

# Introduction to Statistics

---

*Nikesh Bajaj, PhD*

*Lecturer in Data Science, Queen Mary University of London*

*Honorary Research Associate, Imperial Collage London*

[n.bajaj@{qmul,imperial}.ac.uk](mailto:n.bajaj@{qmul,imperial}.ac.uk)

<https://nikeshbajaj.in>

# Overview

---

## Session 1

- Describe your data
  - Descriptive statistics - summarising the data
  - Visualisation (plots and figures)
- Inferential analysis:
  - Inference about population from sample
- Given two groups of data
  - Test the differences between groups (**hypothesis testing**)
  - Test the relationship between two variables (correlation)

## Session 2

- Lab Practice using SPSS

# Introduction to Statistics

---

## Part 1: Describe your Data

*Nikesh Bajaj, PhD*

*Lecturer in Data Science, Queen Mary University of London*

*Honorary Research Associate, Imperial Collage London*

[n.bajaj@{qmul,imperial}.ac.uk](mailto:n.bajaj@{qmul,imperial}.ac.uk)

<https://nikeshbajaj.in>

# Describe your data

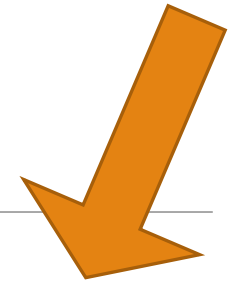
---

# 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?

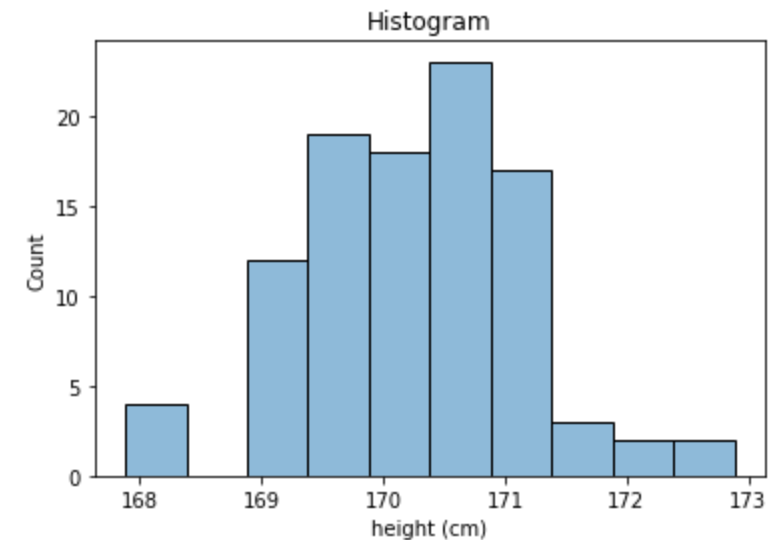
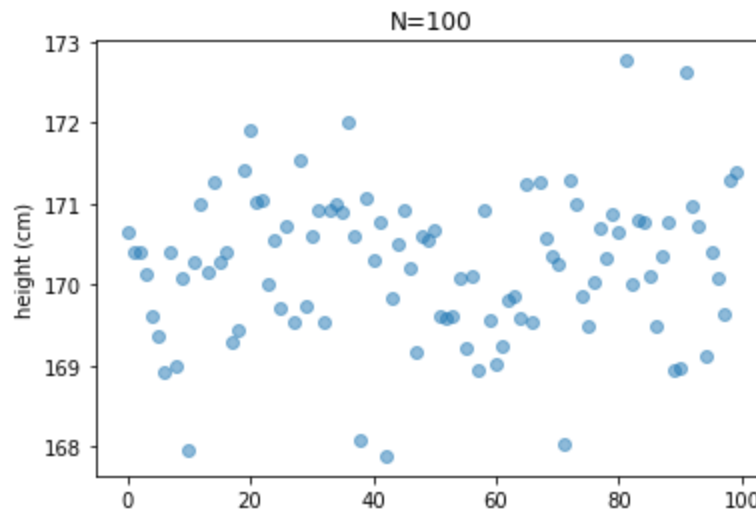


N=9

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

N=100

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

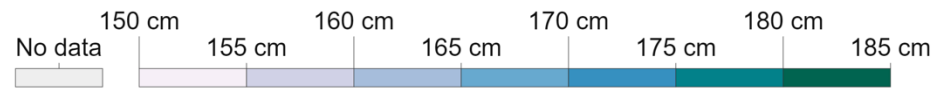
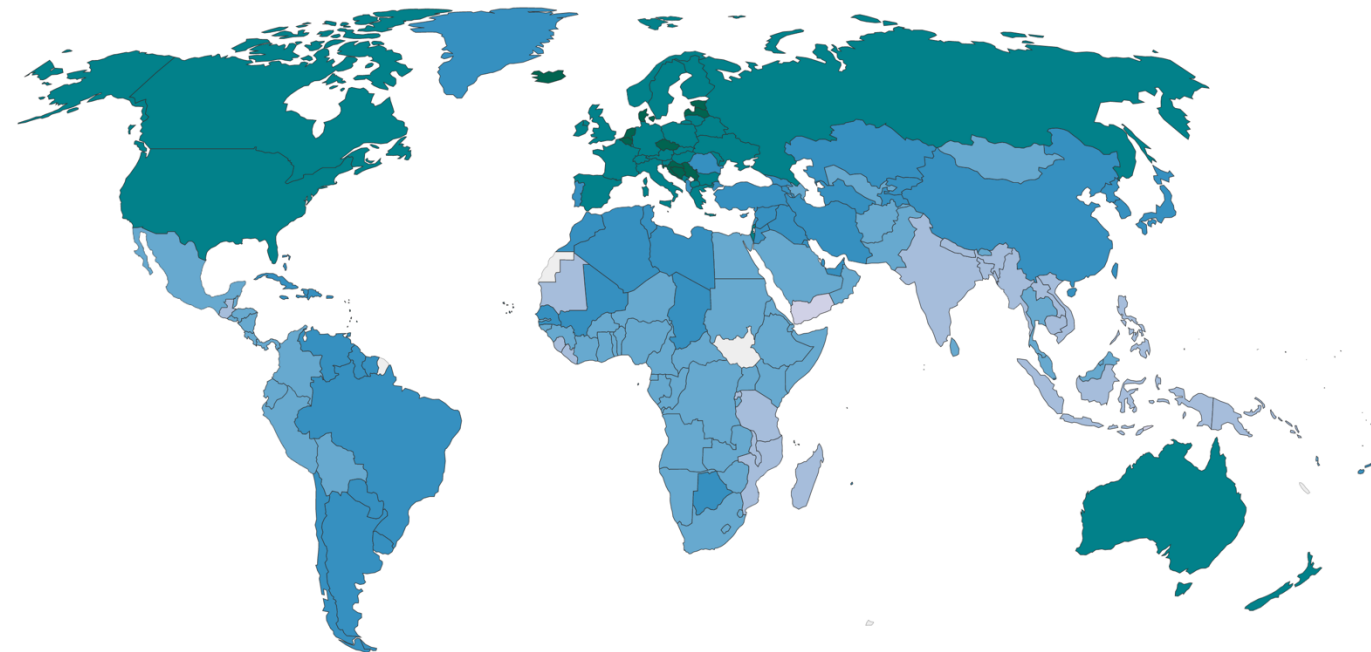


Average – centre tendency of data  
Variability – spread of data  
Nice plots!

# 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.



Source: NCD RisC, Human Height (2017)

OurWorldInData.org/human-height • CC BY

# Types of Variable

## Numerical






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

## Categorical

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

## Operators (where?)

- +, -, X
- >, <
- =, ≠

| ■ Category or Qualitative Data   | Numerical or Quantitative Data  |
|--|---|
|  <p>breed of dog</p>     |  <p>weight</p>  |
|  <p>type of dessert</p> |  <p>time</p>    |
|  |  <p>length</p> |

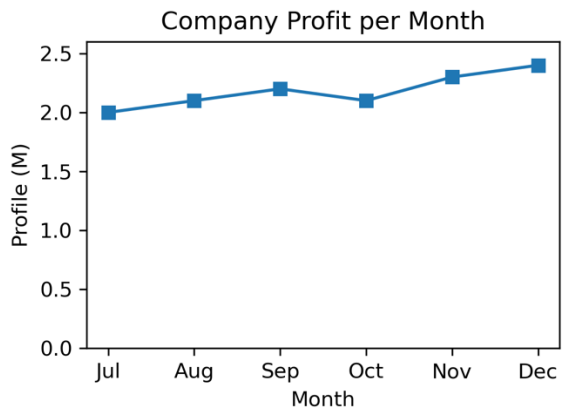


# Visualisation – just plotting the data

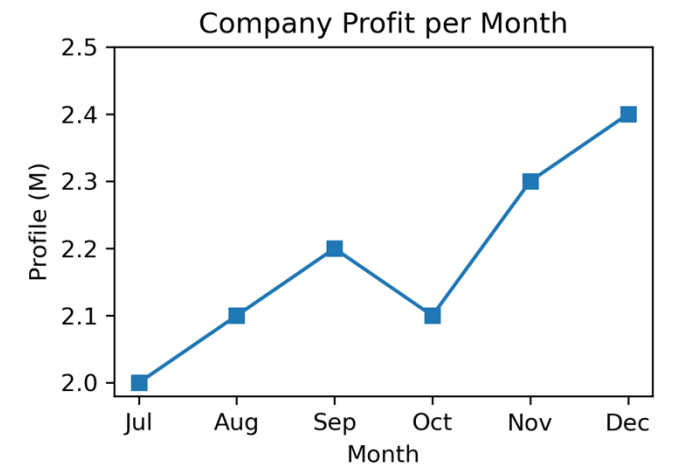
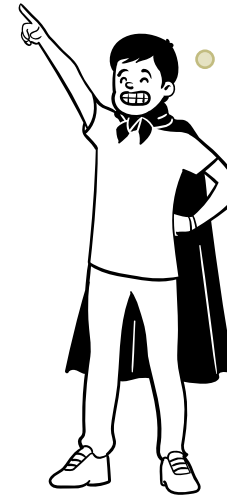
Impression of visualization: Profit of company

| 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!



# Plot & type of variable

## Numerical

- Quantitative: (height)

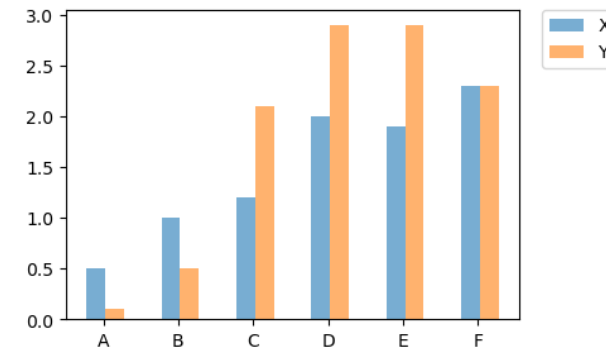
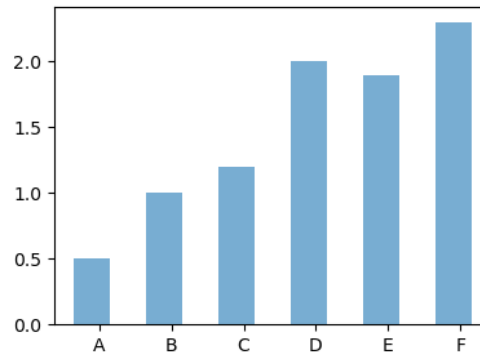
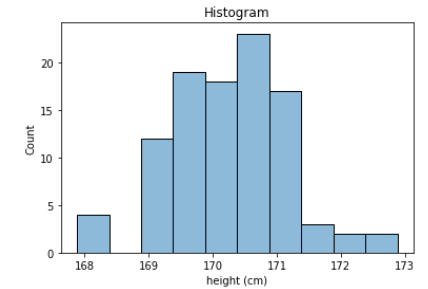
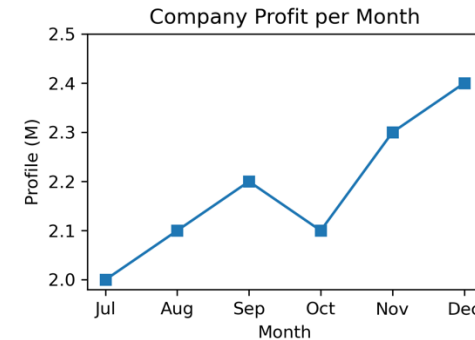
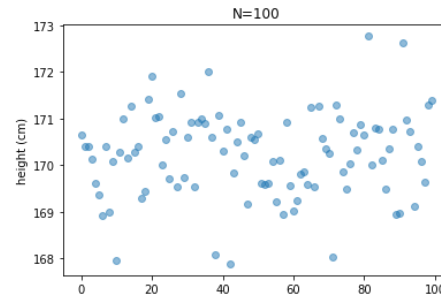
## Categorical

- 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     |



# Descriptive Statistics

---

## Summarizing the data

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

# Average: mean, mode, median

---

Most representative value of data

- 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]

**Mean:** sum of all values/  
number of values

- Preference of drink

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

**Median:** middle value of  
sorted sequence

**Mode :** most frequent value

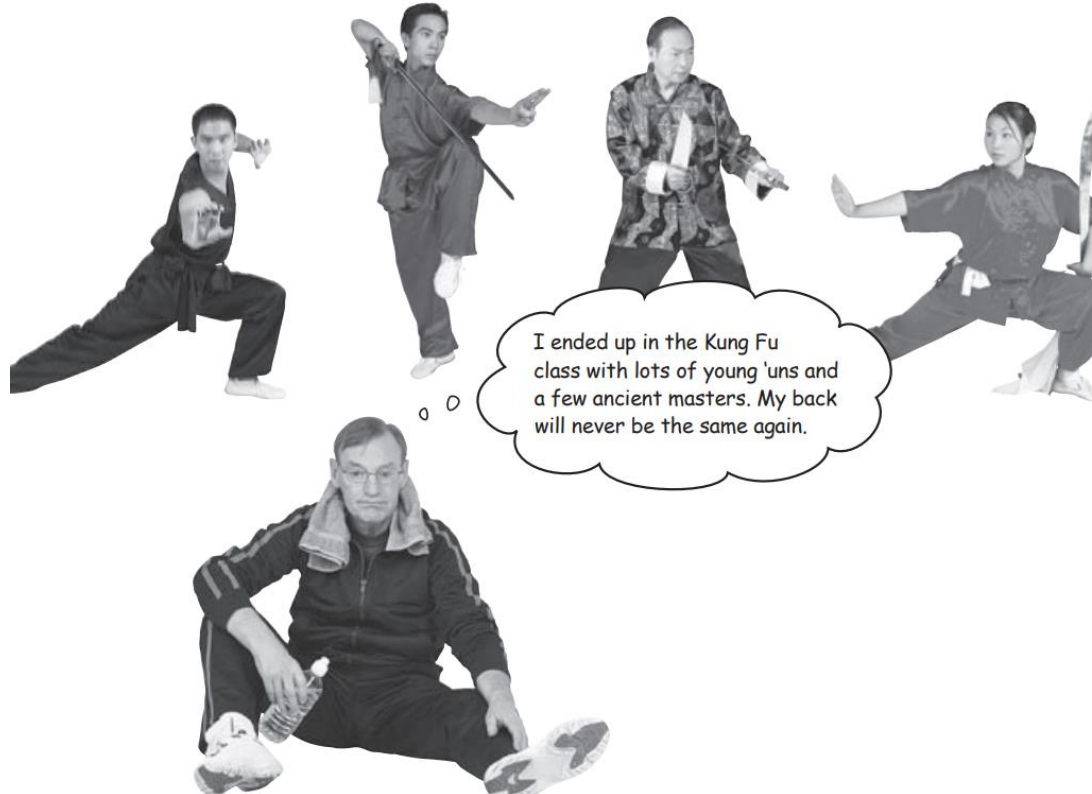
- Age-group

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

# Let's see a case: A Health Club



*“London Health club proud to have class for everybody”*



## ■ Tuesday Evening

**Class**                      **Mean age**

Class 1 :                      17

Class 2 :                      25

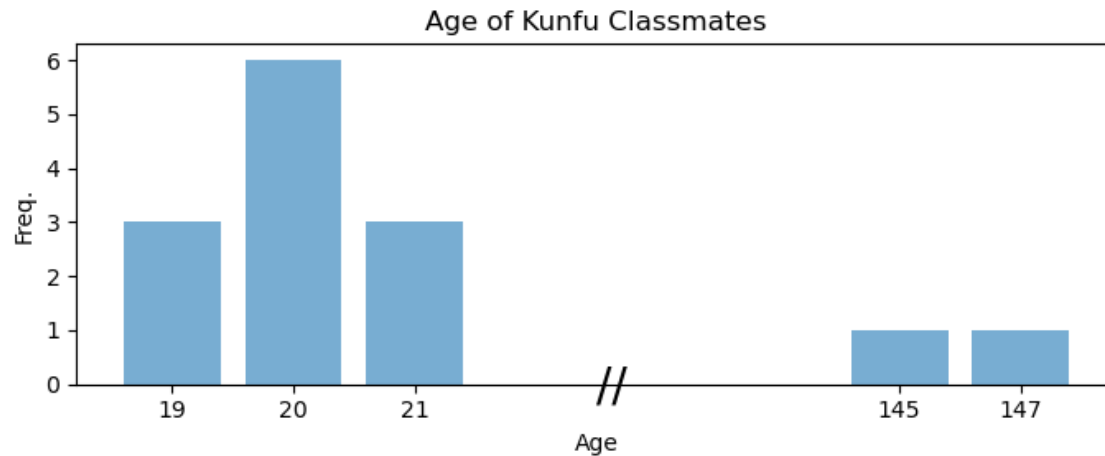
Class 3 :                      38

*Which class new customer should attend??*

New customer in 50s

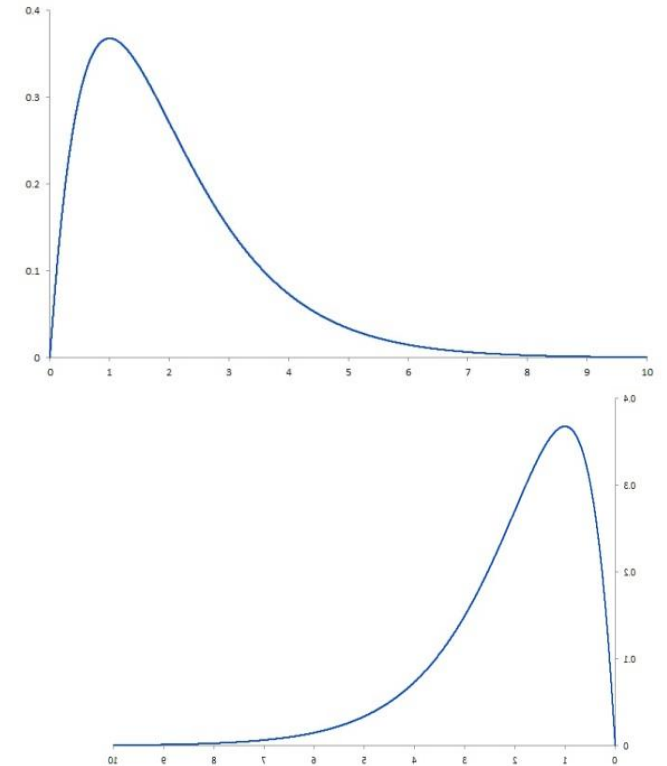
# Health club

|           |    |    |    |     |     |
|-----------|----|----|----|-----|-----|
| 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

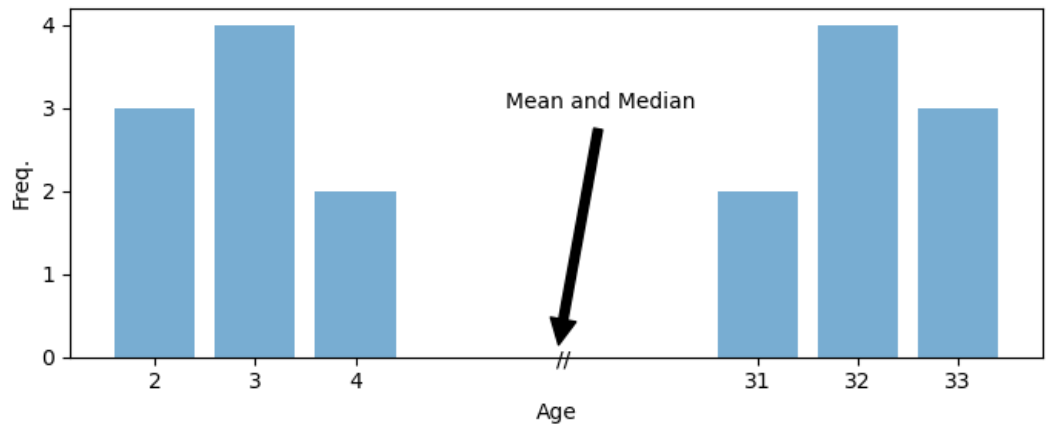
Here's the number in the middle. This is the median, 20.

# Health club



**Swimming Class**  
Median Age: 17

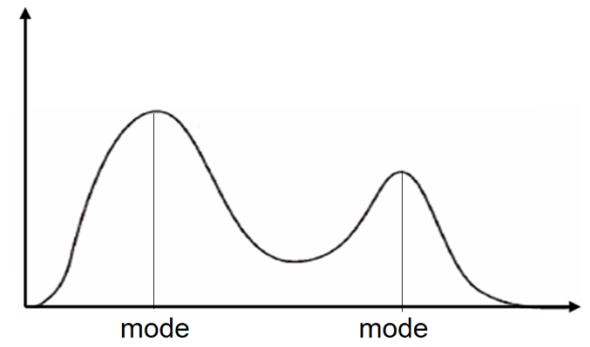
## Can anything go wrong?



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

Mean = 17  
Median = 17

**Mode is our solution**

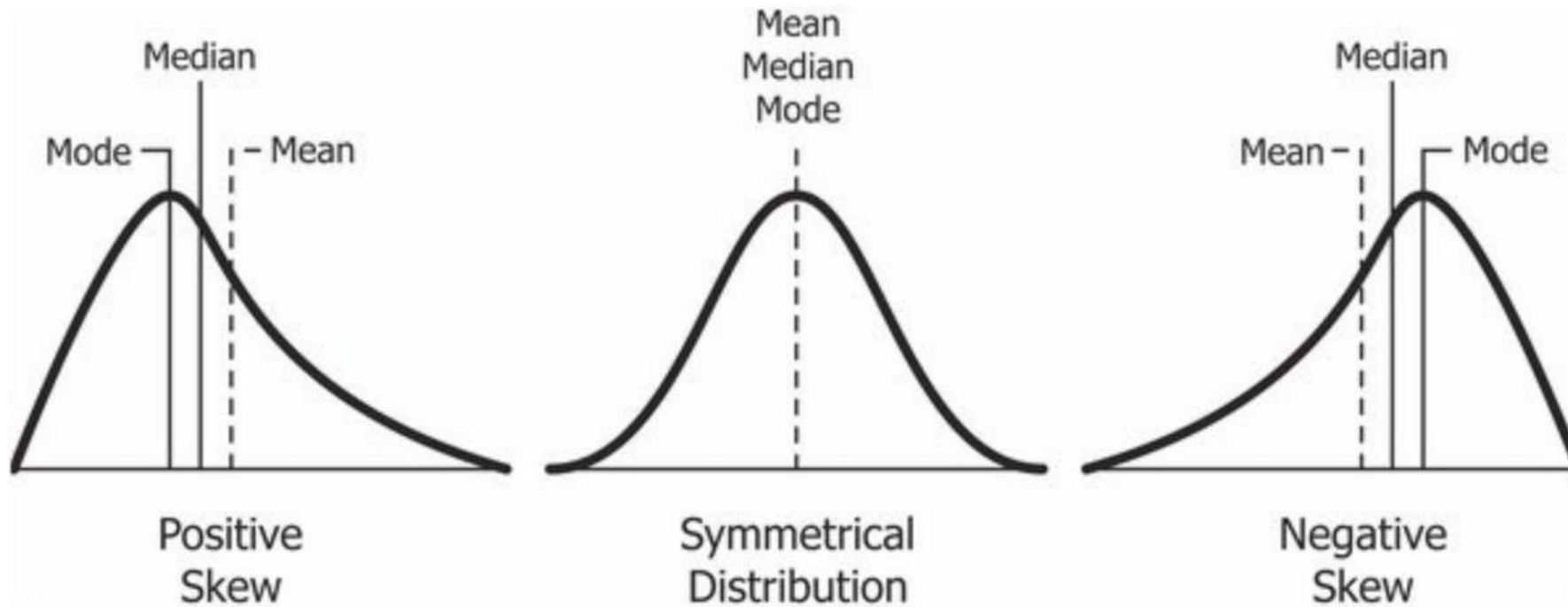


Sounds Cool..  
Sign me up right now!!

teenager

# Skewness

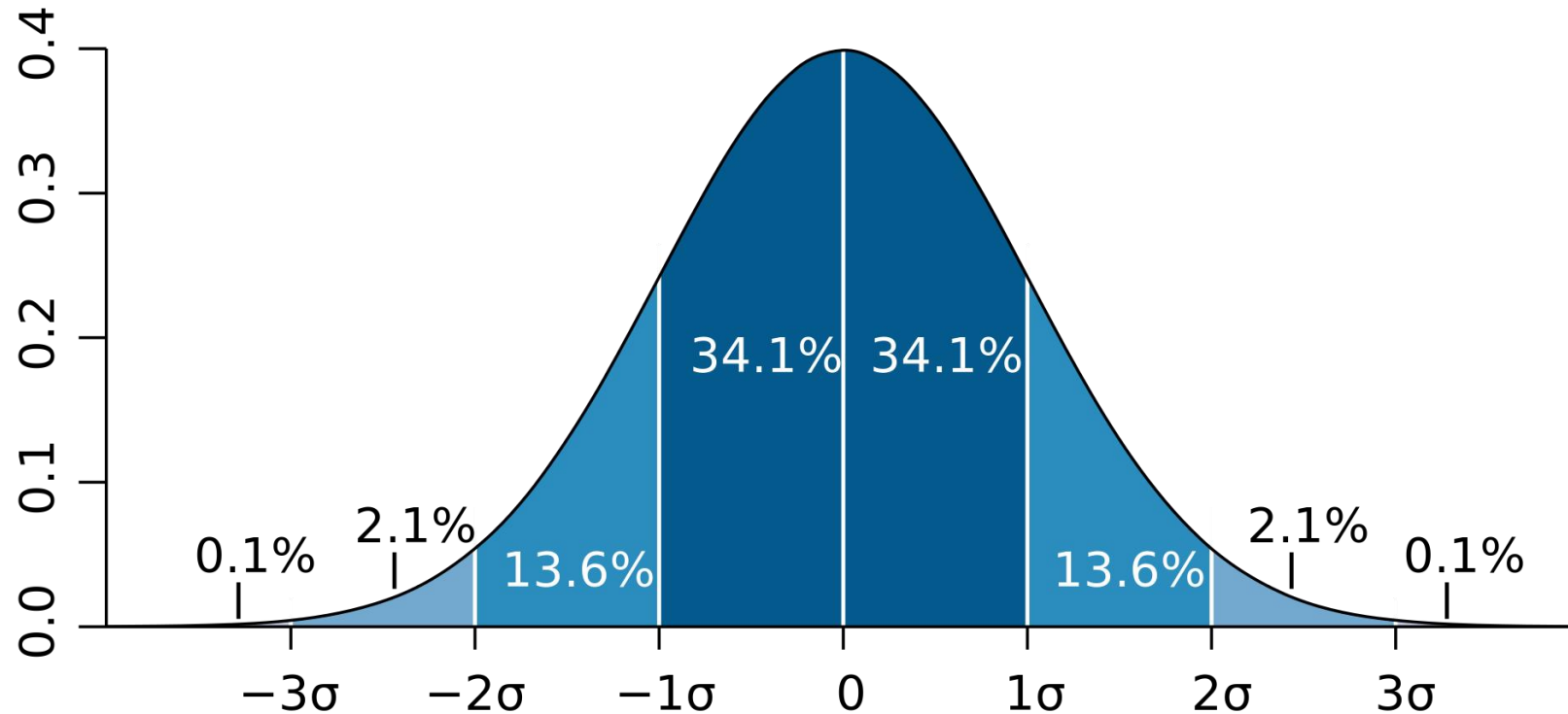
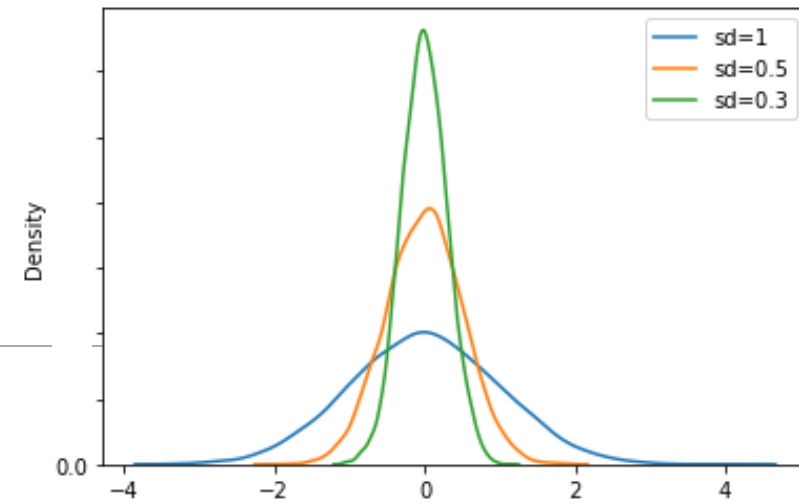
---





# Spread/Variability

Normal distribution



- Standard deviation  $\sigma$  (*sd*)
- Variance  $\sigma^2$  (*var*)
- Range
- Interquartile Range (IQR)

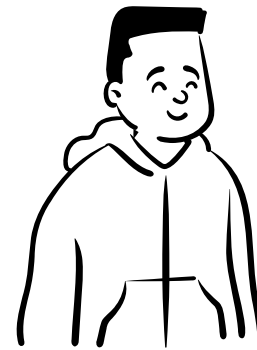
# Question

---

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?

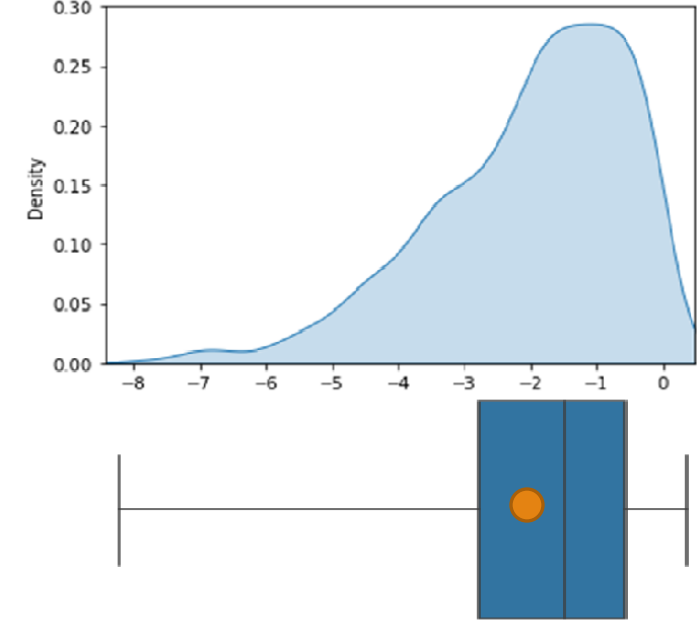
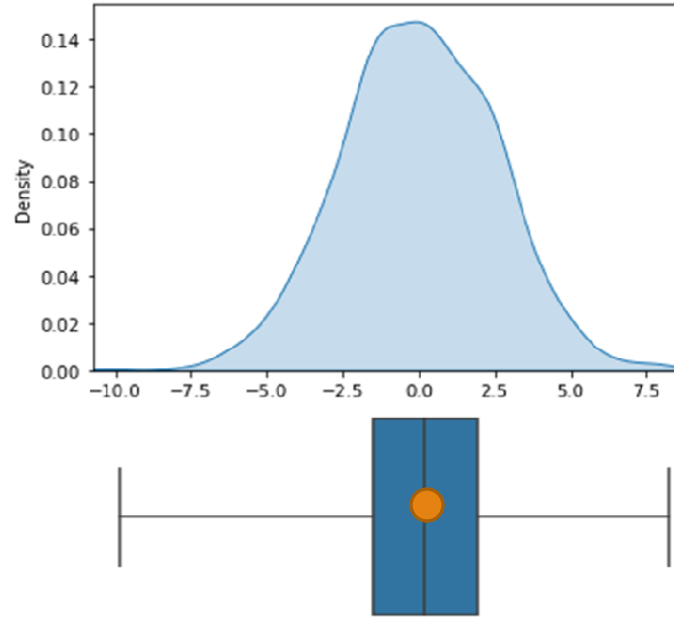
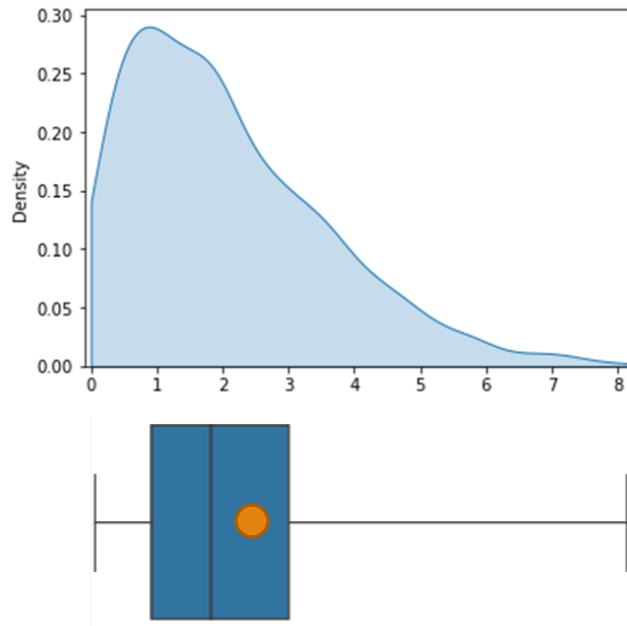
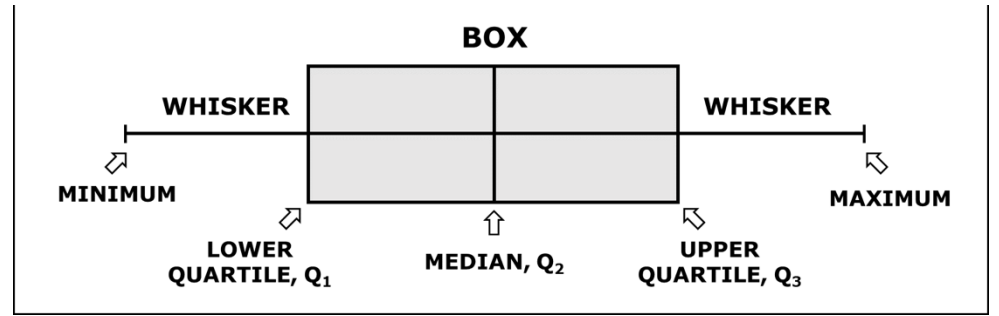


Player 1



Player 2

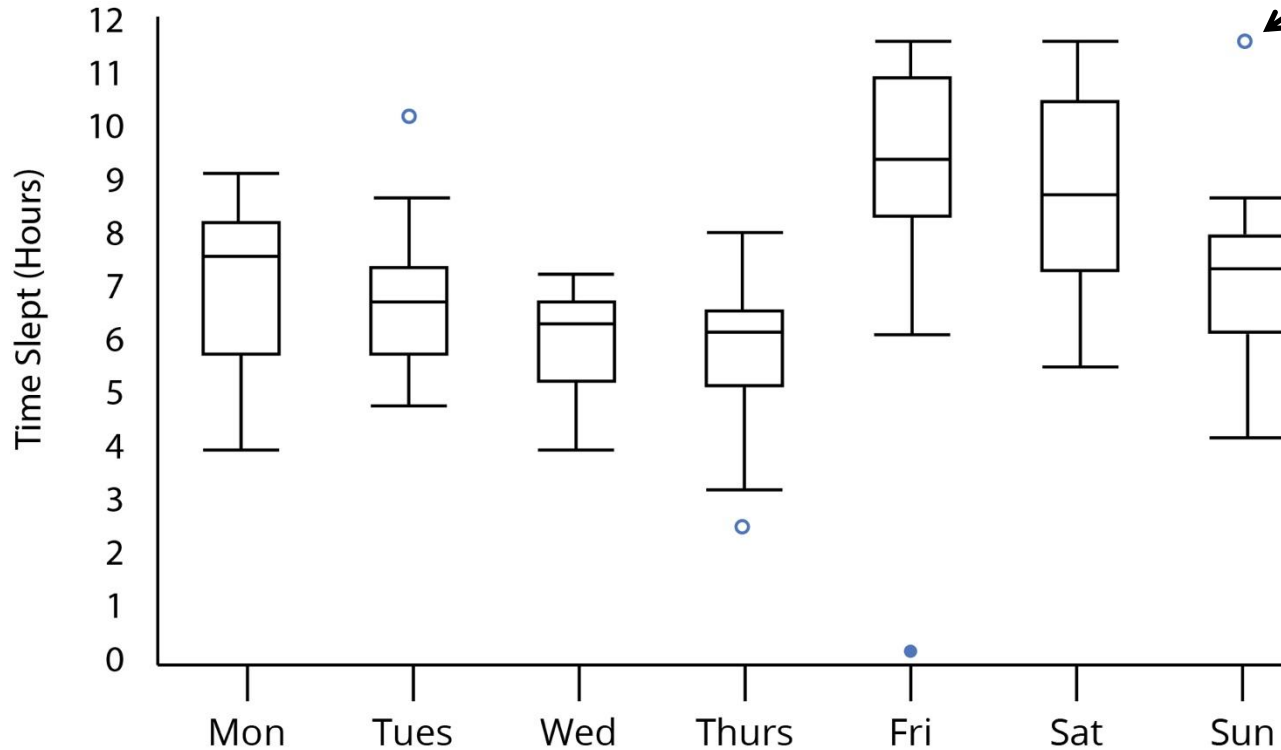
# Box-whisker plot



# Visual comparison with boxplot

Outlier:  
>  $Q3 + 1.5 \times IQR$   
<  $Q1 - 1.5 \times IQR$

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

# Introduction to Statistics

---

## Part 2: Inferential Statistics

*Nikesh Bajaj, PhD*

*Lecturer in Data Science, Queen Mary University of London*

*Honorary Research Associate, Imperial Collage London*

[n.bajaj@{qmul,imperial}.ac.uk](mailto:n.bajaj@{qmul,imperial}.ac.uk)

<https://nikeshbajaj.in>

# Inferential Statistics

---

# 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 & Sample

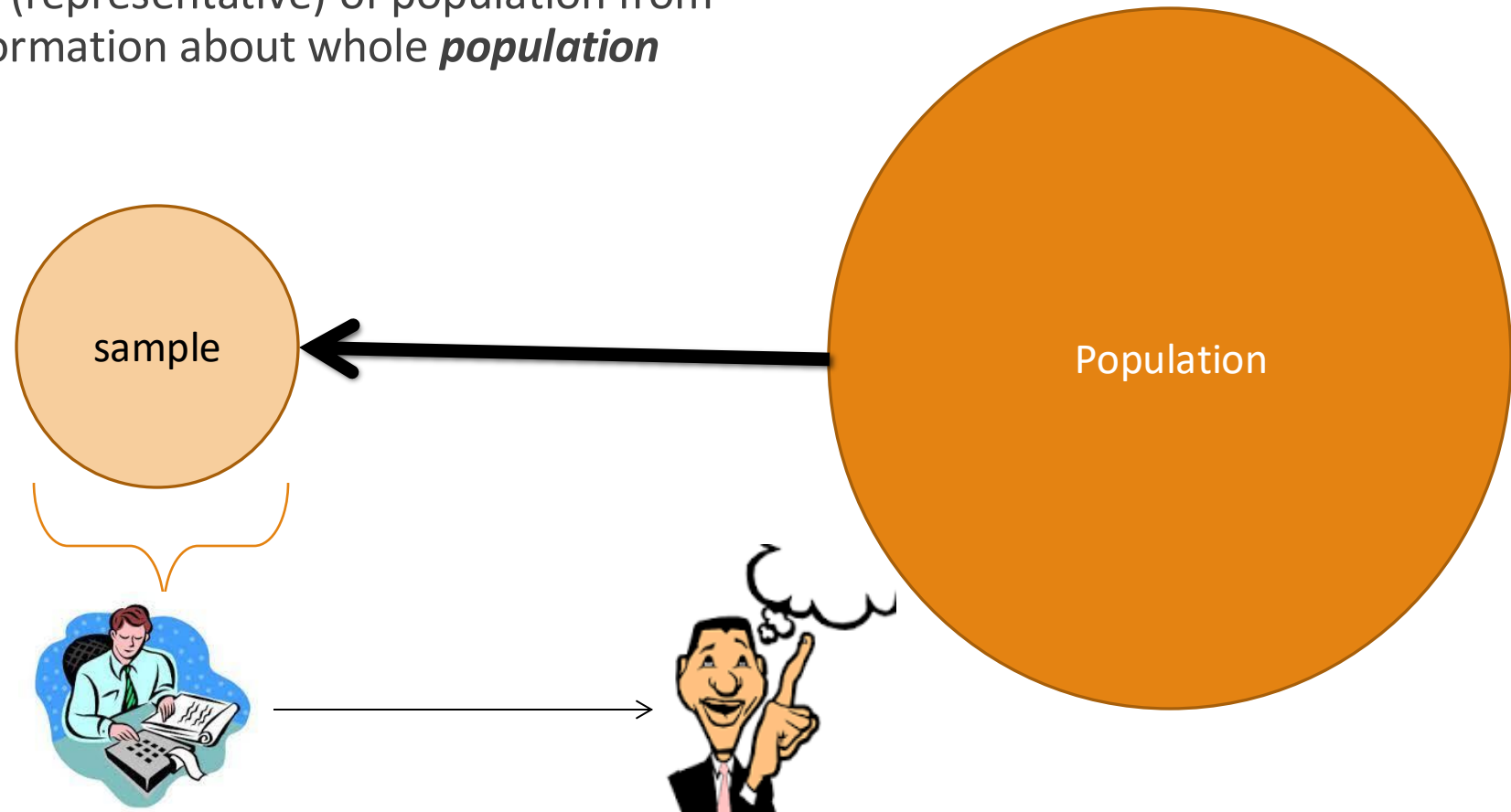
*Population: A complete set of individual (objects)*

*Sample aka sample-data*

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

*Size of sample :  $N$*

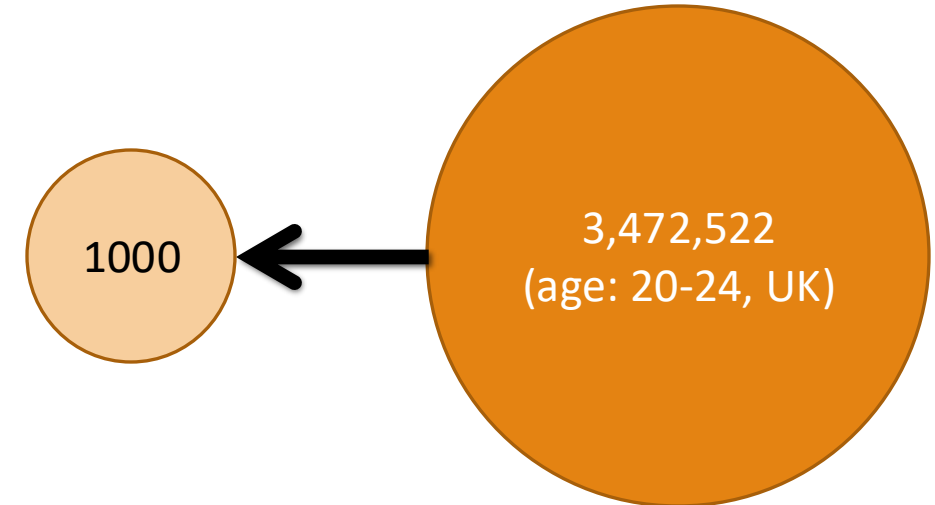
**Sample** is a subset (representative) of population from which we infer information about whole **population**



# 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:

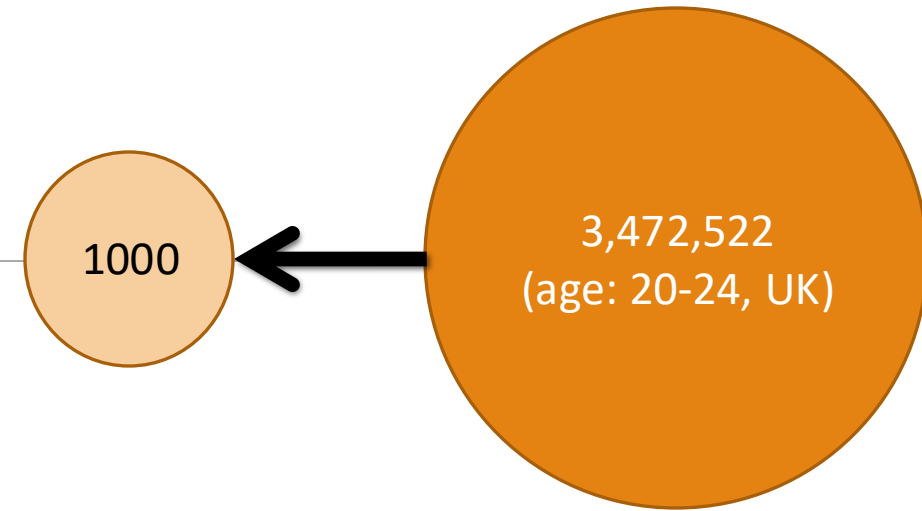
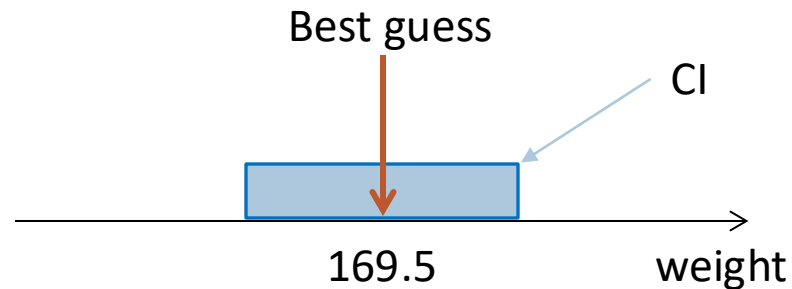
- Sample mean is estimation of population mean
- Uncertainty of this estimation (not exact value)

Accuracy of best guess

- Standard Error (SE)

Plausible values for true unknown

- Confidence Interval (CI)



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

# Estimation of parameter

## Mean weight of population

Best guess:

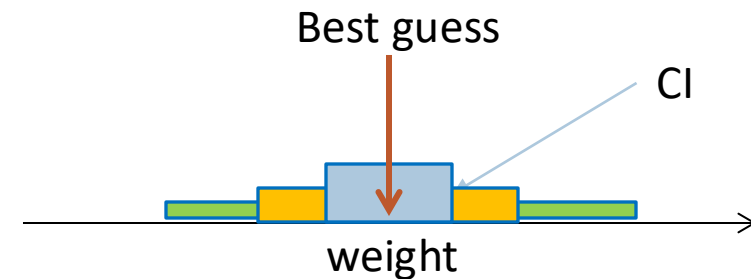
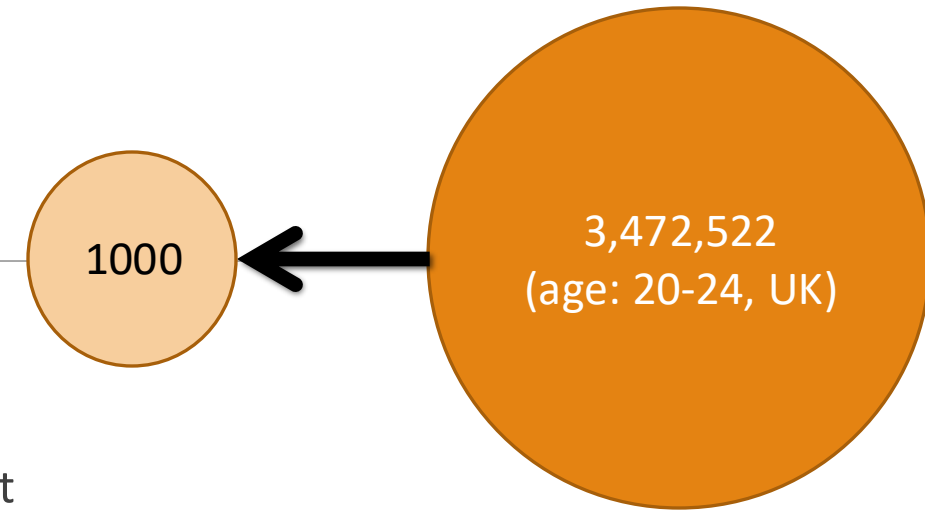
- Sample mean weight of is estimation of population mean weight
- Uncertainty of this estimation (not exact value)

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:

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

95% Confidence Interval

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

***Means of all (hypothetical) samples follow normal distribution and 95% of them lie within mean  $\pm 1.96 \times SE$***

# Standard Deviation Vs Standard Error

---

SD:

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

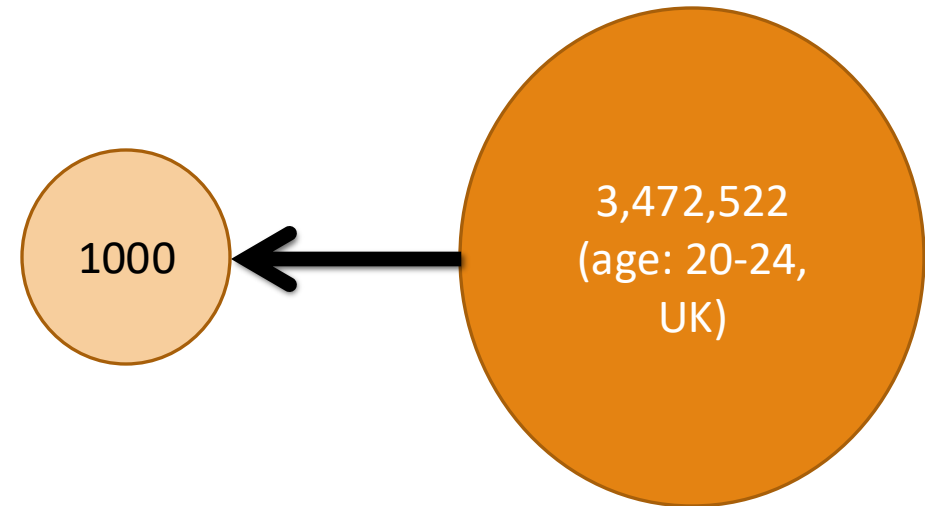
SE:

- 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)



# Let's compute – Lab session

---

Data

Sample mean

Standard deviation

Standard error

Confidence interval



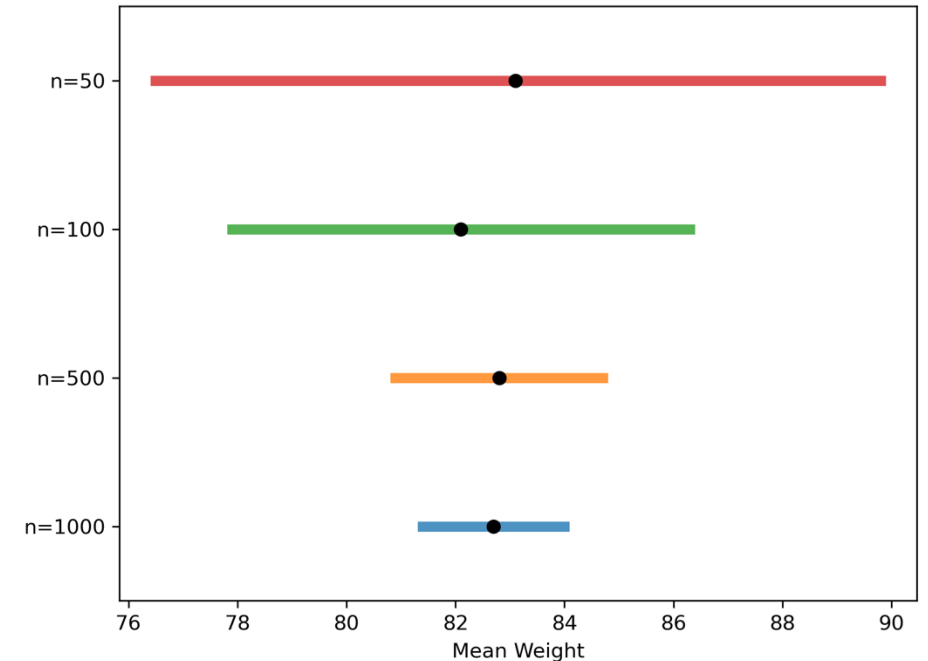
# Effect of sample size: Example 1

## Weight:

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

SE = ?, CI = ?

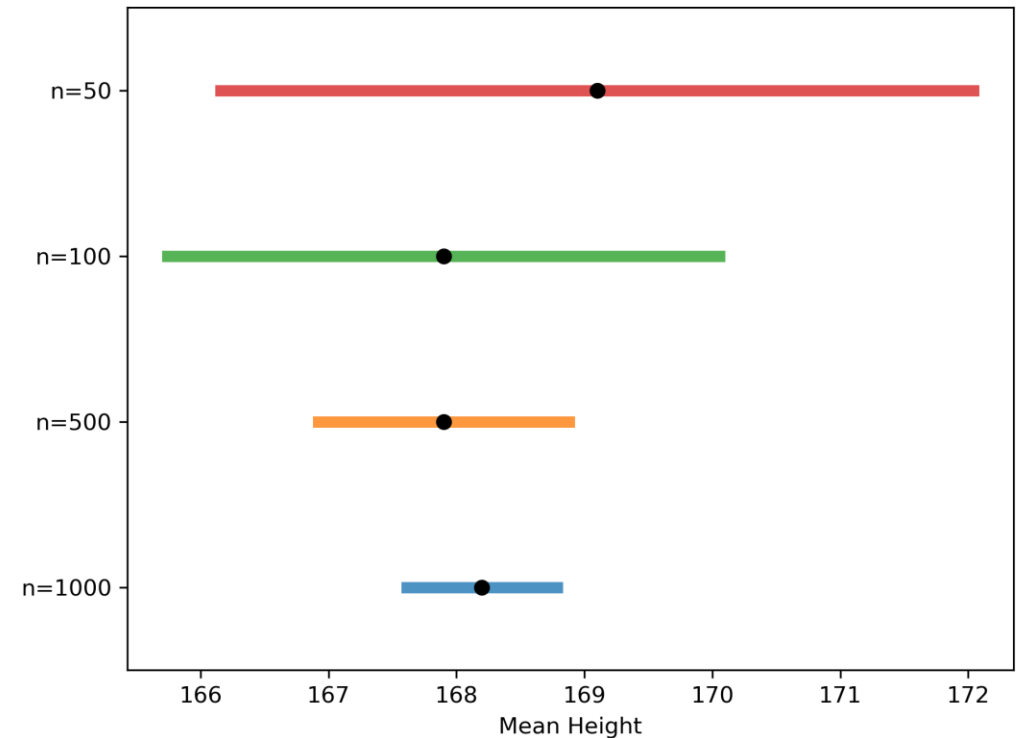
| N of sample | Weight (mean) | Weight (SE) | 95%CI     |
|-------------|---------------|-------------|-----------|
| 50          | 83.1          | 3.3         | 76.5-89.8 |
| 100         | 82.1          | 2.1         | 77.9-86.3 |
| 500         | 82.8          | 0.96        | 80.9-84.7 |
| 1000    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



# Simulation: Sampling

---

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

Others

[https://onlinestatbook.com/stat\\_sim/sampling\\_dist/index.html](https://onlinestatbook.com/stat_sim/sampling_dist/index.html)

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

# Proportion and CI

---

Estimating proportion of population that have particulate condition A.

Find the proportion in sample  $p = \#A/\text{total}$

$$SE \text{ of the proportion } p = \sqrt{\frac{p(1-p)}{n}}$$

$$95\%CI \text{ of the proportion } p = p \pm 1.96 \times SE$$

\* $np > 5$  &  $n(1-p) > 5$

**Example:**

Find proportion of obese people (BMI>30), given sample of 1000 people, among which 391 are obese.

$$p = 391/1000 = 0.391$$

$$SE = 0.0154$$

$$95\%CI = 0.362 \text{ to } 0.422$$

**Question:**

151 have asthma in 1000, compute..?

# Introduction to Statistics

---

## Part 3: Hypothesis: Design & Testing

*Nikesh Bajaj, PhD*

*Lecturer in Data Science, Queen Mary University of London*

*Honorary Research Associate, Imperial Collage London*

[n.bajaj@{qmul,imperial}.ac.uk](mailto:n.bajaj@{qmul,imperial}.ac.uk)

<https://nikeshbajaj.in>

# Given two groups of data

---

TEST FOR DIFFERENCES

TEST FOR ASSOCIATIONS

# Hypothesis Testing (p-value)

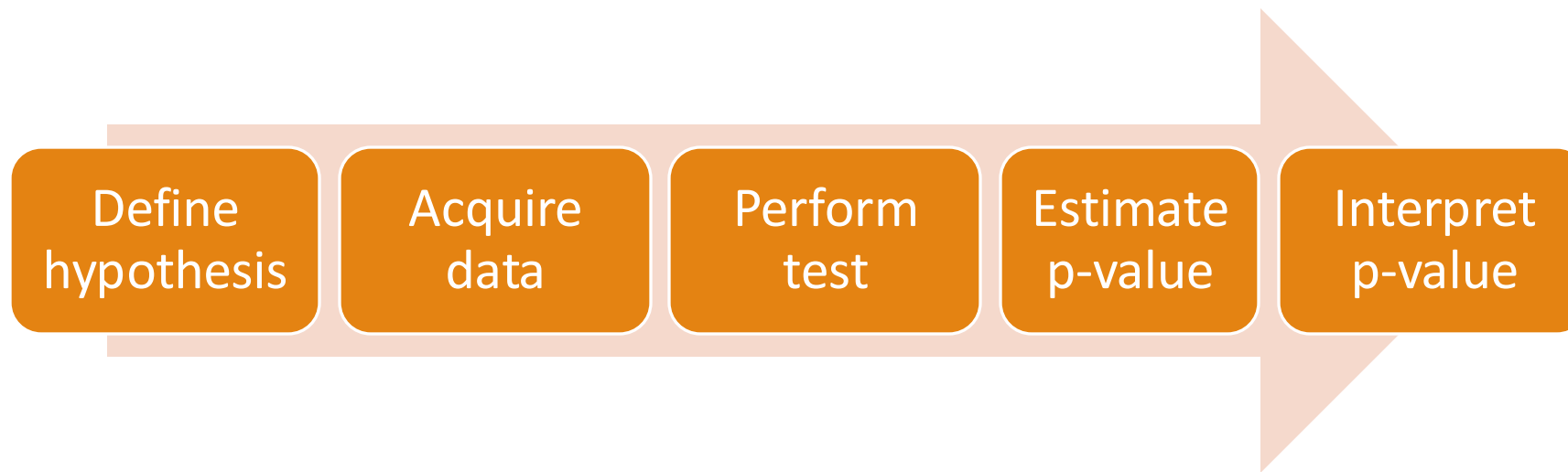
---

- Hypothesis? testing? P-value?
- Type I and type II error,
- Multiple testing and statistical power

# Hypothesis and Testing

---

- Hypothesis: A statement about a true value of parameters and their relationship in a defined **population**
- Testing: The procedure, based on **sample**, to determine if the hypothesis is a ***reasonable statement***\*





# Define Hypothesis

---

You have a  
question/state  
ment? →

Define  
hypothesis

In science to verify if a hypothesis is a reasonable statement, you need to test it against its contrary which is assumed to be true.

## Null Hypothesis $H_0$

Assumed to be true → No true difference or relationship between observed values in the sampled population

## Alternative Hypothesis $H_1$ { “Burden of Proof” }

To be proven → There **IS** true difference or relationship between ..

# Example 1

---

You have a  
question/state  
ment? →

Define  
hypothesis

Is lung function different between genders?

H0: lung function in men = lung function in women

H1: lung function in men  $\neq$  lung function in women

Alternative Hypothesis with sides

- H1: lung function in men  $\neq$  lung function in women - two sided
- H1: lung function in men  $<$  lung function in women - one sided
- H1: lung function in men  $>$  lung function in women - one sided

# Example 2

---

You have a question/statement? →

Define hypothesis

Is height in a class A and class B different?

H<sub>0</sub>: ?

H<sub>1</sub>: ?

**Question:**

**Can an alternative hypothesis be:**

- There is no difference between two samples?
- Two sample groups are same?

# Question:

---

You have a  
question/state  
ment? →

Define  
hypothesis

**Example: Let's make some hypothesis**

- *Your friend is **always** late and never agrees that he is.*

**Statement?:**

# Acquire Data

---

Acquiring Data:

- Conducting experiment in Lab
- Collecting data from other sources

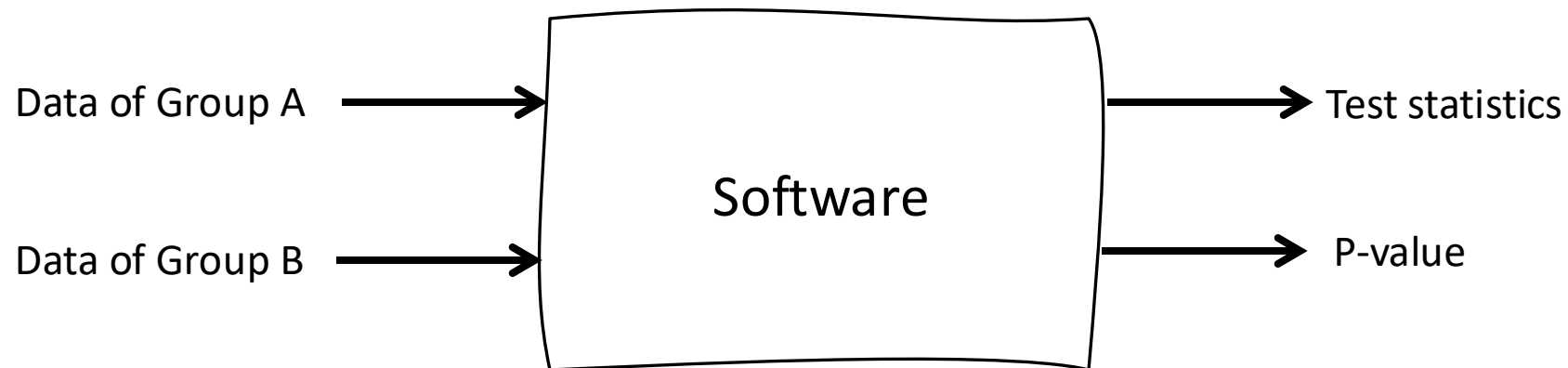
This is your sample data (think of population).

Ideally, this should be representative of population and large enough

Perform  
Test

# Performing a test

Performing a test on a two groups for establishing differences or association, we use software.

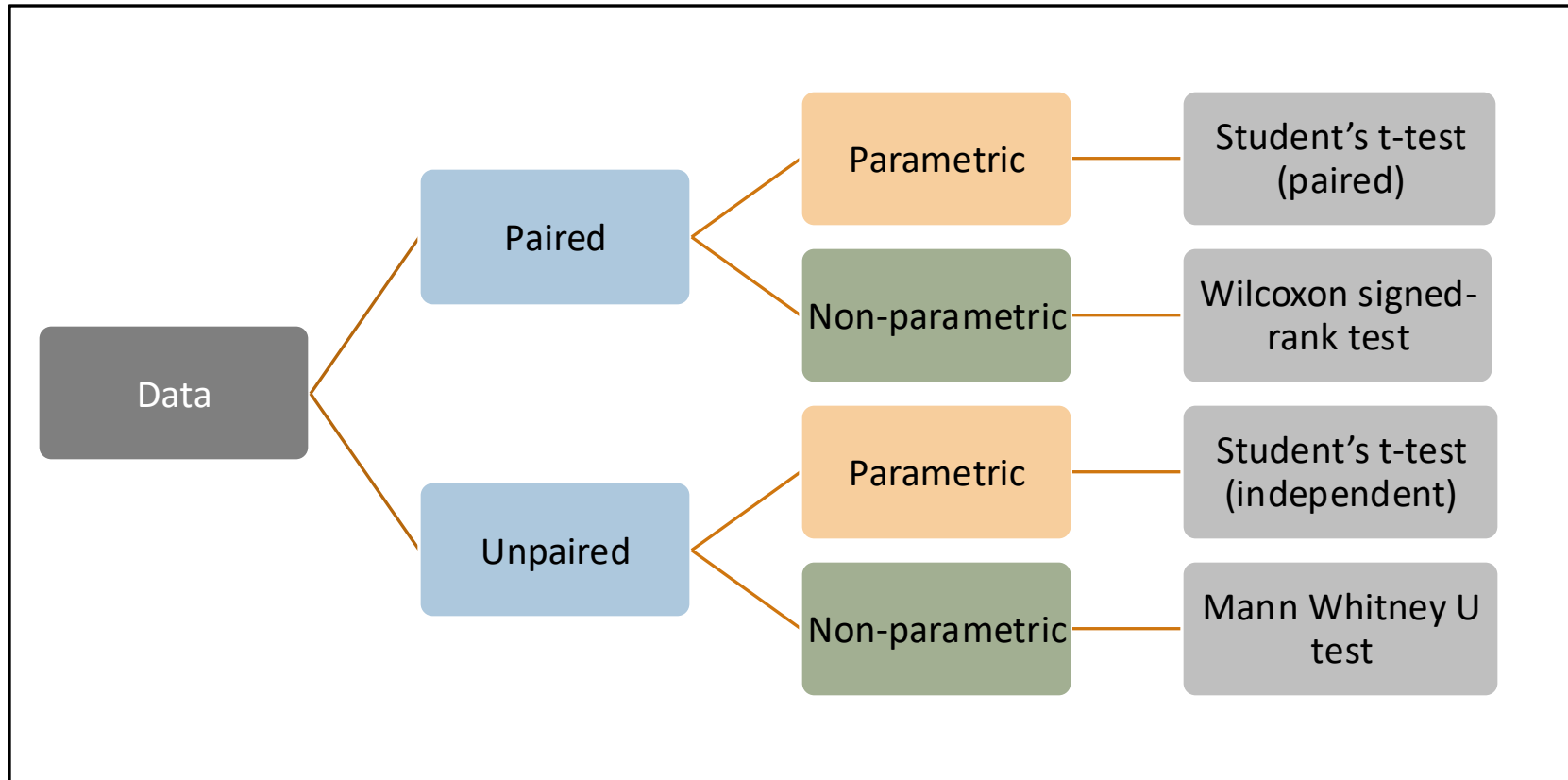


**Good news!** You don't need to remember the formulas  
But you need to know, which test to perform and how to the read results

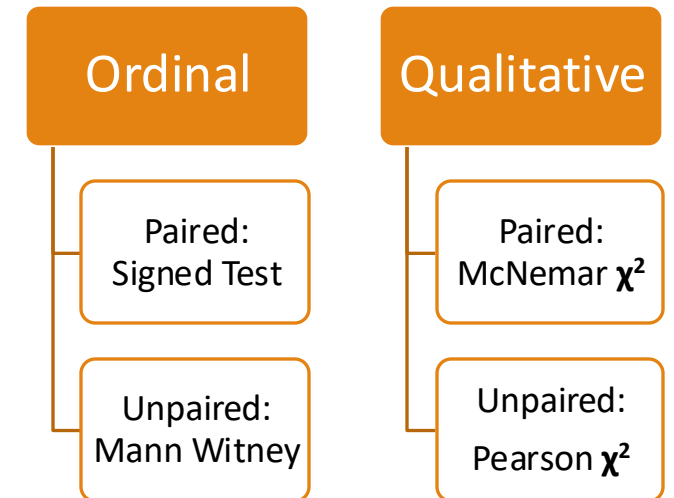
# Perform test

## Choosing a right statistical test

Perform  
Test



*We will see how to choose correct test*



# Test & P-value

---

**With a right test**, compute test statistics, that summarise the difference/relationship in your sample.

- Use test statistics to compute p-value, that tells you either to ***accept or reject null hypothesis***

**P-value:** Probability of obtaining the difference/effect observed in given sample by pure effect of **chance**, when null hypothesis is true.

- If two samples comes from same population, how likely we see a difference between them

- Ranges from 0 to 1.
- To conclude a statistical significance , we need a cut-off value  $\alpha$  (i.e. 0.05)
- NOT the probability of making a mistake!



# P-value

## Example:

- If p-value is  $< 0.05$ , we are confident enough to reject the null hypothesis.
- $\alpha = 0.05 \rightarrow$  5% chance of rejecting null hypothesis, even if it is true.
- $\alpha = 0.05 \rightarrow$  probability of committing a type I error

|   | Null hypothesis H0  |  |
|---|---|--|
|   | True  | Not True   |
| Accept H0<br>(fail to reject H0, $p > \alpha$ ) | Right   | Type 2 Error<br>(False negative with probability $\beta$ ) |
| Reject H0<br>( $p < \alpha$ )                   | Type 1 Error<br>(False positive with probability $\alpha$ ) | Right  |

Interpret  
p-value

# Type 1 and Type 2 Error

---

I think there is a  
tiger over there...



Null Hypothesis



Alternative Hypothesis

Interpret  
p-value

# Type 1 and Type 2 Error



Null Hypothesis



Alternative Hypothesis



Null Hypothesis









Null Hypothesis

H0  
True



H0  
False



|  |   |   |   |
|--|---|---|---|
| <br>...and be right                    |  | <br>and be wrong – a very bad Type 2 error | Type 2<br>FN  |
| <br>...and be wrong – a Type 1 error | Type 1<br>FP  | <br>...and be right                      |  |

# Type 1 and Type 2 Error

---

## Type 1 Error ( $\alpha$ ):

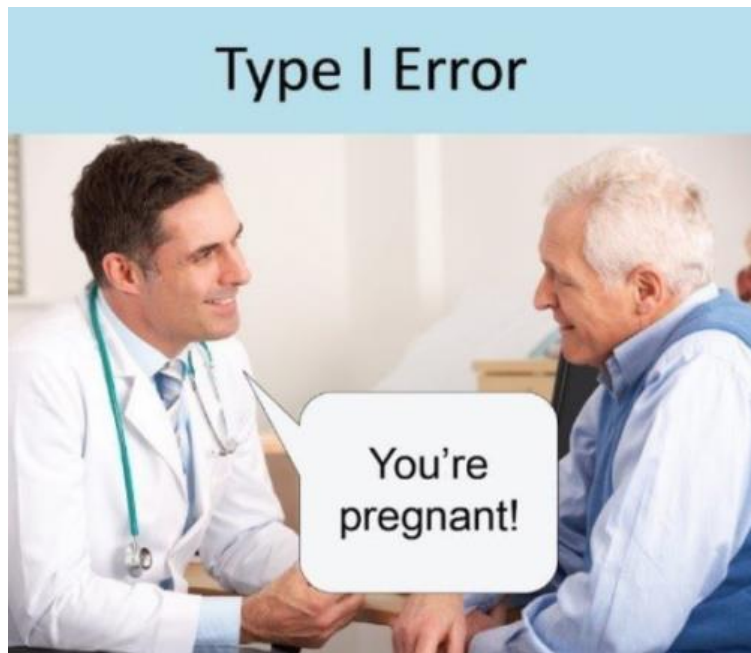
- Rejecting TRUE null hypothesis
- False Positive
- $\alpha=0.05$ , 5% probability of false positive

## Type 2 Error ( $\beta$ )

- Failing to reject FALSE null hypothesis
- False Negative
- Power is probability that we correctly reject the Null Hypothesis
- Power = 0.8,  $\beta =0.20$  (1-power), power 80%

# A fun way to remember Type of Error

---



# P-value: summary

---

Smaller the p-value, stronger the evidence against null hypothesis

If p-value is  $<0.05$ :

- It is unlikely that any difference found in samples are due to chance
- Reject the null hypothesis in favour of alternative hypothesis
- **Statistical significance**

If p-value is  $<0.001$ :

- Strong evidence of significant results

What is  $p=0.049$  or  $p=0.051$ ?

- p-value is a guideline to decide if results deserves second look
- Ref: [Scientific method: Statistical errors](#)

# Multiple testing

---

Each test has a 5% chance of 1 false positive

So, running multiple tests increases the probability of false positive

On same dataset, testing for multiple outcomes, single outcomes in multiple sub-groups, or multiple effects.

- Before testing, limit your objectives and outcomes to be tested.
- If you have to apply for multiple testing, apply  $\alpha$  correction methods (e.g. Bonferroni,  $\alpha/n$ )

# Significance and Meaningful

---

Statistically significant results does not always have meaningful relevance and vice-versa

- For that, we check
  - **Effect Size**
    - **Difference in Mean**
    - **Cohen's D**

Example:

- Two class groups A and B, have statistical significant ( $p < 0.001$ ) difference of 2 marks in given course.
- Men and women have a statistically significant difference of 0.5mL in lung function



# Power and Sample size (N)

- Low power (due small sample size ) increase the probability of False Negative
- You might find no difference between groups, and that might be False Negative, due to high  $\beta$  or low power (small sample size)

### Example

| Group  | N   | Mean    | sd     |
|--------|-----|---------|--------|
| Male   | 421 | 3555.20 | 909.75 |
| Female | 407 | 2500.77 | 625.99 |

Difference in mean = 1054.43  
p-value <0.0001  
T-statistics = 0.81

| Group  | N  | Mean    | sd    |
|--------|----|---------|-------|
| Male   | 12 | 3548.08 | 917.3 |
| Female | 8  | 2993.00 | 571.3 |

Difference in mean = 555.08  
p-value = 0.4319  
T-statistics = 0.81

# Sample Size calculation

---

For two groups with mean  $m_1$  and  $m_2$ , and standard deviation of  $sd$ , we need  $N$  samples in each group to be able to reject a null hypothesis with probability of False positive as 5% and probability of False negative of 80%

$$N \text{ in each group} = f(\alpha, \beta) \times \frac{2(sd^2)}{(m_2 - m_1)^2}$$

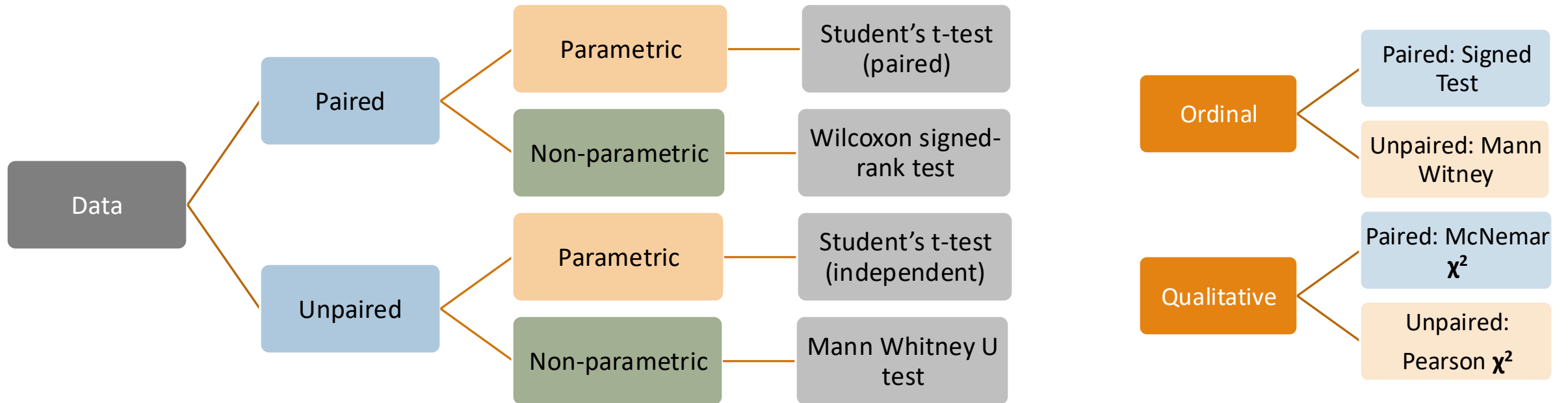
$$f(\alpha, \beta) = f(0.05, 0.20) = 7.85$$

Don't worry about the formula, it is available in all the software.

Important thing to notice  $\rightarrow$  smaller the difference you like to detect, more samples you need, smaller the  $sd$  is less sample you need

# Choosing the correct test

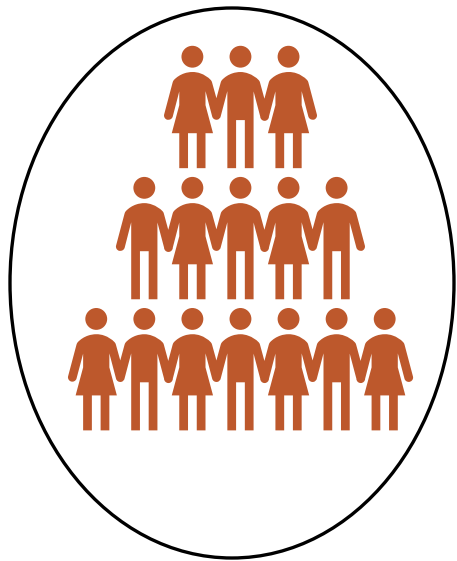
---



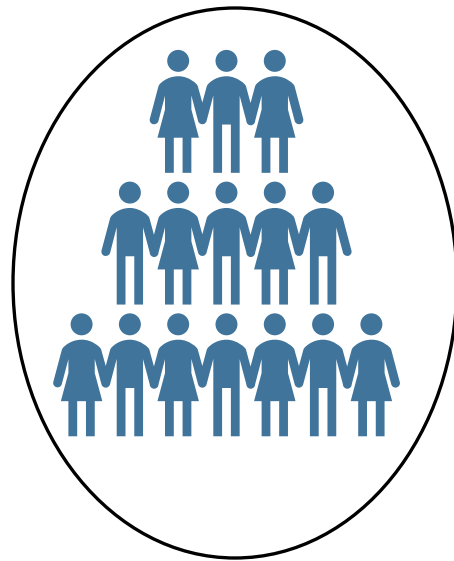
# Paired & Unpaired

Examples?

Unpaired

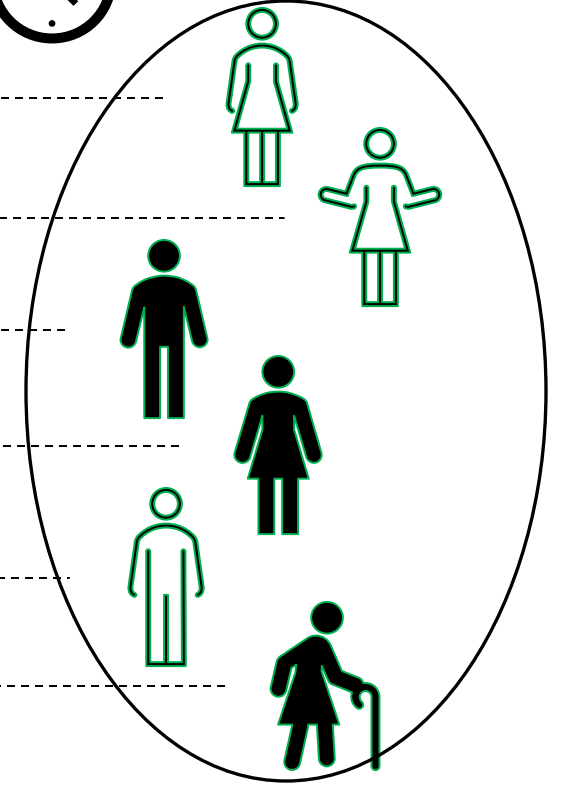
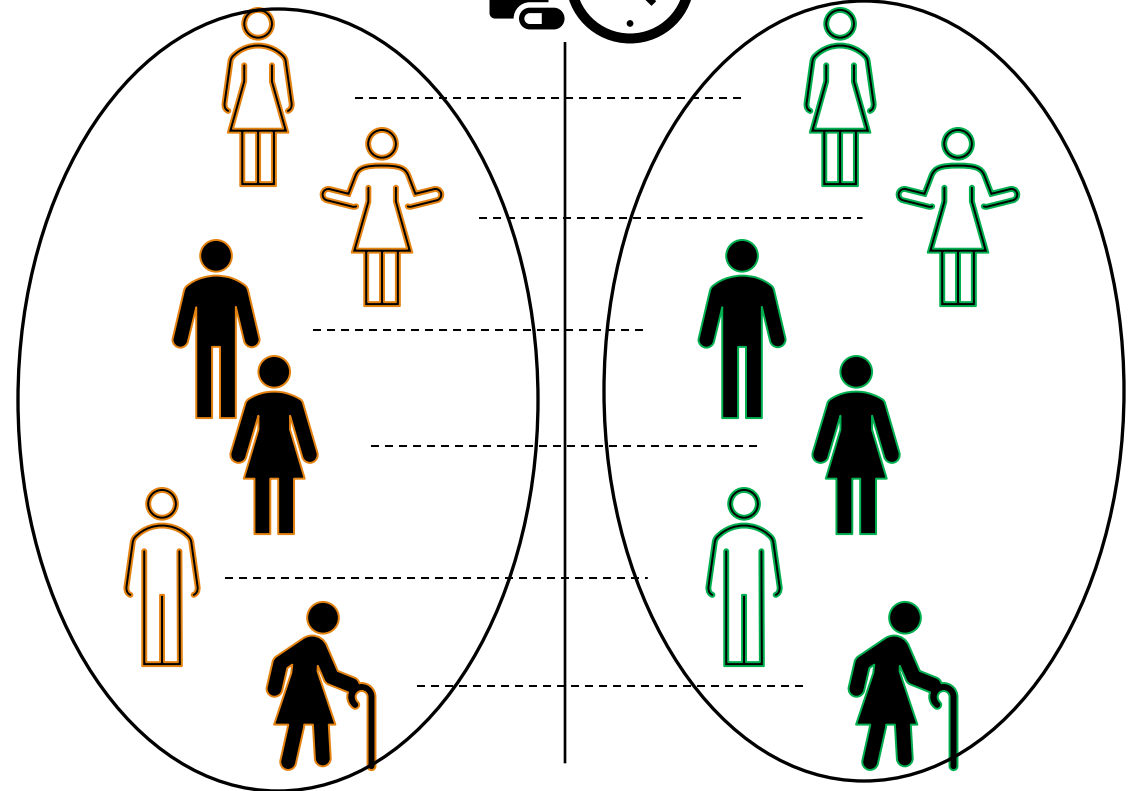


A



B

Paired



# Paired & Unpaired

---

## **Unpaired (independent)**

- Data collected from each sample is independent of time (usually collected once)
- Different subjects in different groups
- BMI, age, height of subject

## **Paired (dependent)**

- Data collected from same subjects at different time (before and after treatment)
- Same subjects in different groups
- BMI, before and after a treatment

# Parametric & Non-parametric

---

## Parametric Tests

- Relies on the underlying statistical distribution
- Normally distributed data (**normality test**)

## Non-parametric Tests

- Do not depend on any distribution

# Normality and Homogeneity of Variance

---

- **Normality test** : Shapiro-test
- If test is not significant ( $p < \alpha$ ) that indicates the sampling distribution is normally distributed.
  
- **Test for Homogeneity of Variance:** Levene test
- The small p-value suggests that the populations do not have equal variances

# For Categorical Data

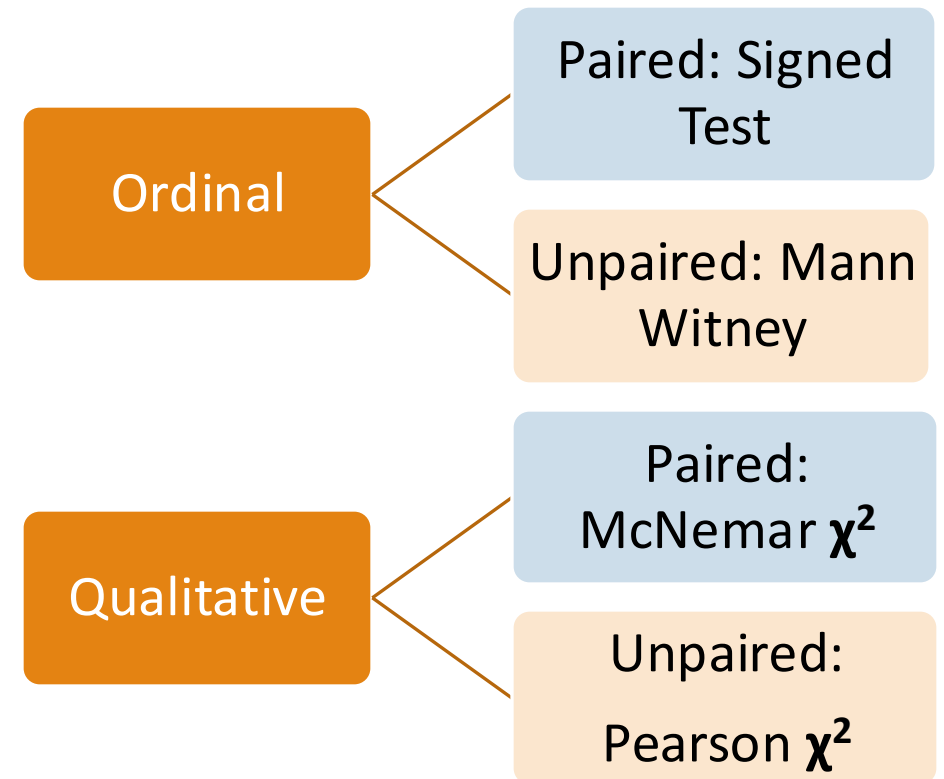
---

## Numerical

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

## Categorical

- **Qualitative:** ethnicity, disease or not? , sex?
  - Binary (2), nominal (>2)
- **Ordinal:** satisfaction-rating, age-group





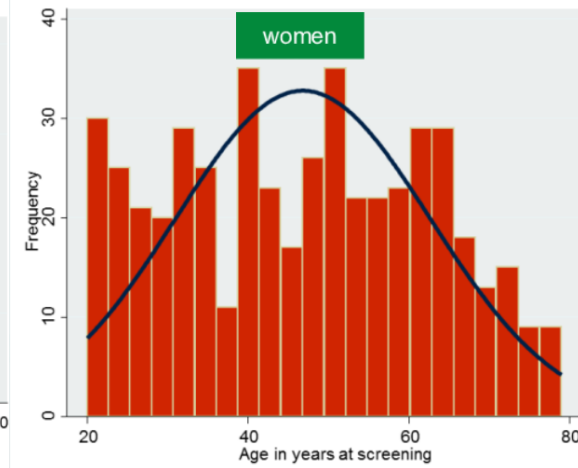
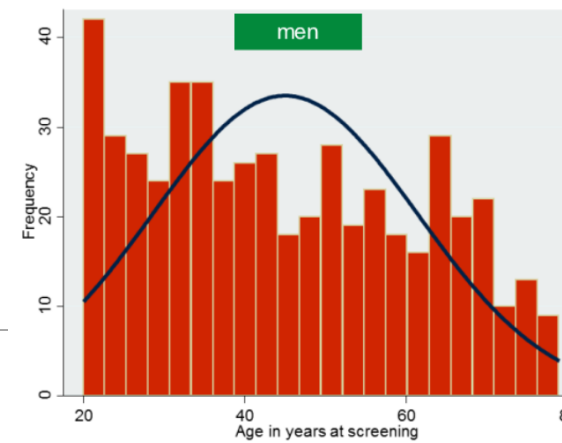
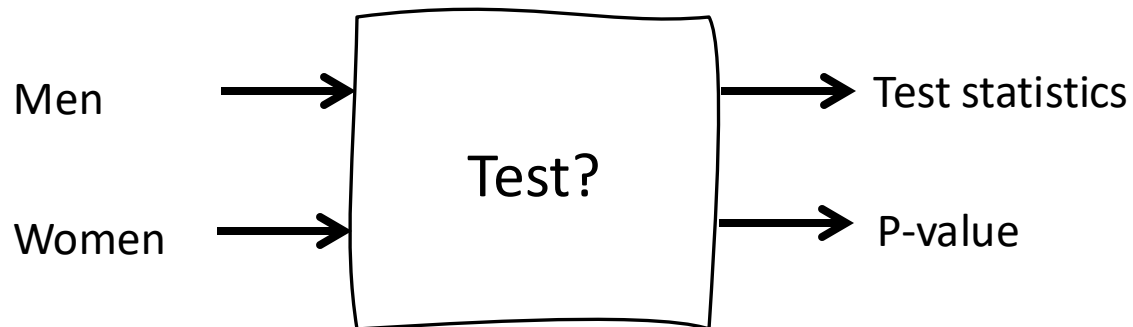
# Example 1

Is average age of men and women different in given sample (dataset\*)?

H0:  $\mu_{men} = \mu_{women}$

H1:  $\mu_{men} \neq \mu_{women}$

*Paired? Normality? Equal variance?*



| Group | N   | mean age | sd    | sd <sup>2</sup> |
|-------|-----|----------|-------|-----------------|
| Men   | 514 | 45       | 16.41 | 269.38          |
| Women | 486 | 46.83    | 15.87 | 251.81          |

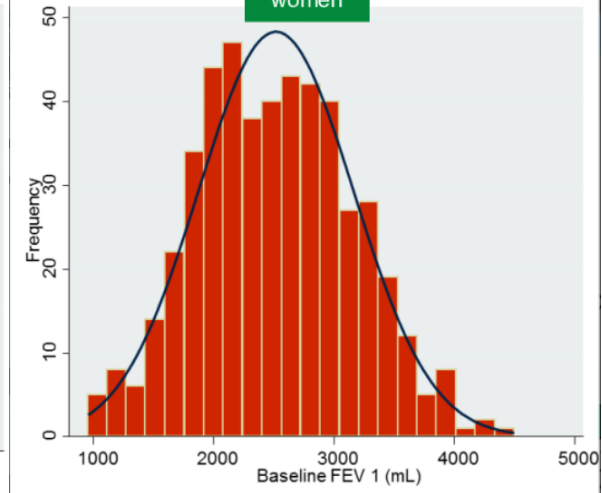
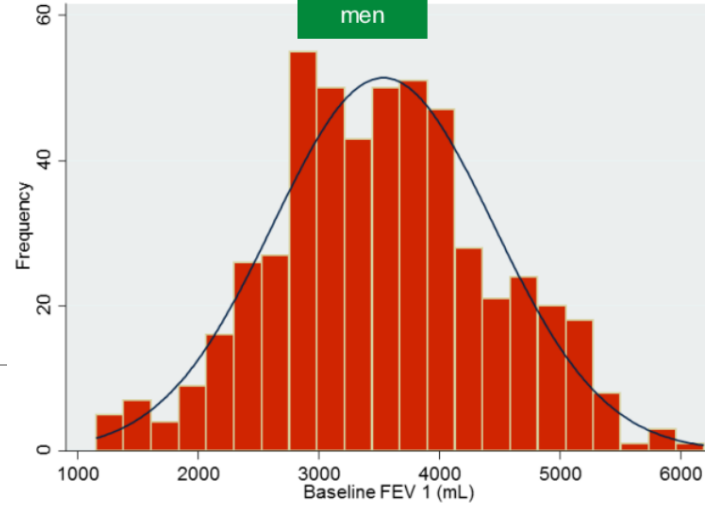
Test statistics: 1.79

P-value: 0.0741

Age difference?

# Example 2

Is the average lung function in men different to one in women in general population

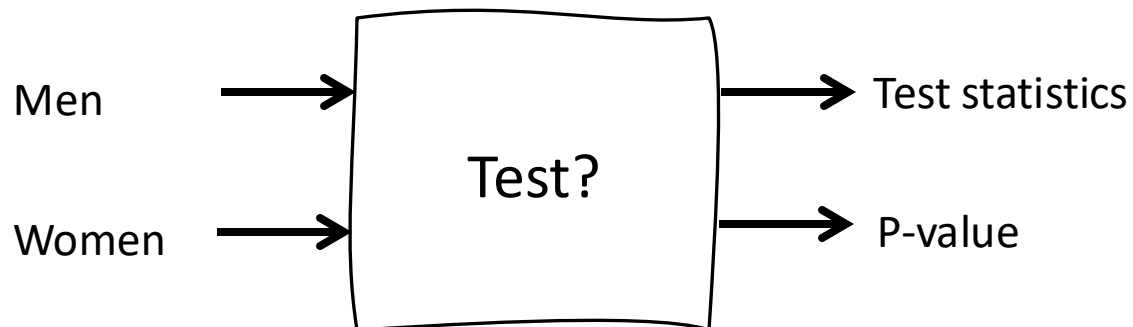


H0:  $\mu_{men} = \mu_{women}$

H1:  $\mu_{men} \neq \mu_{women}$

| Group | N   | mean fev1(mL) | s d    |
|-------|-----|---------------|--------|
| Men   | 514 | 3535.08       | 915.08 |
| Women | 486 | 2515.17       | 646.08 |

Paired? Normality? Equal variance?



Test statistics: 20.26

P-value: 0.001

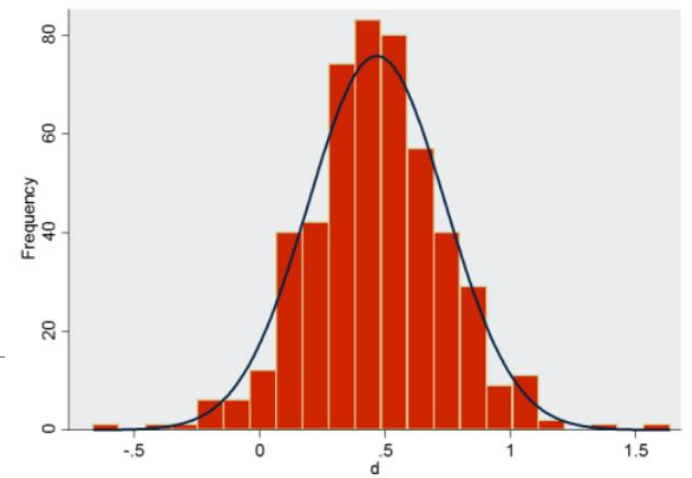
difference?

# Example 3

Has the lung function changed after intense exercise in the study

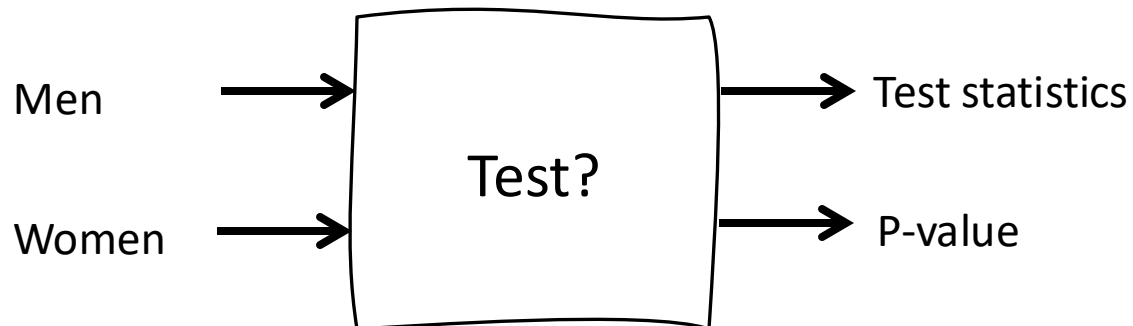
H0  $\mu_{before} = \mu_{after} : \mu_d = 0$

H1:  $\mu_{before} \neq \mu_{after} : \mu_d \neq 0$



|   | N   | mean | median | sd   | sd   |
|---|-----|------|--------|------|------|
| d | 496 | 0.47 | 0.47   | 0.27 | 0.01 |

*Paired? Normality? Equal variance?*



Test statistics: 38.13

P-value < 0.001

difference?

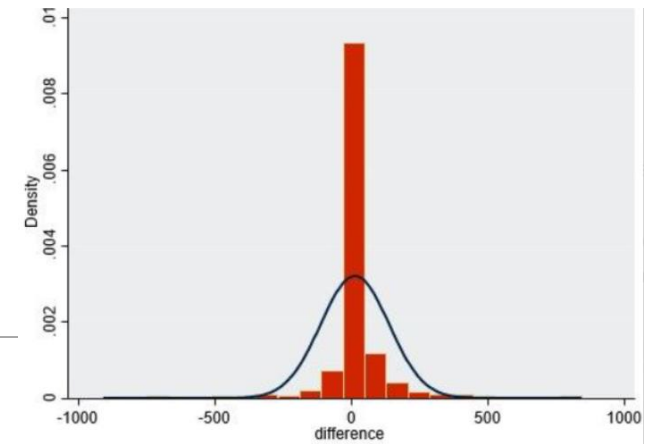
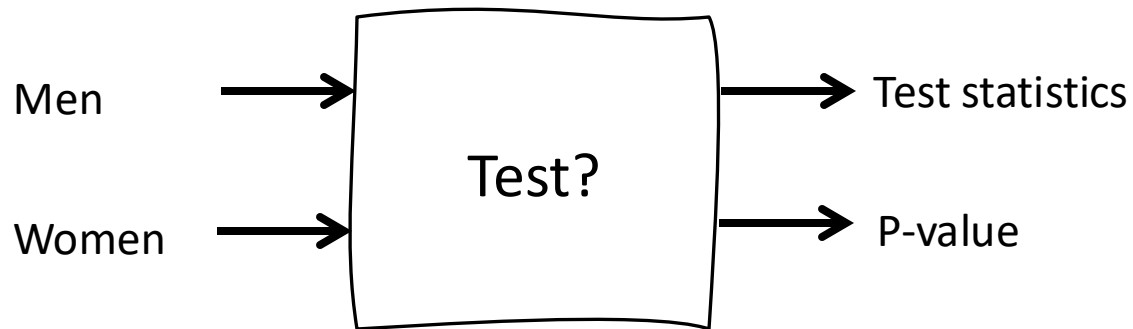
# Example 4

*Has total immunoglobulin E (IgE) changed over time (over 10 years)*

H0  $\mu_{before} = \mu_{after} : \mu_d = 0$

H1:  $\mu_{before} \neq \mu_{after} : \mu_d \neq 0$

*Paired? Normality? Equal variance?*



Test statistics: 10

P-value > 0.05

# Let's give it a try

---

From the statement from list, let's define a hypothesis, testing approach:

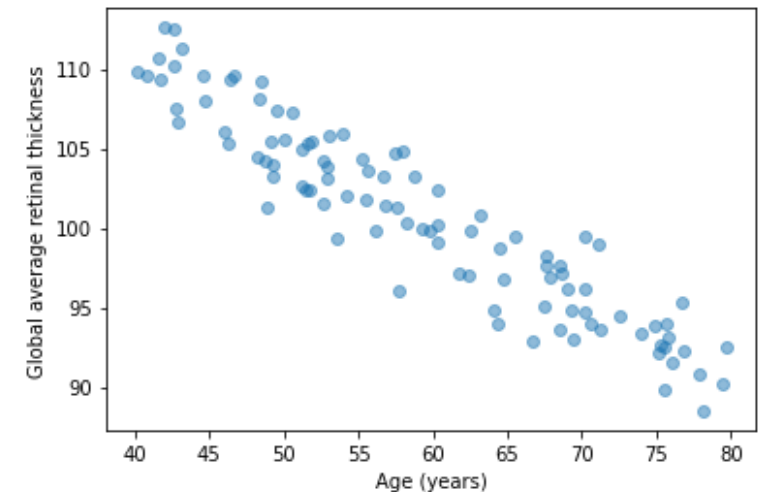
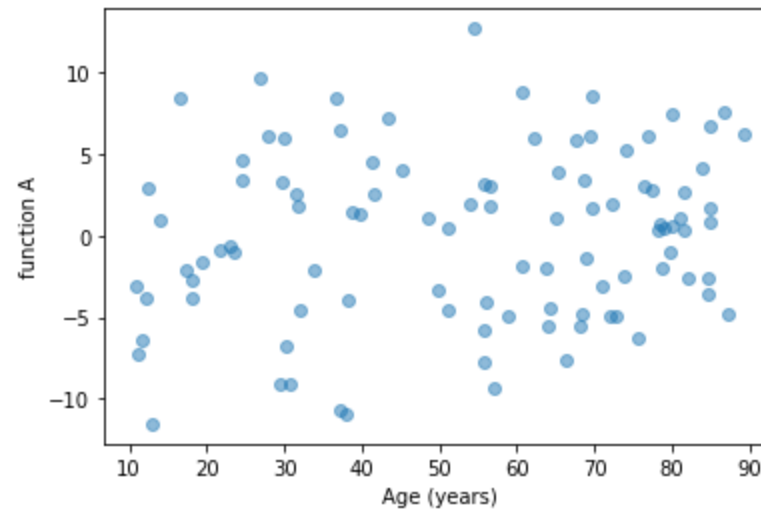
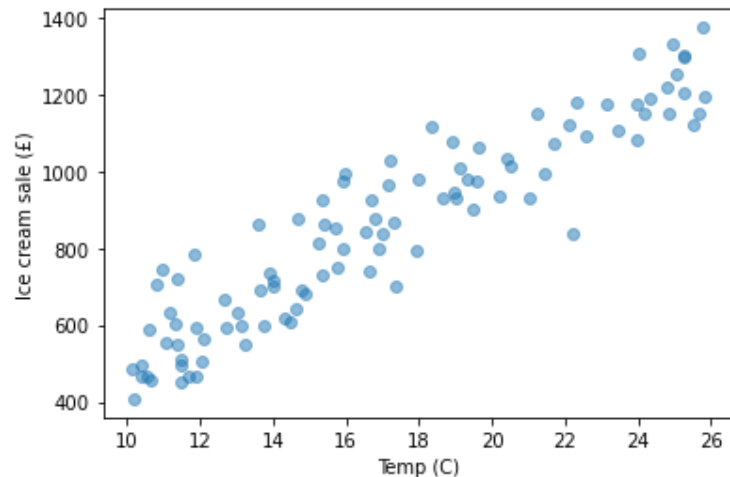
1. Effectiveness of a teaching method for a subject
2. Effectiveness of a drug on elderly  $>50$  for lung function
3. Lung function of smokers and non-smokers

# Association between two

## Correlation

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

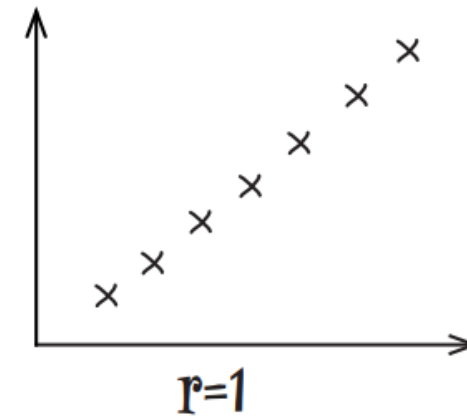
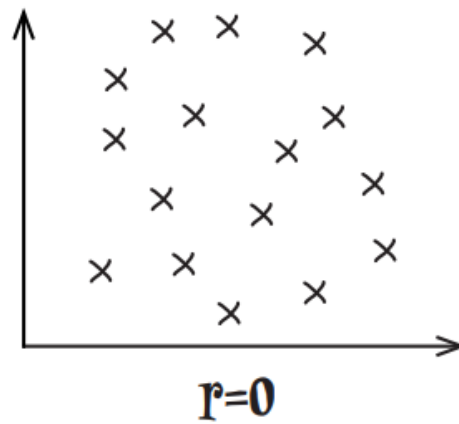
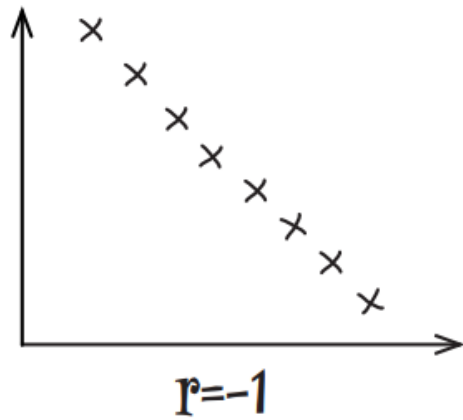
One simple way is to plot scatter graph and see



# Quantifying Correlation

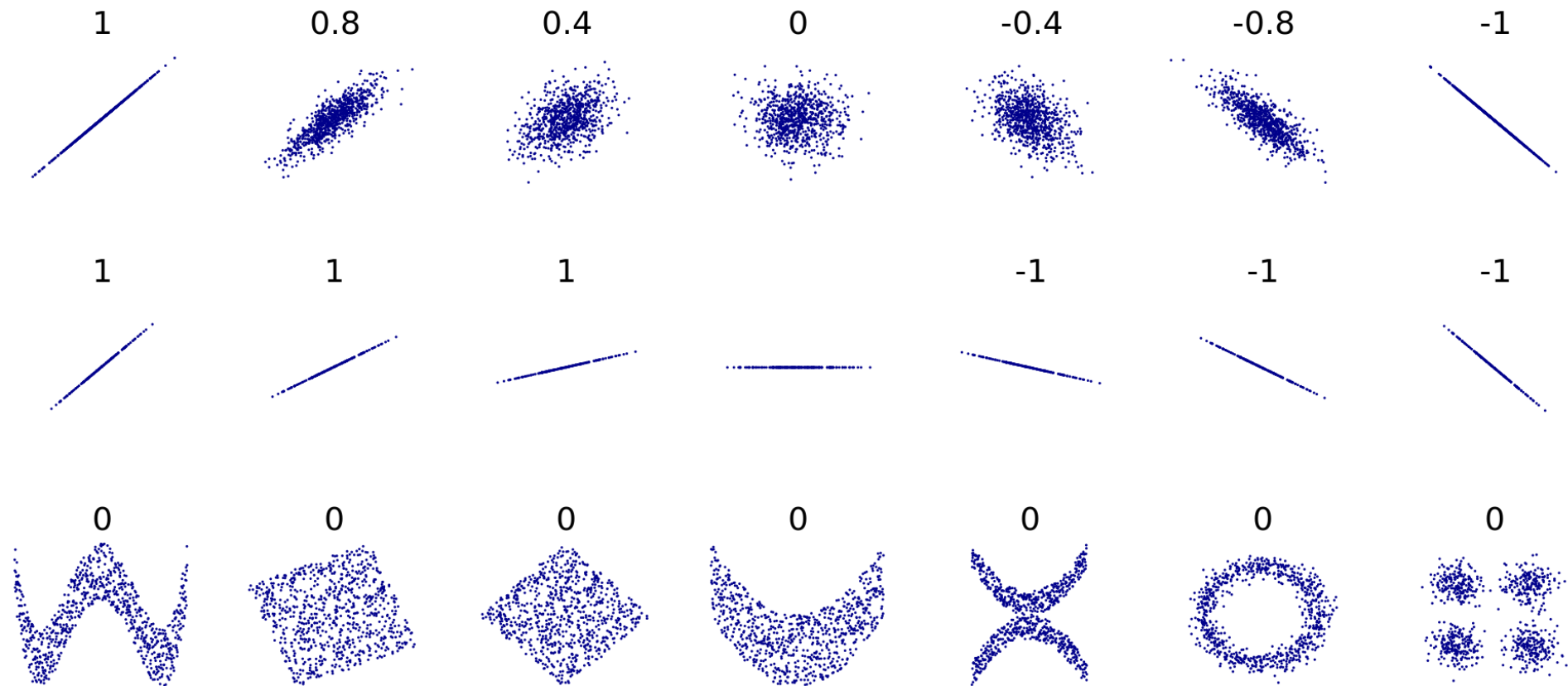
---

Pearson Correlation Coefficient  $r$  or  $\rho$  (rho)



# Pearson Correlation Coefficient

---





# Correlation

## Pearson Correlation Coefficient

### Parametric test

- x, y: normally distributed
- linear relationship

Generally:

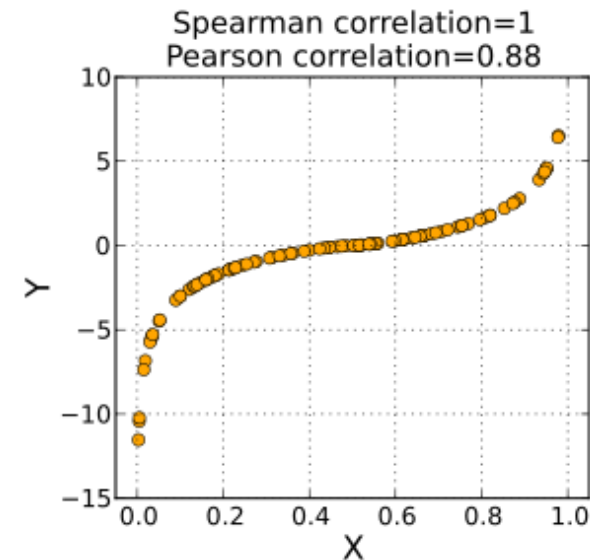
- $|r| < 0.4 \rightarrow$  weak
- $0.4 < |r| < 0.7 \rightarrow$  moderate
- $0.7 < |r| \rightarrow$  strong

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

## Spearman Rank correlation

### Non-parametric test

Based on ranks rather than exact values



# Correlation and P-value

---

P-value can be obtained from correlation with

Null Hypothesis  $H_0 : r = 0$

Alternative Hypothesis  $H_1 : r \neq 0$

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

# Examples

---

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

# Correlation

---

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

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

*Correlation is not causation*

# For single or more than 2 samples

---

## One Sample t-Test

- Also known as single sample t-test
- When you have only one sample, to test if the mean of sample is equal to some well established/ hypothesized mean
  
- Null hypothesis  $H_0 : \mu_s = P$
- Alternative hypothesis  $H_1 : \mu_s \neq P$

## ANOVA

- Analysis of Variance, generalised for more than 2 groups
- Tests whether mean of all the groups are similar or not
  
- Null hypothesis: There is no difference in means between groups
- Alternative hypothesis: At least one group mean is significantly different from the overall mean

Stats Demo links:

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

<https://nikeshbajaj.github.io/teaching/demos/Stats/hypothesis>

Resources/Video/Material:

<https://nikeshbajaj.github.io/teaching/stats>

If you have any question or doubt, please  
contact me via email

---

*Nikesh Bajaj, PhD*

*Lecturer in Data Science, Queen Mary University of London*

*Honorary Research Associate, Imperial Collage London*

[n.bajaj@{qmul, imperial}.ac.uk](mailto:n.bajaj@qmul.ac.uk)

<https://nikeshbajaj.in>