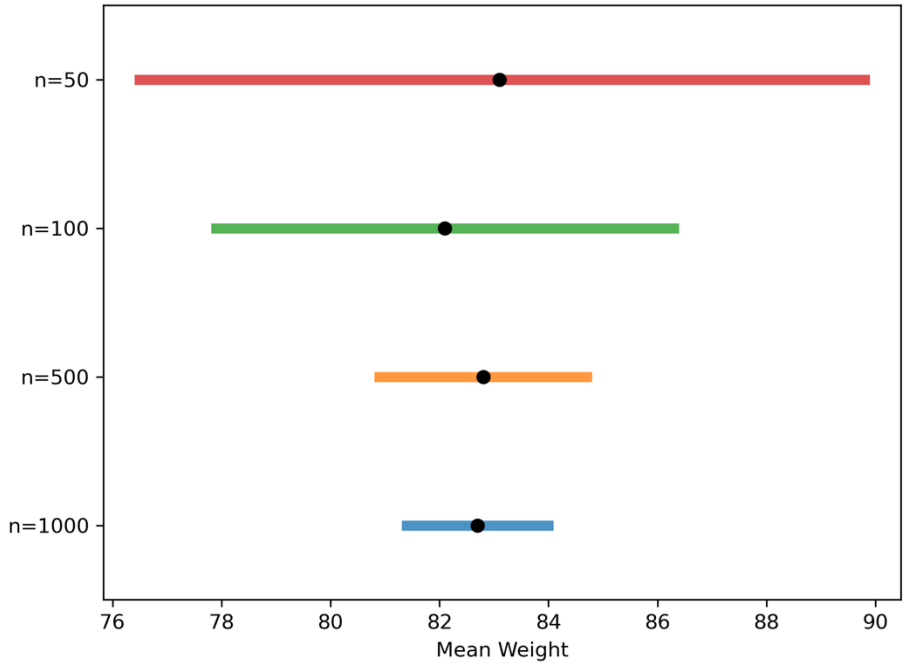
Effect of sample size: Example 1

Weight:

Sample mean = 81.4 kg, Standard Deviation = 21.4kg, n=1000

SE = ?, CI= ?

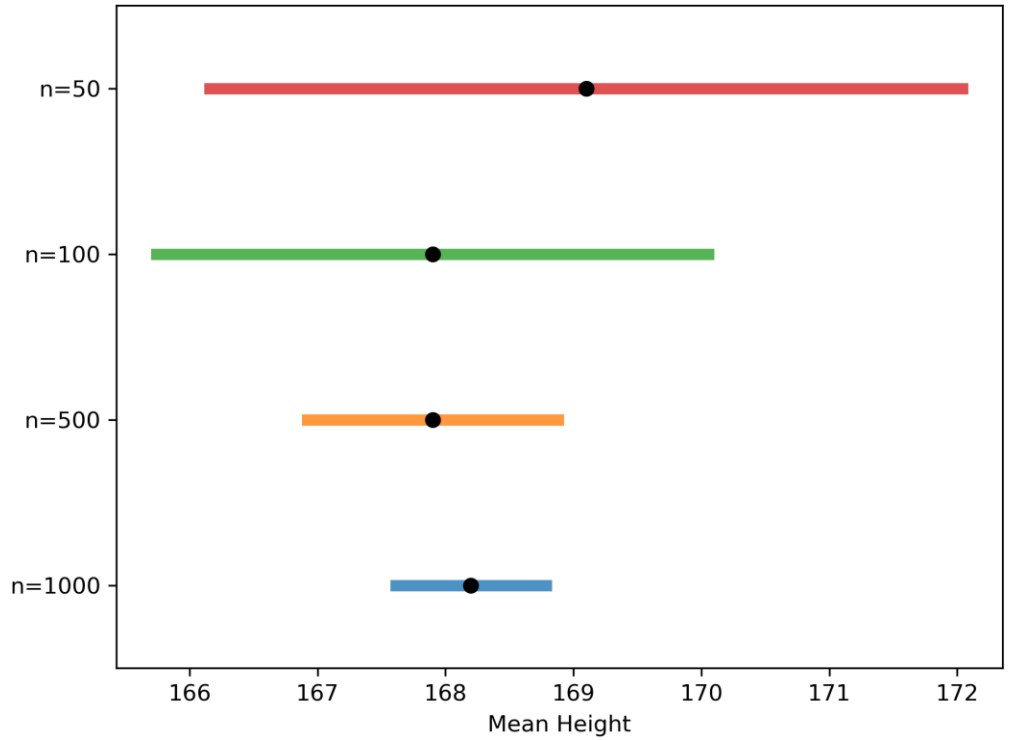
N of sample	Weight (mean)	Weight (SE)	95%CI
50	83.1	3.3	76.5-89.8
100	82.1	2.1	77.9-86.3
500	82.8	0.96	80.9-84.7
1000	82.7	0.68	81.4-84.0



Effect of sample size: Example 2

Height:

N	Mean	SE	CI
50	169.1	1.5	166.16-172.04
100	167.9	1.1	165.744-170.056
500	167.9	0.5	166.92-168.88
1000	168.2	0.3	167.612-168.788

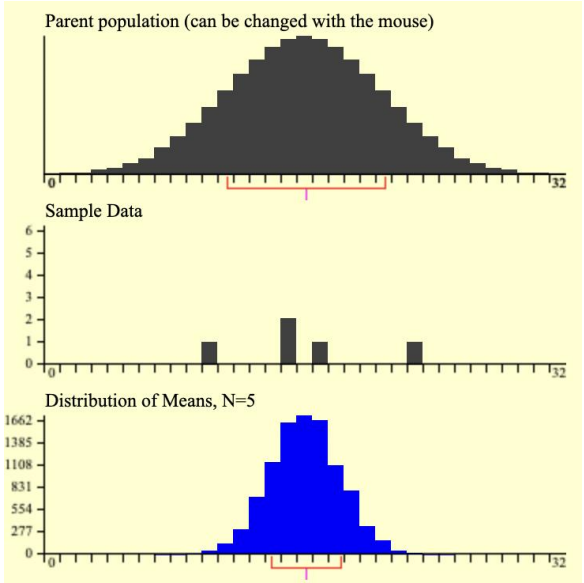


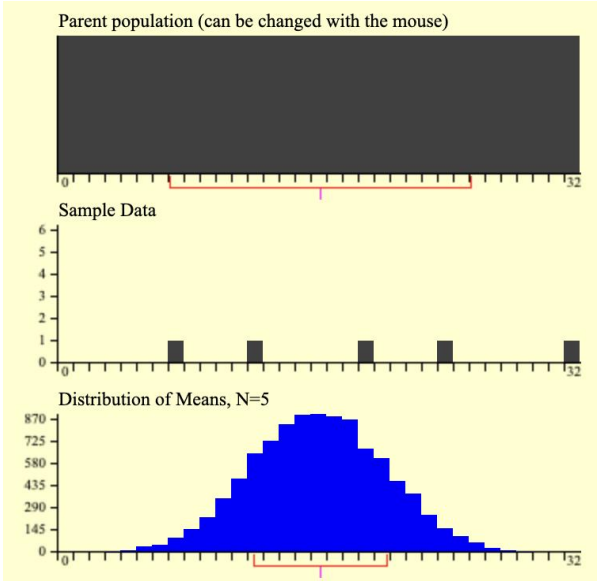
Normal Distribution $X \sim N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$

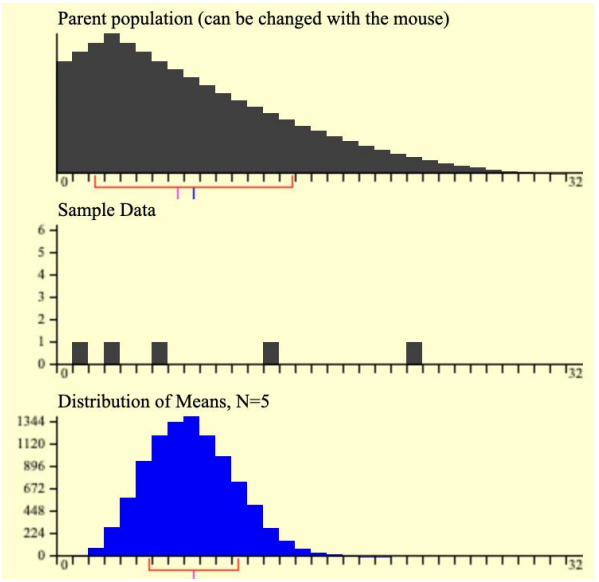
Distribution of means

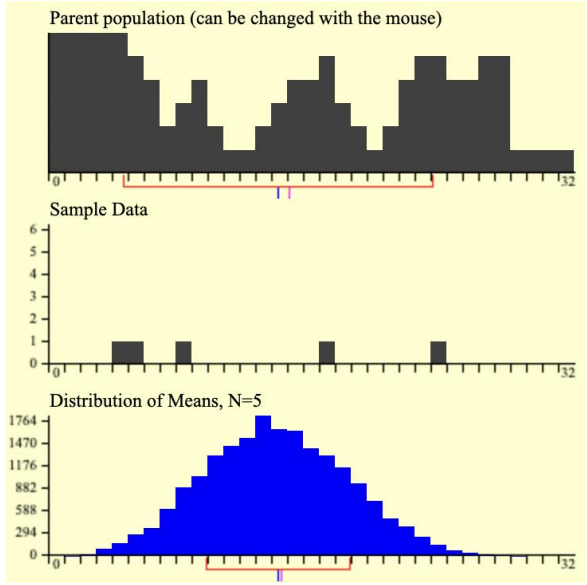
Central limit theorem

$$\tilde{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$







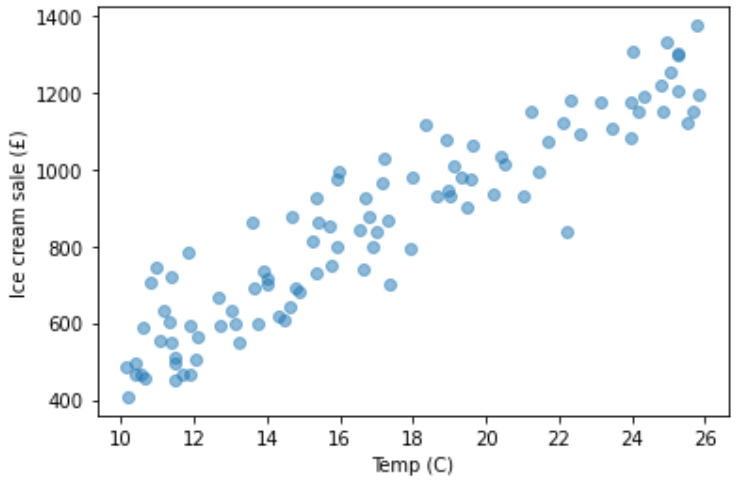


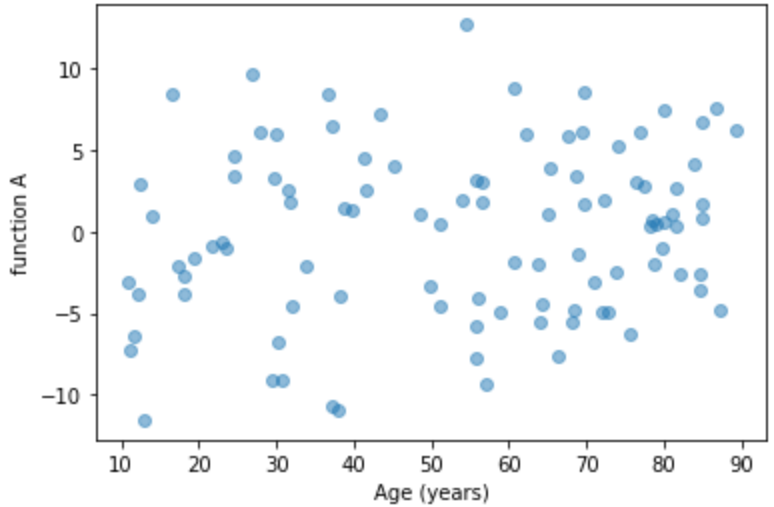
Association between two variables

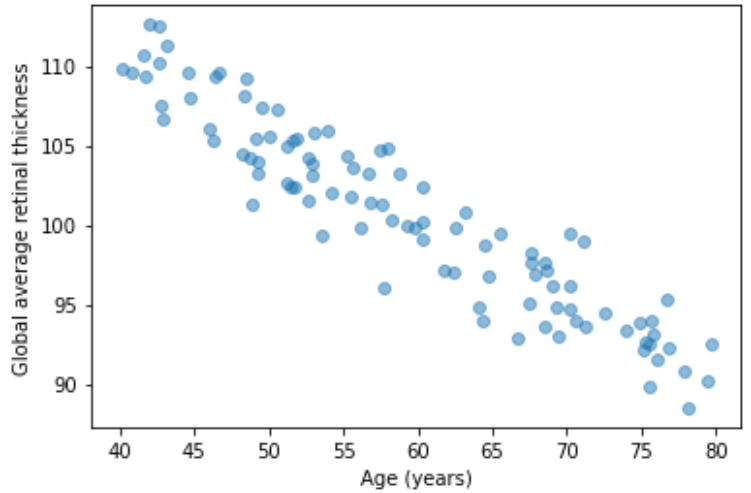
Correlation

- Investigate a relationship between two independent variables (i.e. x and y)
- Does x increase as y or vice-versa?
- Is relation linear?

One simple way is to plot scatter graph and see



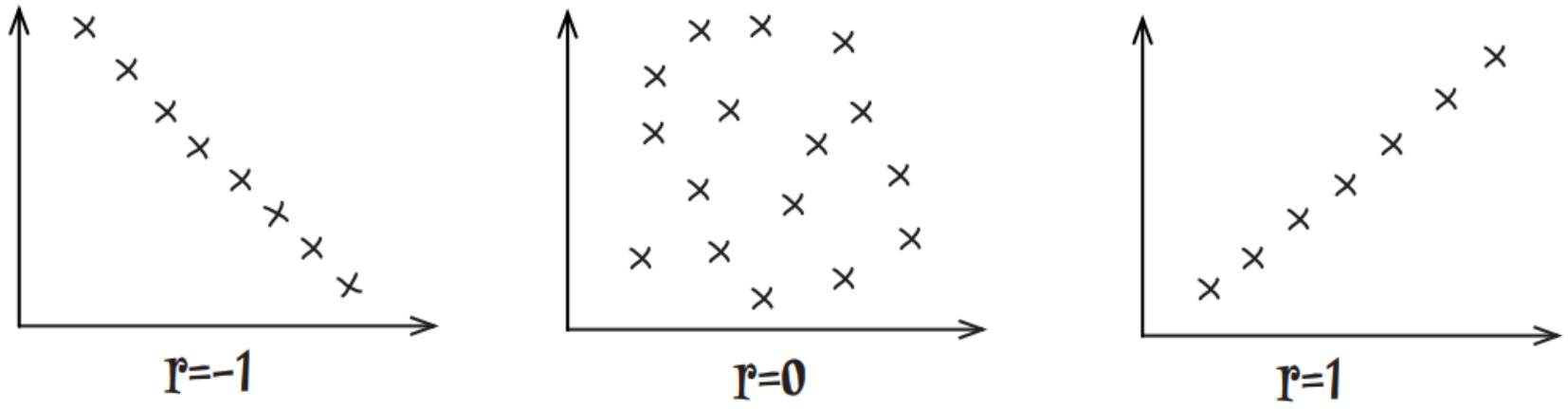




Quantifying: Pearson Correlation

Pearson Correlation Coefficient r or ρ (rho)

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$



Ref: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Pearson Correlation: Example

Example 1:

x = [1, 3, 5,]

y = [1, 2, 3]

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Example 2:

x = [1, 3, 5,]

y = [3, 2, 1]

$\overline{x} =$

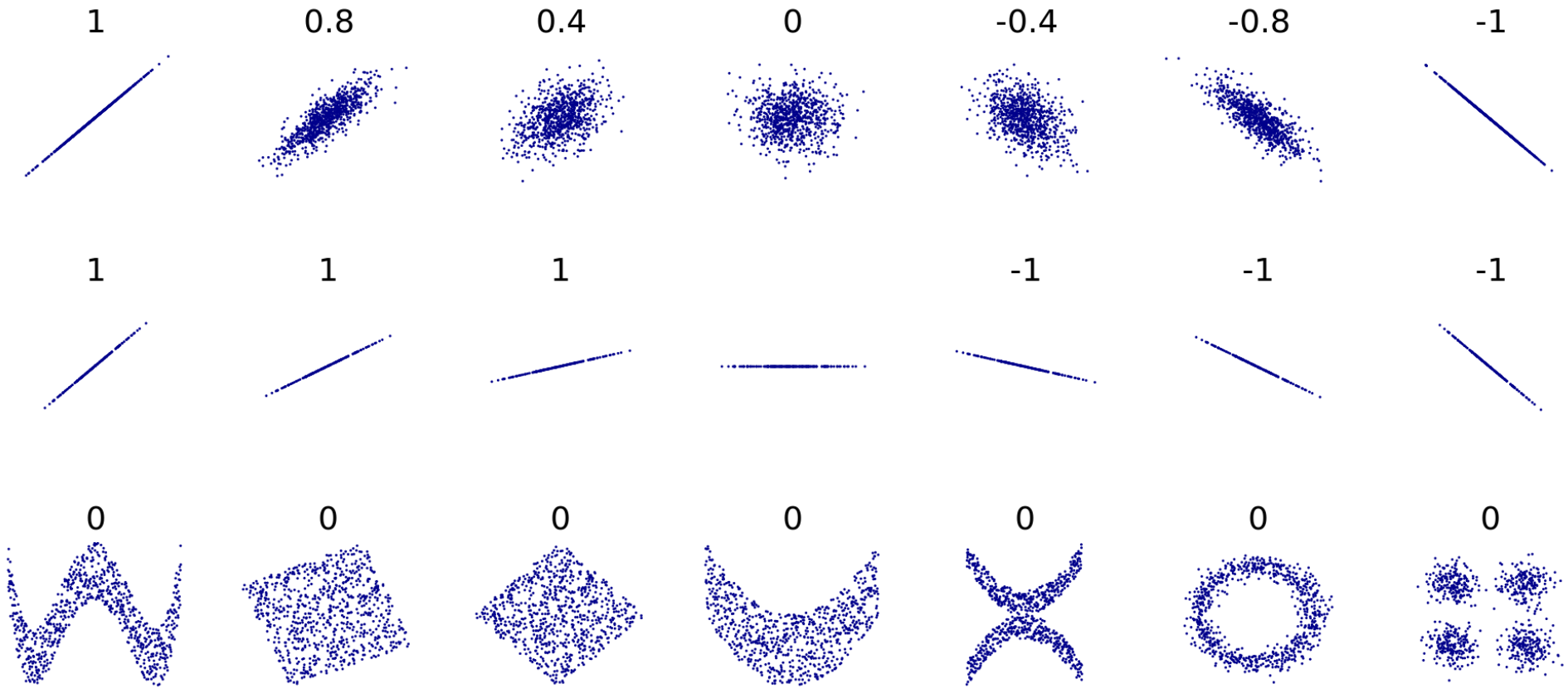
$\overline{y} =$

$\sum(x - \overline{x})^2 =$

$\sum(y - \overline{y})^2 =$

$r =$

Pearson Correlation Coefficient



https://en.wikipedia.org/wiki/Correlation

Correlation

Pearson Correlation Coefficient

Parametric test

- x, y: normally distributed
- linear relationship

Generally:

- $|r| < 0.4$ → weak
- $0.4 < |r| < 0.7$ → moderate
- $0.7 < |r|$ → strong

r = 0 , no ***linear*** relationship

Spearman Rank correlation

Non-parametric test

Based on ranks rather than exact values

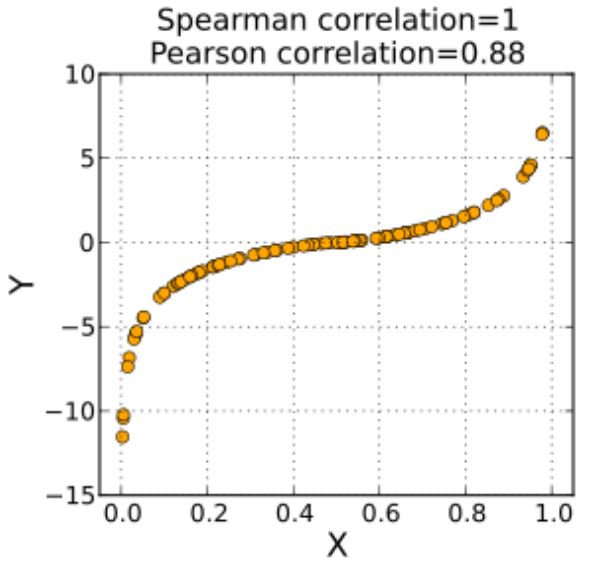


Image: https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

Correlation and P-value

P-value can be obtained from correlation with

Null Hypothesis $H0 : r = 0$

Alternative Hypothesis $H1 : r \neq 0$

P-value tells us the probability of getting high correlation between x and y by pure chance

Examples

1. BMI vs Age, r= 0.13, p-value =0.08
2. BMI vs Age, r= 0.13, p-value =0.04
3. lung function before vs after exercise, r= 0.93, p-value = 0.001

Correlation

A strong correlation between x and y does **Not mean**

- x causes y : X → Y
- y causes x : Y → X
- x and y are caused by one or more other variables z: Z→X, Z → Y

Correlation is not causation

Interacting web-widgets

Sampling

https://nikeshbajaj.github.io/teaching/demos/Stats/sampling

Others

https://onlinestatbook.com/stat_sim/sampling_dist/index.html

https://onlinestatbook.com/2/index.html

Queen Mary

University of London