

# Apply statistics for Big Data Analytics



## Need of statistics in Data Science & Big Data Analytics



## Measures of Central Tendency



## Measures of Dispersion



## Need of statistics in Data Science & Big Data Analytics



### Measures of Central Tendency



Mean, Median, Mode, Mid-range

### Measures of Dispersion



Range, Variance, Mean Deviation, Standard Deviation

Bayes theorem, Basics and need of hypothesis and hypothesis testing, Pearson Correlation, Sample Hypothesis testing, Chi-Square Tests, t-test.

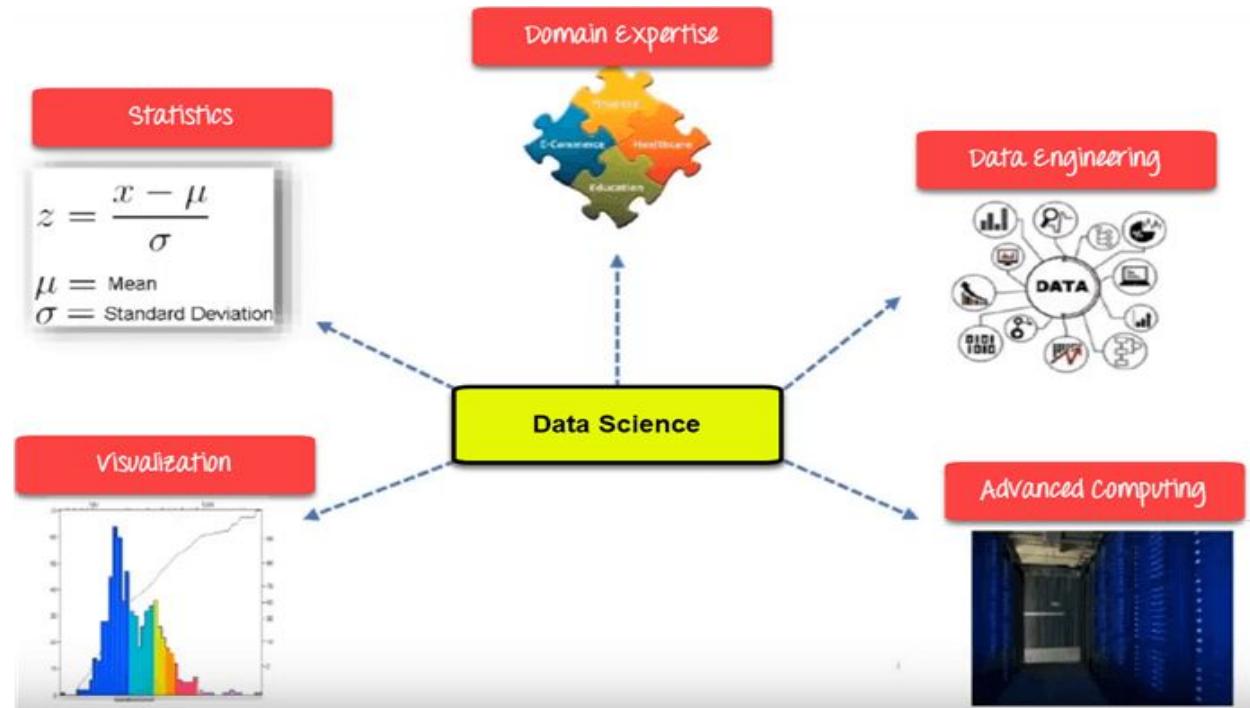


# Need of Statistics

Need of statistics in Data Science  
& Big Data Analytics



## Components of Data Science



# Need of Statistics

Need of statistics in Data Science  
& Big Data Analytics



## Statistics

Statistic

is

most critical unit of Data  
Science

It is

method or science of  
collecting and analyzing  
numerical data

useful insights

To get

large quantities

In

# Need of Statistics

## Need of statistics in Data Science & Big Data Analytics



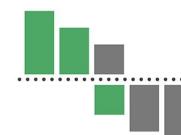
1



Identify the Importance

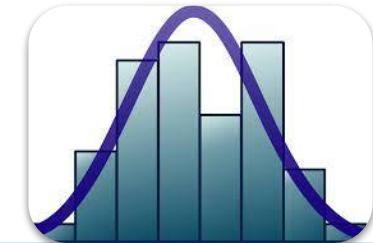
of

Relevant Features



various statistical tests.

By using



# Need of Statistics

Need of statistics in Data Science  
& Big Data Analytics



2



Finding the Relationships

between

Features

| #     | Car Make | Car Model | Car Year | ... | ... | Sell Year | ... |
|-------|----------|-----------|----------|-----|-----|-----------|-----|
| 1     | Toyota   | Camry     | 2018     | ... | ... | 2018      | ... |
| 2     | Toyota   | Corolla   | 2019     | ... | ... | 2019      | ... |
| 3     | Toyota   | Camry     | 2018     | ... | ... | 2018      | ... |
| 4     | Toyota   | Corolla   | 2019     | ... | ... | 2019      | ... |
| ...   | ...      | ...       | ...      | ... | ... | ...       | ... |
| ...   | ...      | ...       | ...      | ... | ... | ...       | ... |
| ...   | ...      | ...       | ...      | ... | ... | ...       | ... |
| 9999  | Toyota   | Camry     | 2018     | ... | ... | 2018      | ... |
| 10000 | Toyota   | Camry     | 2018     | ... | ... | 2018      | ... |

Duplicate Features.

To eliminate

# Need of Statistics

## Need of statistics in Data Science & Big Data Analytics



Convert  
the

Features  
into

Required Format



Designed by PO

# Need of Statistics

## Need of statistics in Data Science & Big Data Analytics



Normalizing &  
Scaling the data

This step also involves

the identification of the distribution  
of data and the nature of data.

# Need of Statistics

## Need of statistics in Data Science & Big Data Analytics



5

Tacking the

Data for further  
processing

By using

Required adjustment  
in the data



# Need of Statistics

## Need of statistics in Data Science & Big Data Analytics

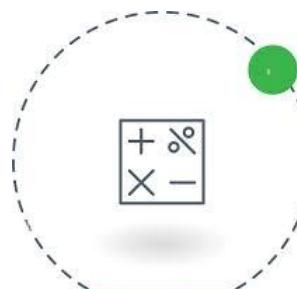


After



Processing  
the data

Identify the



Right mathematical  
approach/model

# Need of Statistics

## Need of statistics in Data Science & Big Data Analytics



Once the

Results are obtained

the

Results are verified

On the

Different accuracy  
measurement scales

# Need of Statistics

## Know your Data



- collection of **objects and their attributes**
- An attribute is a **property or characteristic of an object**
  - Examples: eye color of a person, temperature, cost, etc.
  - also known as **variables, fields, characteristics, dimensions, or features**
- A collection of attributes describe an object
  - Objects are also known as **records, points, cases, samples, entities, or instances**

Objects

Attributes

| Tid | Refund | Marital Status | Taxable Income | Cheat |
|-----|--------|----------------|----------------|-------|
| 1   | Yes    | Single         | 125K           | No    |
| 2   | No     | Married        | 100K           | No    |
| 3   | No     | Single         | 70K            | No    |
| 4   | Yes    | Married        | 120K           | No    |
| 5   | No     | Divorced       | 95K            | Yes   |
| 6   | No     | Married        | 60K            | No    |
| 7   | Yes    | Divorced       | 220K           | No    |
| 8   | No     | Single         | 85K            | Yes   |
| 9   | No     | Married        | 75K            | No    |
| 10  | No     | Single         | 90K            | Yes   |

## Know your Data



**Dimension**

- Data Warehousing

**Feature**

- Machine learning

**Variable**

- Statisticians

**Attribute**

- Data mining and database professional

## Attribute Values



- Attribute values are numbers or symbols assigned to an attribute

|                    |     |       |     |       |     |     |     |
|--------------------|-----|-------|-----|-------|-----|-----|-----|
| Heights<br>(in cm) | 164 | 167.3 | 170 | 174.2 | 178 | 180 | 186 |
|--------------------|-----|-------|-----|-------|-----|-----|-----|

Univariate Data

| TEMPERATURE(IN CELSIUS) | ICE CREAM SALES |
|-------------------------|-----------------|
| 20                      | 2000            |
| 25                      | 2500            |
| 35                      | 5000            |
| 43                      | 7800            |

Bivariate Data

## Attribute Values



- Attribute values are numbers or symbols assigned to an attribute

| Height | Hair   | Eyes  | CLASS |
|--------|--------|-------|-------|
| short  | blonde | blue  | ⊕     |
| short  | dark   | blue  | ⊖     |
| tall   | dark   | brown | ⊖     |
| tall   | blonde | brown | ⊖     |
| tall   | dark   | blue  | ⊖     |
| short  | blonde | brown | ⊖     |
| tall   | red    | blue  | ⊕     |
| tall   | blonde | blue  | ⊕     |

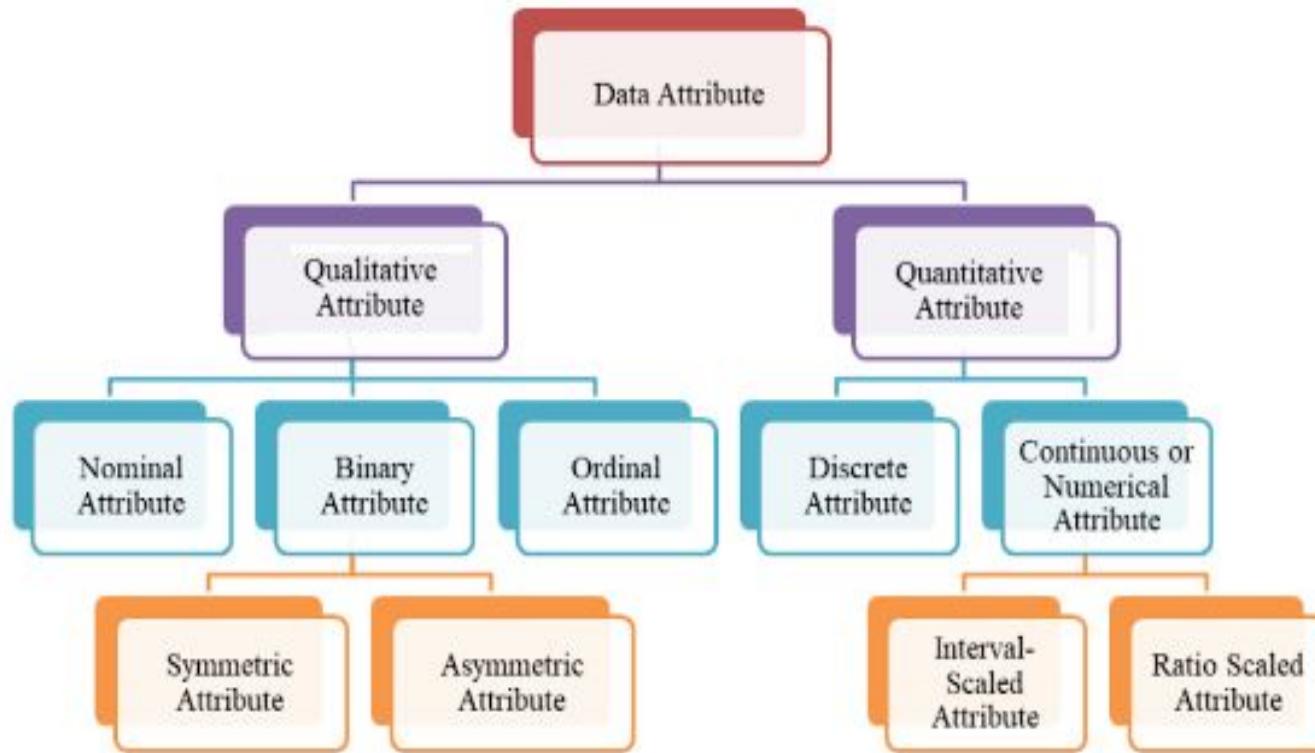
Multivariate Data

# Need of Statistics

## Measures of Central Tendency



## Types of Data Attribute



# Data Attribute

## Quantitative Attributes



| Age          | Dates | Distance | IQ  | Weight    |
|--------------|-------|----------|-----|-----------|
| 10 years old | 1066  | 3 metres | 80  | 80 grams  |
| 20 years old | 1492  | 6 metres | 100 | 100 grams |
| 30 years old | 1776  | 9 metres | 120 | 120 grams |

# Data Attribute

## Qualitative Attributes



| Socioeconomic status | Opinion         | Nationality | Genre     | Hair colour |
|----------------------|-----------------|-------------|-----------|-------------|
| Lower class          | Agree           | British     | Rock      | Red         |
| Middle class         | Mostly agree    | American    | Hip-Hop   | Brown       |
| Upper class          | Neutral         | Spanish     | Jazz      | Blonde      |
|                      | Mostly disagree |             | Classical |             |
|                      | Disagree        |             |           |             |

## Quantitative Attributes



- attributes can be measured and assigned a number
- measurable and can be expressed in integer or real values
- tends to answer questions about the ‘how many’ or ‘how much’
- Eg. height, width, and length. Temperature and humidity. Prices. Area and volume.

## Qualitative Attributes



- Named or described in words
- Sometimes it is not easily reduced to numbers.
- tends to answer questions about the ‘what’, ‘how’ and ‘why’
- smells, tastes, textures, attractiveness, and color.

# Data Attribute

## Qualitative Attributes



1

Nominal



# Data Attribute

## Qualitative Attributes



1

**Nominal**

| Nationality | Genre     | Hair colour | Favourite animal | Pizza topping |
|-------------|-----------|-------------|------------------|---------------|
| British     | Rock      | Red         | Aardvark         | Olives        |
| American    | Hip-Hop   | Brown       | Koala            | Anchovies     |
| Spanish     | Jazz      | Blonde      | Sloth            | Pepperoni     |
|             | Classical |             |                  | Banana        |

## Qualitative Attributes



1

Nominal



### Nominal Data Definition

Nominal data is the simplest form of data, and is defined as data that is used for naming or labelling variables

# Data Attribute

## Qualitative Attributes



1

**Nominal**

**NOMINAL DATA**  
characteristics

| Measured | Ordered | Equidistant | Meaningful Zero |
|----------|---------|-------------|-----------------|
|          |         |             |                 |
|          |         |             |                 |

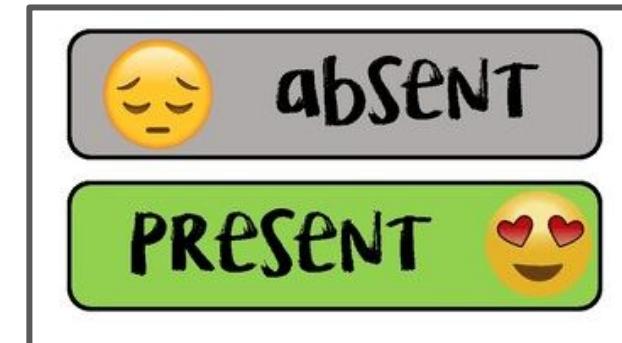
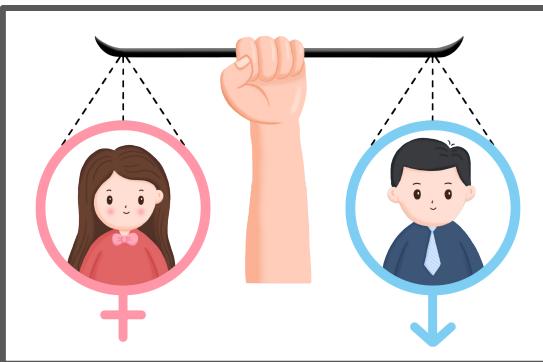
## Qualitative Attributes



1.1

Binary

a special nominal attribute with  
only two states: 0 or 1



## Qualitative Attributes

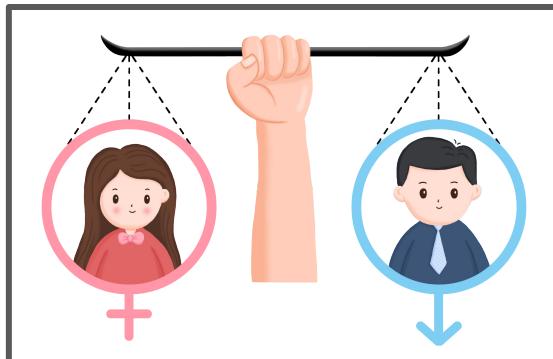


1.1

**Binary**

a special nominal attribute with  
only two states: 0 or 1

|  |                  |  |
|--|------------------|--|
|  | Symmetric Binary |  |
|--|------------------|--|



Equal Important

0 Female  
1 Male

1 Female  
0 Male

# Data Attribute

## Qualitative Attributes

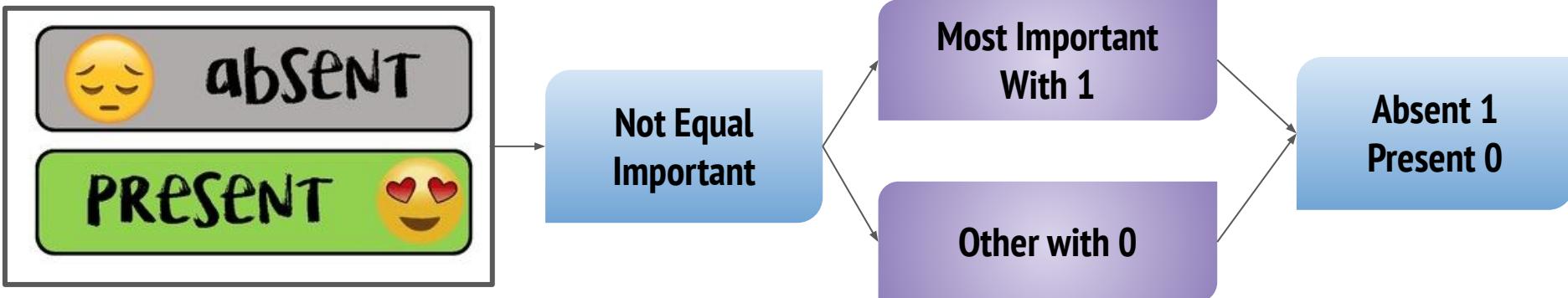


1.1

Binary

a special nominal attribute with only two states: 0 or 1

Asymmetric Binary



Depending on the task 0 or 1 mapped with attribute values :  
Here the task is to identify absent students

# Data Attribute

## Qualitative Attributes



References:

<https://www.scribbr.com/statistics/nominal-data/>

1

Nominal

Data Collection

### Examples of closed-ended questions

What is your gender?

- Male
- Female
- Other
- Prefer not to answer

Do you own a smartphone?

- Yes
- No

What is your favorite movie genre?

- Romance
- Action
- Mystery
- Animation
- Musical
- Comedy
- Thriller

### Examples of open-ended questions

1. What is your student ID number?
2. What is your zip code?
3. What is your native language?

# Data Attribute

## Qualitative Attributes



References:

<https://www.scribbr.com/statistics/nominal-data/>

1

### Nominal Data Analysis

Example: Nominal data set

You distribute a survey with a question asking respondents to select their political preferences from a list. Your data set is a list of response values.

Data set

|             |             |             |
|-------------|-------------|-------------|
| Republican  | Independent | Democrat    |
| Democrat    | Republican  | Republican  |
| Independent | Democrat    | Democrat    |
| Independent | Democrat    | Democrat    |
| Republican  | Democrat    | Independent |
| Republican  | Democrat    | Republican  |
| Republican  | Republican  | Republican  |
| Democrat    | Democrat    | Democrat    |
| Democrat    | Democrat    | Democrat    |
| Independent | Democrat    | Democrat    |

Simple Frequency Distribution

Percentage Frequency Distribution

# Data Attribute

## Qualitative Attributes



1

### Nominal Data Analysis

#### Simple Frequency Distribution

| Political preference | Frequency |
|----------------------|-----------|
| Democrat             | 13        |
| Republican           | 9         |
| Independent          | 5         |

#### References:

<https://www.scribbr.com/statistics/nominal-data/>

#### Percentage Frequency Distribution

| Political preference | Percent |
|----------------------|---------|
| Democrat             | 48.1%   |
| Republican           | 33.3%   |
| Independent          | 18.5%   |

# Data Attribute

## Qualitative Attributes



References:

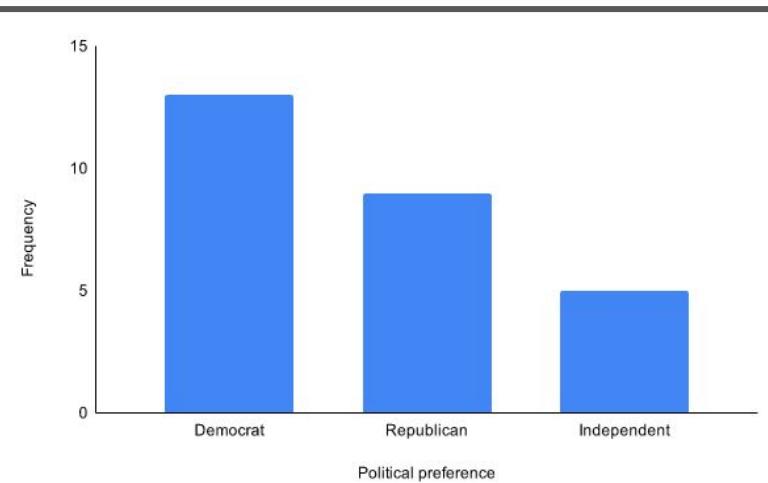
<https://www.scribbr.com/statistics/nominal-data/>

1

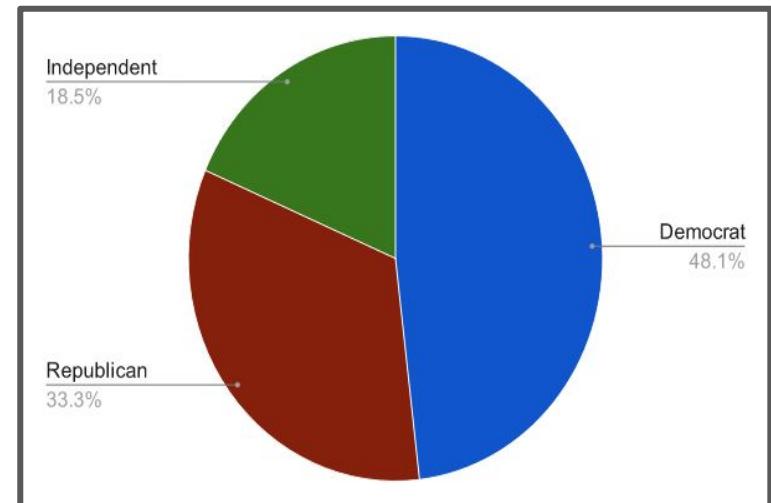
Nominal

Data Analysis

Bar Chart



Pie Chart



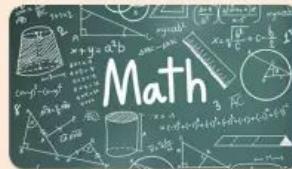
# Data Attribute

## Qualitative Attributes



1

Nominal



### NOMINAL DATA

Mathematical features

#### Grouping

 $\equiv \neq$ 

Same /  
Different



#### Sorting

 $< >$ 

Greater /  
Less Than



#### Difference

 $+ -$ 

Add /  
Subtract



#### Magnitude

 $\times \div$ 

Multiply /  
Divide



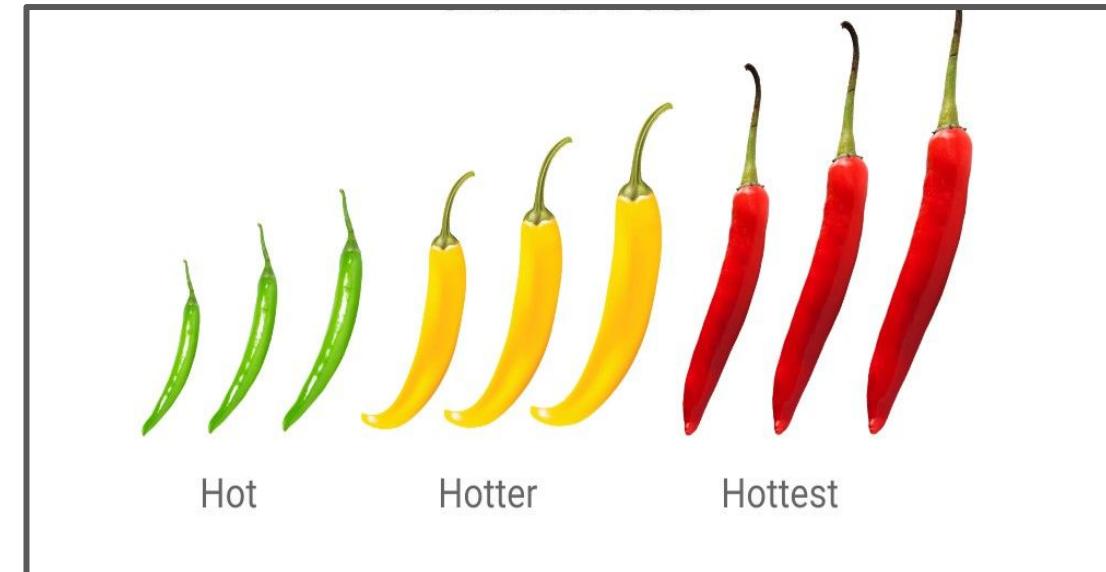
# Data Attribute

## Qualitative Attributes



2

Ordinal



# Data Attribute

## Qualitative Attributes



2

### Ordinal

| Socioeconomic status | Opinion         | Tumour Grade | Political orientation | Time of day |
|----------------------|-----------------|--------------|-----------------------|-------------|
| Lower class          | Agree           | 1            | Left                  | Morning     |
| Middle class         | Mostly agree    | 2            | Middle                | Noon        |
| Upper class          | Neutral         | 3            | Right                 | Night       |
|                      | Mostly disagree |              |                       |             |
|                      | Disagree        |              |                       |             |

## Qualitative Attributes



2

Ordinal



### Ordinal Data Definition

Ordinal data is a type of categorical data in which the values follow a natural order

## Qualitative Attributes



2

Ordinal



### ORDINAL DATA

characteristics

Measured



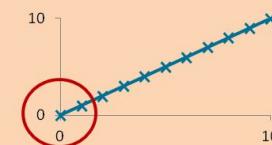
Ordered



Equidistant



Meaningful Zero



# Data Attribute

## Qualitative Attributes



2

Ordinal

Data Collection

**Question****Options****What is your age?**

- 0-18
- 19-34
- 35-49
- 50+

**What is your education level?**

- Primary school
- High school
- Bachelor's degree
- Master's degree
- PhD

**In the past three months, how many times did you buy groceries online?**

- None
- 1-4 times
- 5-9 times
- 10-14 times
- 15 or more times

# Data Attribute

## Qualitative Attributes



2

Ordinal

Data Visualization

### Example

You ask 30 survey participants to indicate their level of agreement with the statement below:

**Regular physical exercise is important for my mental health.**

Strongly disagree

Disagree

Neither disagree nor agree

Agree

Strongly agree

# Data Attribute

## Qualitative Attributes



2

Ordinal

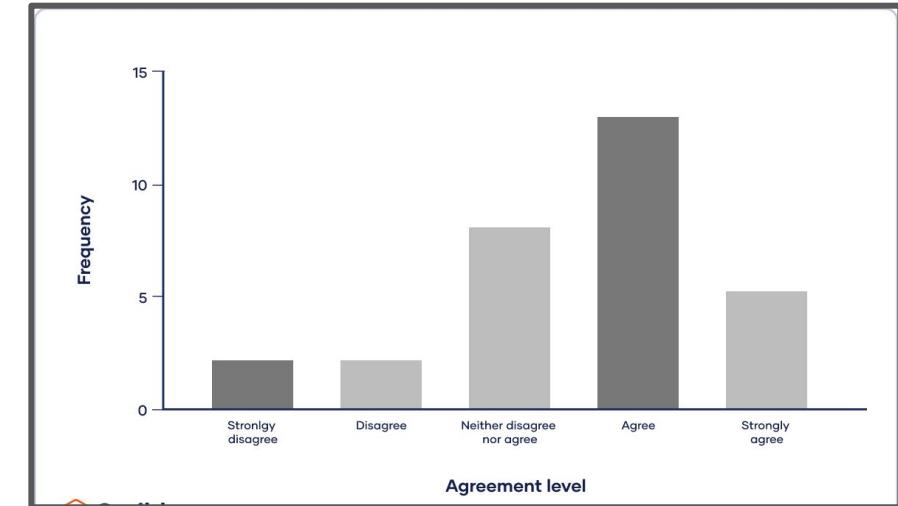
Data Analysis

### Simple Frequency Distribution

Example: Frequency distribution table

| Agreement level            | Frequency |
|----------------------------|-----------|
| Strongly disagree          | 2         |
| Disagree                   | 2         |
| Neither disagree nor agree | 8         |
| Agree                      | 13        |
| Strongly agree             | 5         |

### Bar Graph



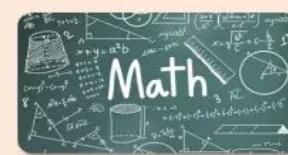
# Data Attribute

## Qualitative Attributes



2

### Ordinal



#### ORDINAL DATA

Mathematical features

##### Grouping

= ≠

Same /  
Different



##### Sorting

< >

Greater /  
Less Than



##### Difference

+ -

Add /  
Subtract



##### Magnitude

× ÷

Multiply /  
Divide

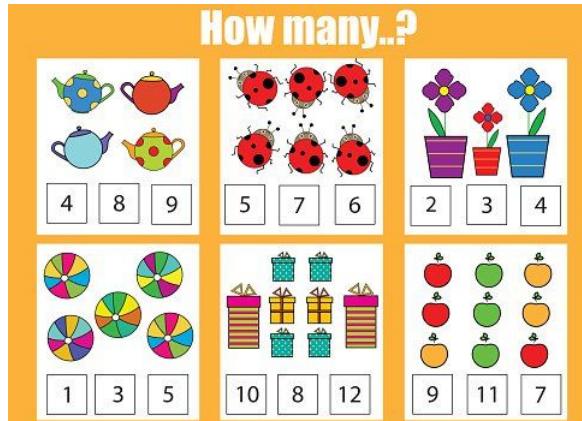


# Data Attribute

## Quantitative Attributes



3

**Discrete**

4

**Continues/Numeric**

| SI Base Units             |                      |           |        |
|---------------------------|----------------------|-----------|--------|
| Base quantity             | Typical symbol       | Base unit |        |
| Name                      | Symbol               | Name      | Symbol |
| time                      | t                    | second    | s      |
| length                    | <i>l, x, r, etc.</i> | meter     | m      |
| mass                      | <i>m</i>             | kilogram  | kg     |
| electric current          | <i>I, i</i>          | ampere    | A      |
| thermodynamic temperature | T                    | kelvin    | K      |
| amount of substance       | <i>n</i>             | mole      | mol    |
| luminous intensity        | <i>I<sub>v</sub></i> | candela   | cd     |

Source: NIST Special Publication 330:2019, Table 2.

**Discrete data is counted****Continuous data is measured**

## Quantitative Attributes



3

Discrete

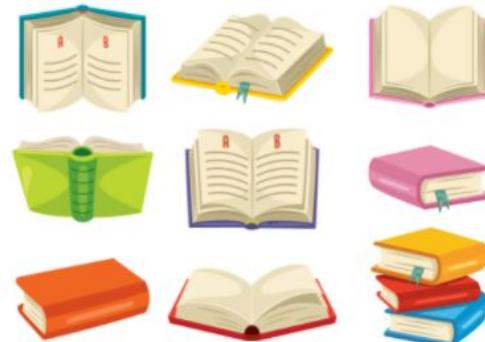
Number of books in a bookshelf



4

Continues/Numeric

Length of pages of books present in a bookshelf



## Quantitative Attributes



3

Discrete

Number of students present in a class



4

Continues/Numeric

Weight of each student in a class



## Quantitative Attributes



### Check your Knowledge

Temperature in a city on different days

Continues

Number of people travel in trains on different days of the week

Discrete

Sum of numbers on rolling three dice together.

Discrete

Volume of water in a water tank

Continues

# Data Attribute

## Quantitative Attributes

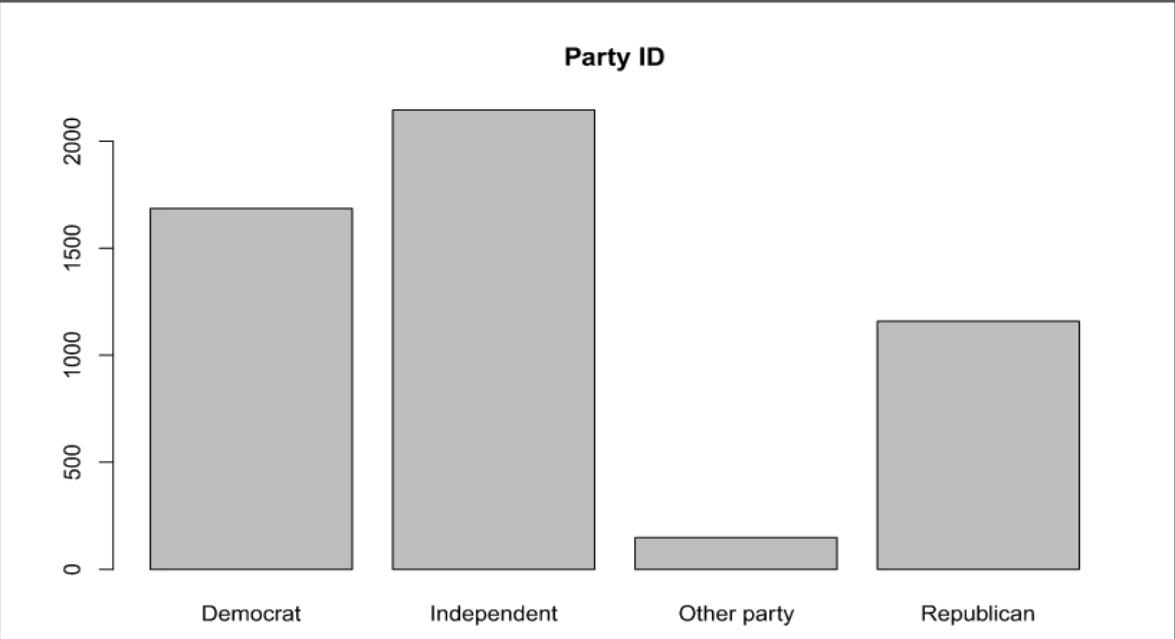


3

Discrete

Data Analysis

Bar Graph



## Quantitative Attributes

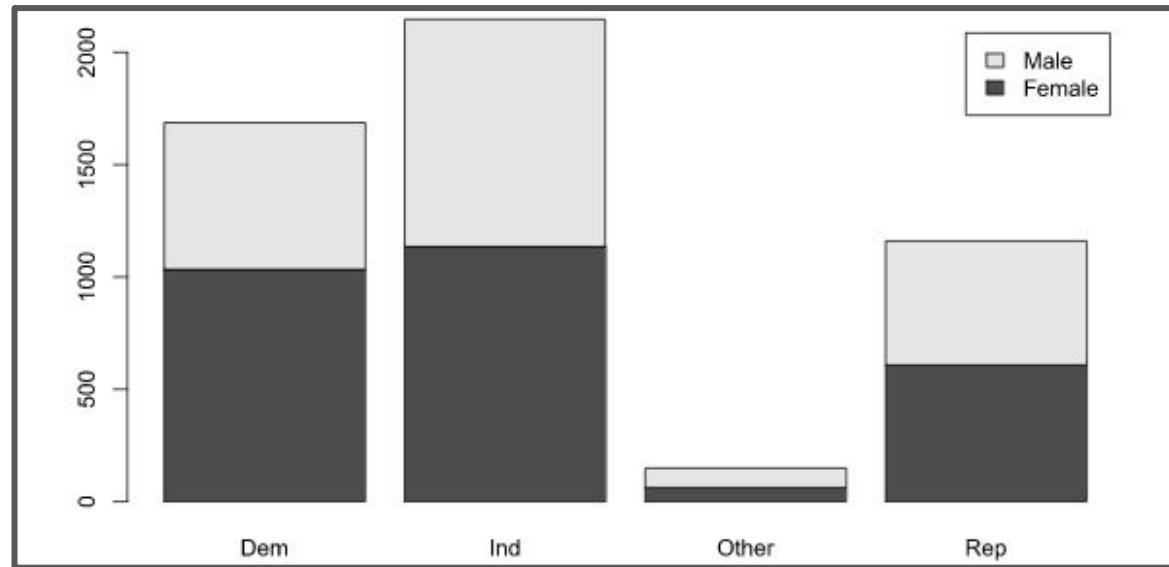


3

Discrete

Data Analysis

Bar Graph



## Quantitative Attributes



### Meaning of True Zero/ Absolute Zero



No Money

Absence of  
Property that is  
Money

Absolute/true/Meaningful zero means that the zero point represents the absence of the property being measured

## Quantitative Attributes



### Meaning of True Zero/ Absolute Zero

0  
celsius  
temperature

Temperature is  
present with  
value= 0 Celsius

Not Absolute/true/Meaningful zero means that  
the zero point , is the value of that property

## Quantitative Attributes



4

Continues/Numeric  
: Interval



### Interval Data Definition

Interval data is measured numerical data that has equal distances between adjacent values, but no meaningful zero

# Data Attribute

## Quantitative Attributes



4

Continues/Numeric  
: Interval



## INTERVAL DATA examples

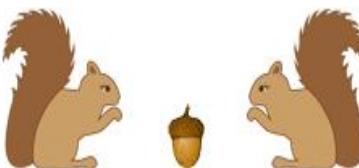
| Temperature | Time      | Dates | pH                 | IQ  |
|-------------|-----------|-------|--------------------|-----|
| 10°C        | 1 0'clock | 1066  | 2.5 (e.g. vinegar) | 80  |
| 20°C        | 2 0'clock | 1492  | 7 (e.g. water)     | 100 |
| 30°C        | 3 0'clock | 1776  | 12.5 (e.g. bleach) | 120 |

## Quantitative Attributes



4

Continues/Numeric  
: Interval



### Equidistance

## INTERVAL DATA

characteristics



Measured



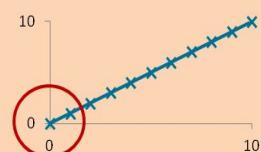
Ordered



Equidistant



Meaningful Zero



## Quantitative Attributes



4

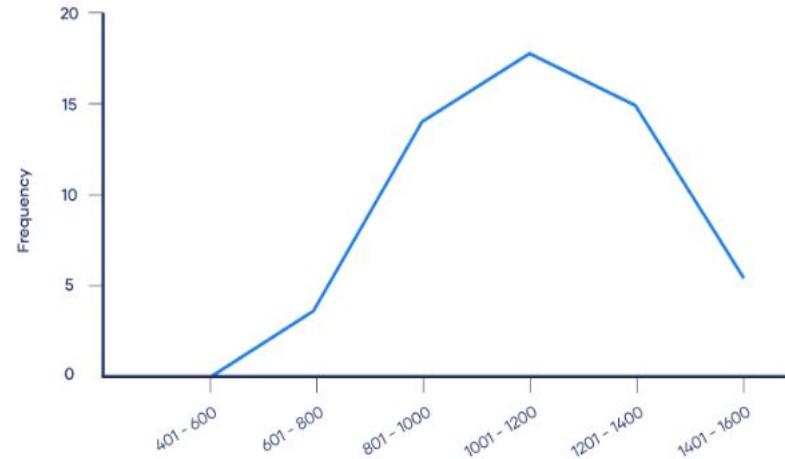
Continues/Numeric  
: Interval

Data Analysis

To organize your data, enter it into a grouped frequency distribution table.

| SAT score   | Frequency |
|-------------|-----------|
| 401 - 600   | 0         |
| 601 - 800   | 4         |
| 801 - 1000  | 15        |
| 1001 - 1200 | 19        |
| 1201 - 1400 | 16        |
| 1401 - 1600 | 5         |

Frequency distribution SAT scores



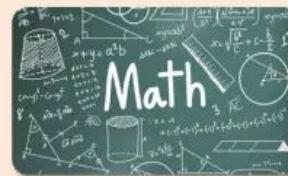
# Data Attribute

## Quantitative Attributes



4

Continues/Numeric  
: Interval



### INTERVAL DATA

Mathematical features

#### Grouping

$\neq$

Same /  
Different



#### Sorting

$<>$

Greater /  
Less Than



#### Difference

$+-$

Add /  
Subtract



#### Magnitude

$\times \div$

Multiply /  
Divide



## Quantitative Attributes



4

Continues/Numeric:  
Ratio Scaled



### Ratio Data Definition

Ratio data is measured numerical data  
that has equal distances between  
adjacent values and a meaningful zero

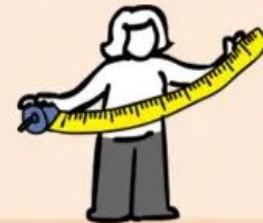
# Data Attribute

## Quantitative Attributes



4

Continues/Numeric  
: Interval



### RATIO DATA examples

| Age          | Temperature | Distance | Time Interval | Weight    |
|--------------|-------------|----------|---------------|-----------|
| 10 years old | 200 K       | 3 metres | 2.5 seconds   | 80 grams  |
| 20 years old | 300 K       | 6 metres | 7 seconds     | 100 grams |
| 30 years old | 400 K       | 9 metres | 12.5 seconds  | 120 grams |

## Quantitative Attributes



4

Continues/Numeric:  
Ratio Scaled



### RATIO DATA characteristics

Measured



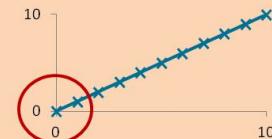
Ordered



Equidistant



Meaningful Zero



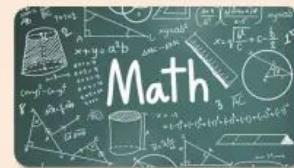
# Data Attribute

## Quantitative Attributes



4

Continues/Numeric:  
Ratio Scaled



### RATIO DATA

Mathematical features

#### Grouping

$\neq$

Same /  
Different

#### Sorting

$<>$

Greater /  
Less Than

#### Difference

$+-$

Add /  
Subtract

#### Magnitude

$\times \div$

Multiply /  
Divide



# Data Attribute

## Quantitative Attributes



I am type of Interval

10 \$ less  
Than Rajan

20 \$ /hour



Poojan

## Interval Vs Ratio

I am type of Ratio

66% greater  
than  
Poojan



Rajan

30 \$ /hour

## Quantitative Attributes



### Interval Vs Ratio

As part of a test preparation course, students are asked to take a practice version of the Graduate Record Examination (GRE). This is a standardized test.

- Scores can range from 200 to 800
- with a population mean of 500
- and a population standard deviation of 100.

Interval

## Quantitative Attributes



### Interval Vs Ratio

- If Frank is 20 years old and Paul is twice as old as Frank. How old will Paul be in the next 10 years?

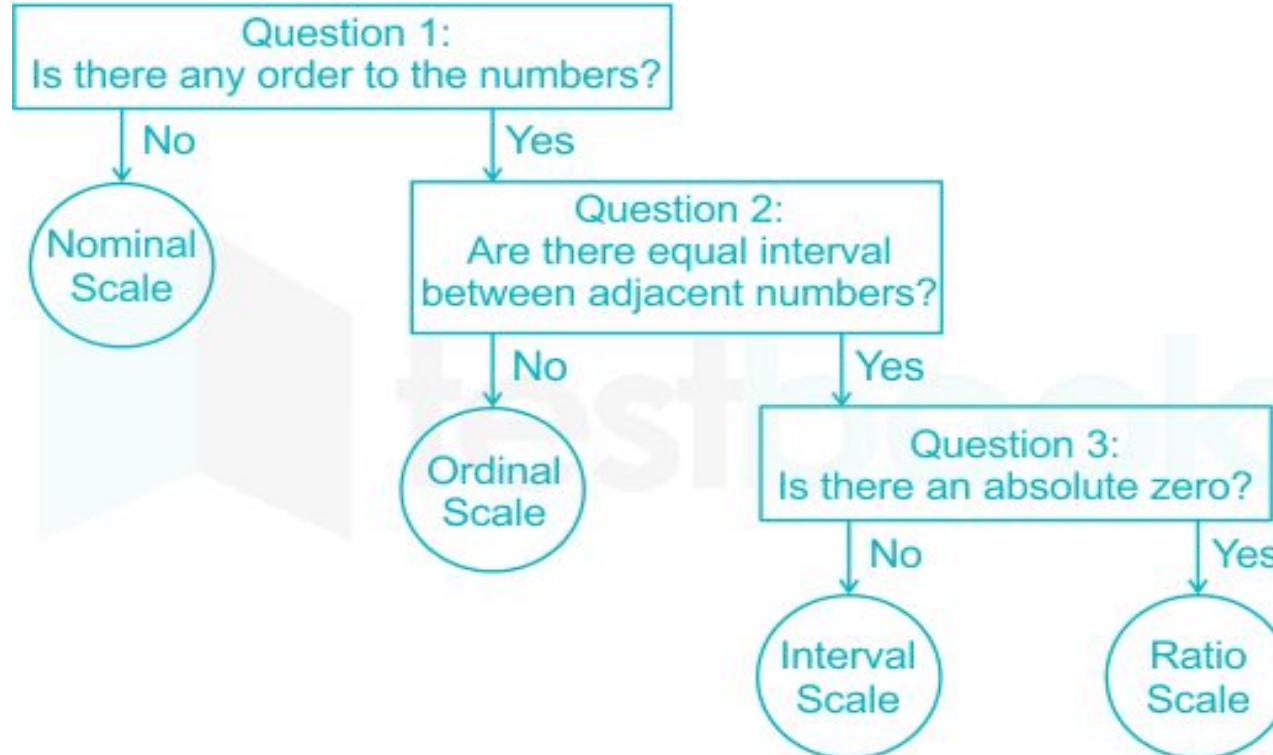
Ratio

# Need of Statistics

## Measures of Central Tendency



## Types of Data Attribute



# Need of Statistics

Measures of Central Tendency



## Case Study: Shopping Mall



# Need of Statistics

## Measures of Central Tendency



## Case Study: Shopping Mall



What is your Age?

I am type of  
Continuous

Can not count

Not equal distance

No Meaning zero

# Need of Statistics

## Measures of Central Tendency



## Case Study: Shopping Mall

What is your Marital Status

I am type of  
Nominal

Not in order

Not equal distance

No Meaning zero

# Need of Statistics

## Measures of Central Tendency



## Case Study: Shopping Mall



I am type of  
Ratio

equal distance

Meaning zero

# Need of Statistics

Measures of Central Tendency



## Case Study: Shopping Mall

Select your Income level scale

(10K, 20K, 30K, 40K, 50K, 60K, 70K, 80K, 90K)



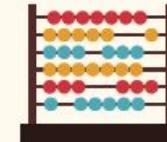
I am type of  
Interval

In order

Equal distance

No Meaning zero

# Key points to focus

| Points                    | Discrete Data  | Continuous Data                                 |   |
|---------------------------|--|---|---|
| Meaning                   | Discrete data has clear spaces between values.       | Continuous data falls on a continuous sequence. |  |
| Can you count the data?   | Yes, data is usually units counted in whole numbers. | Generally, NO                                   |  |
| Can you measure the data? | NO   | YES   |  |

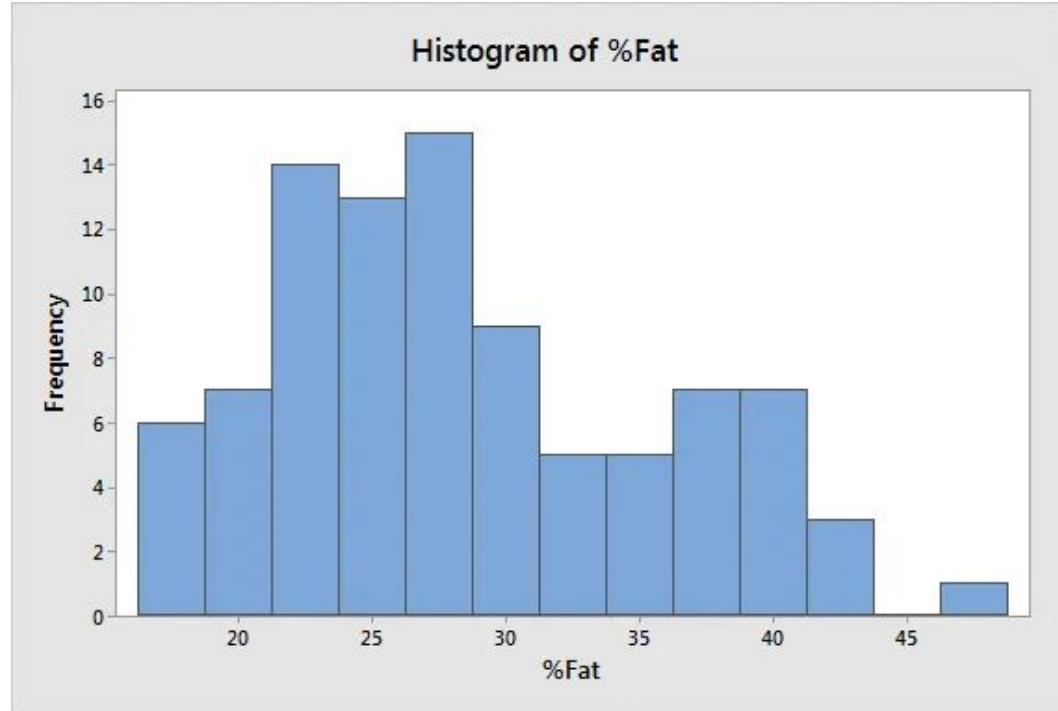
# Key points to focus

| Points                   | Discrete Data   | Continuous Data  |   |      |       |        |    |        |    |        |    |        |    |        |    |
|--------------------------|---|--|---|------|-------|--------|----|--------|----|--------|----|--------|----|--------|----|
| Values                   | It has a finite number of possible values. The values cannot be divided into smaller pieces and add additional meaning. | It has an infinite number of possible values within an interval. The values can be subdivided into smaller and smaller pieces. |    |      |       |        |    |        |    |        |    |        |    |        |    |
| Graphical Representation | Bar Chart   | Histogram  |  <table border="1"><caption>Data represented in the Histogram</caption><thead><tr><th>Item</th><th>Value</th></tr></thead><tbody><tr><td>Item 1</td><td>10</td></tr><tr><td>Item 2</td><td>20</td></tr><tr><td>Item 3</td><td>30</td></tr><tr><td>Item 4</td><td>40</td></tr><tr><td>Item 5</td><td>50</td></tr></tbody></table> | Item | Value | Item 1 | 10 | Item 2 | 20 | Item 3 | 30 | Item 4 | 40 | Item 5 | 50 |
| Item                     | Value   |  |   |      |       |        |    |        |    |        |    |        |    |        |    |
| Item 1                   | 10  |  |   |      |       |        |    |        |    |        |    |        |    |        |    |
| Item 2                   | 20  |  |   |      |       |        |    |        |    |        |    |        |    |        |    |
| Item 3                   | 30  |  |   |      |       |        |    |        |    |        |    |        |    |        |    |
| Item 4                   | 40  |  |   |      |       |        |    |        |    |        |    |        |    |        |    |
| Item 5                   | 50  |  |   |      |       |        |    |        |    |        |    |        |    |        |    |

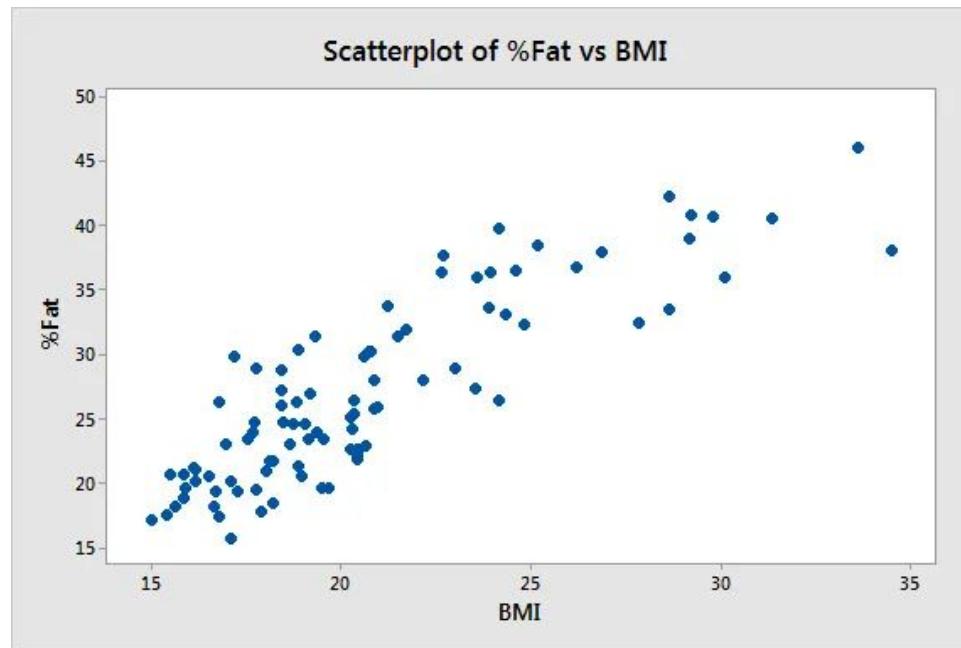
# Key points to focus

| Points   | Discrete Data  | Continuous Data   |
|----------|--|---|
| Examples | <ul style="list-style-type: none"><li>The number of students in a class.</li><li>The number of workers in a company.</li><li>The number of parts damaged during transportation.</li><li>Shoe sizes.</li><li>Number of languages an individual speaks.</li><li>The number of home runs in a baseball game.</li><li>The number of test questions you answered correctly.</li></ul> | <ul style="list-style-type: none"><li>The amount of time required to complete a project.</li><li>The height of children.</li><li>The amount of time it takes to sell shoes.</li><li>The amount of rain, in inches, that falls in a storm.</li><li>The square footage of a two-bedroom house.</li><li>The weight of a truck.</li><li>The speed of cars.</li><li>Time to wake up.</li></ul> <br> |

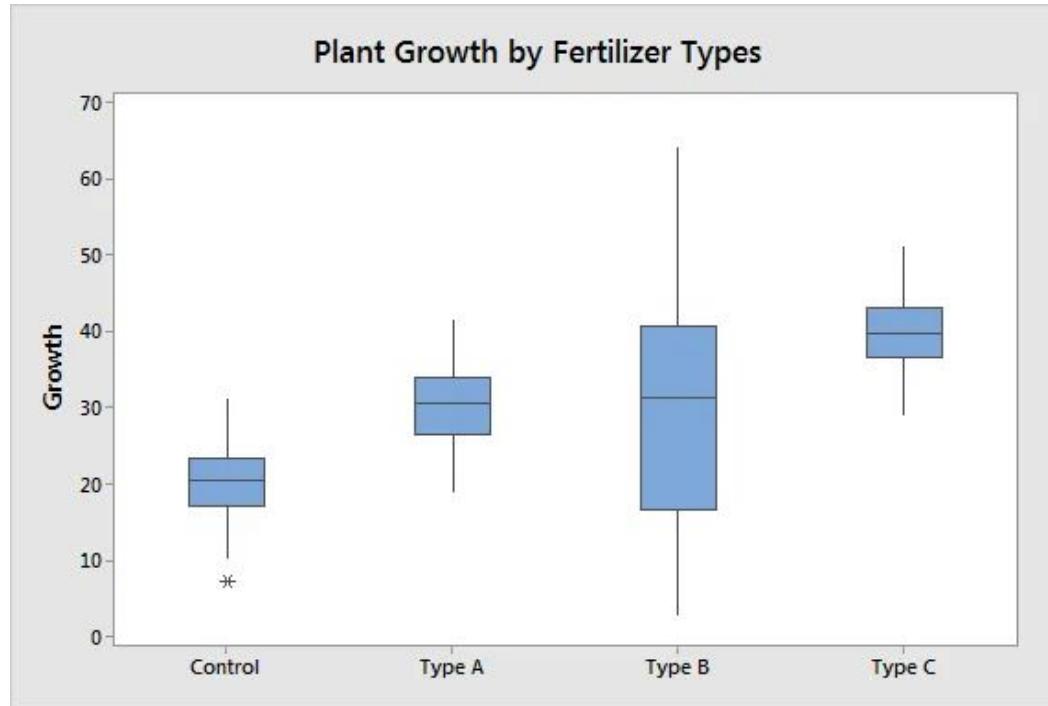
continuous variable



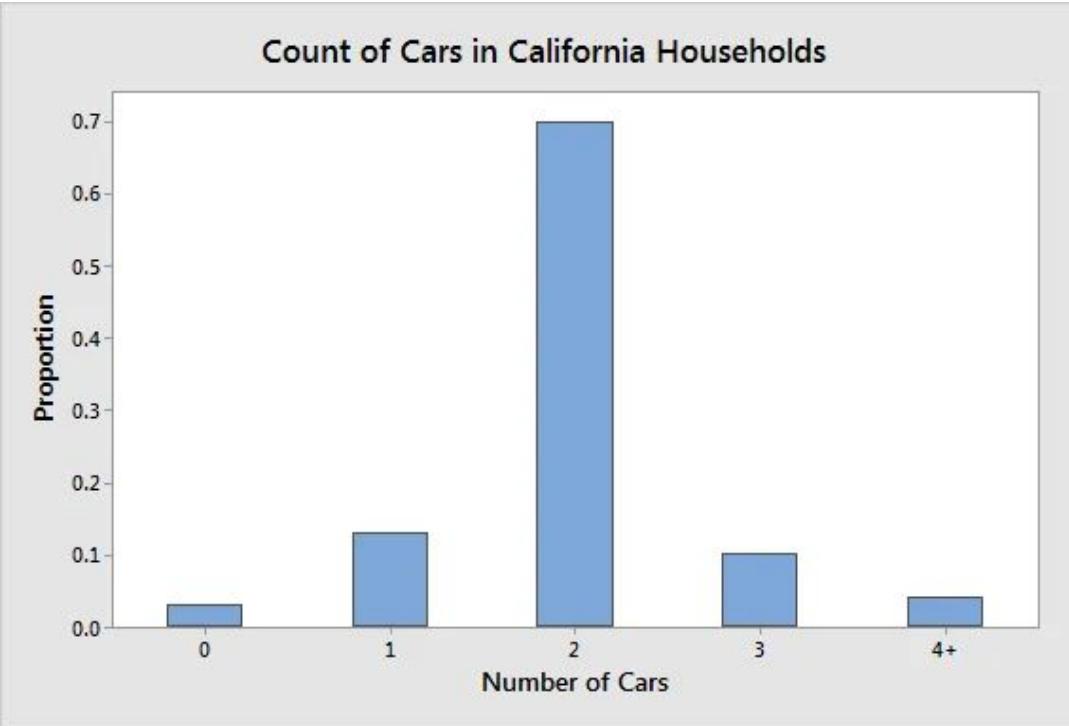
two continuous variable



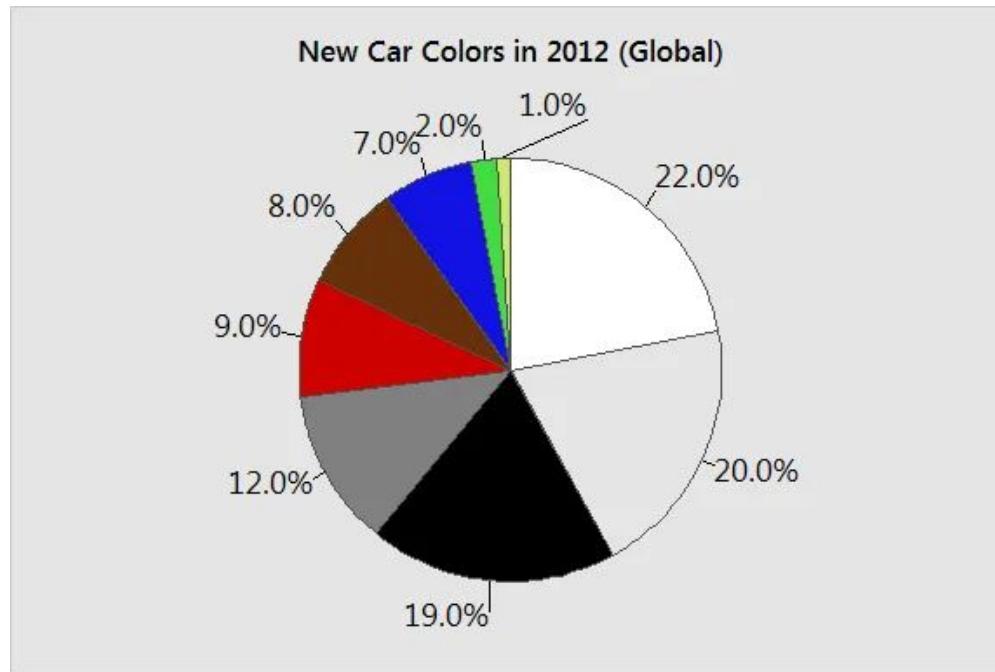
Groupwise continuous variable



## Discrete data

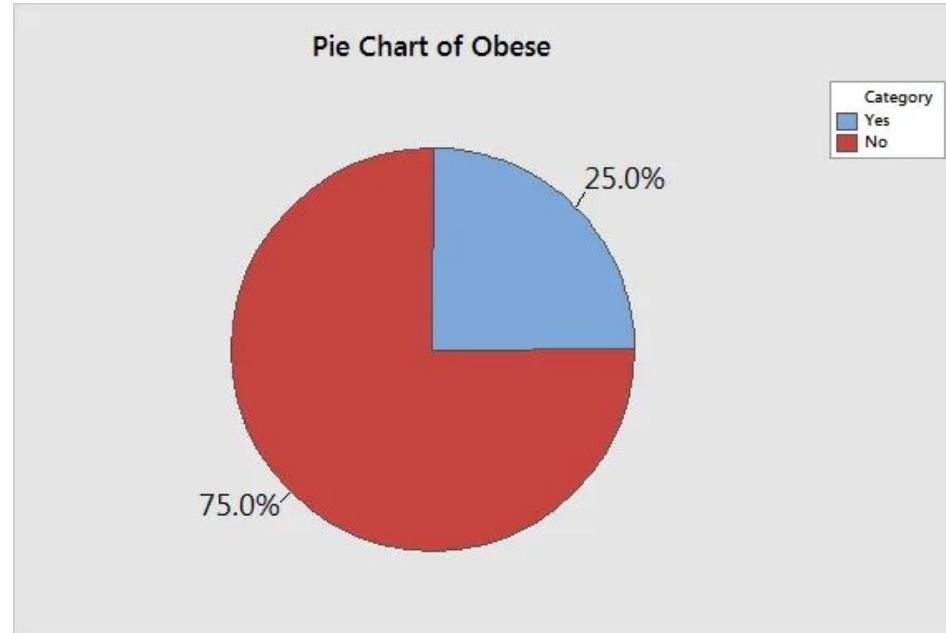


## Categorical data



| Color  |
|--------|
| White  |
| Silver |
| Black  |
| Gray   |
| Red    |

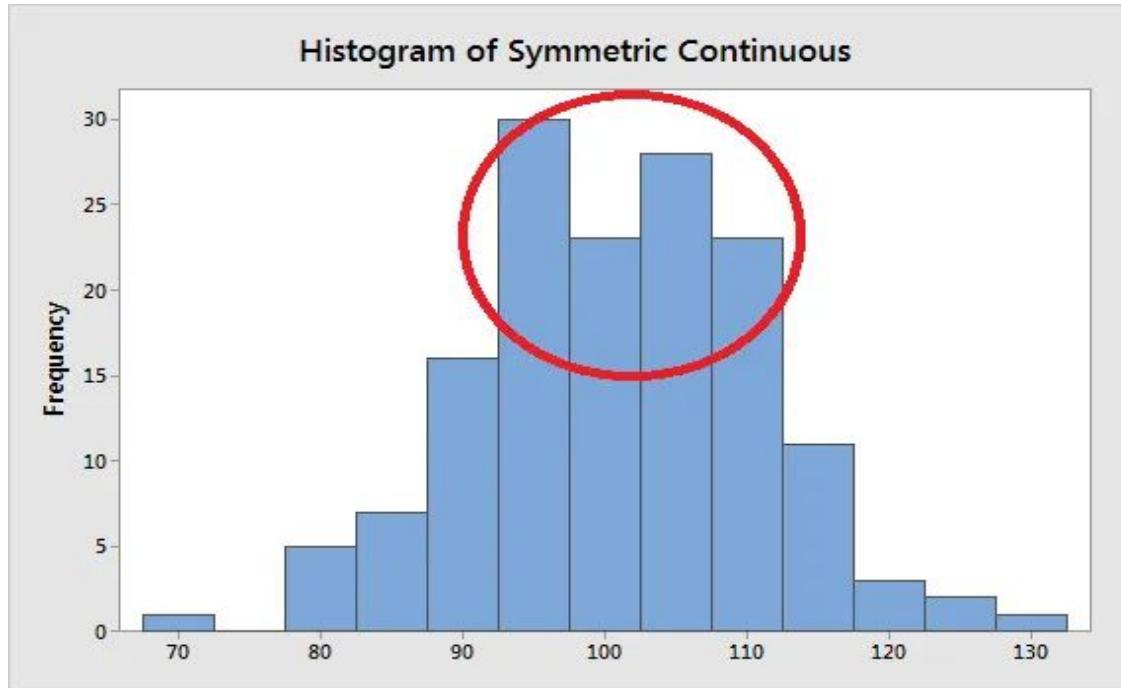
## Binary data



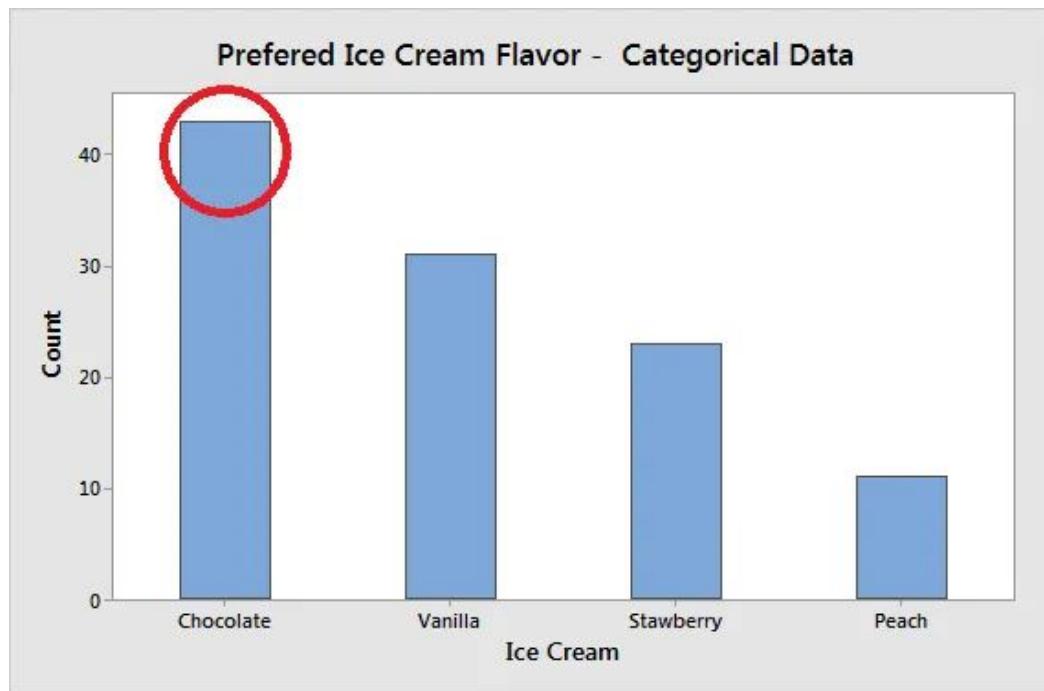
## Ordinal data



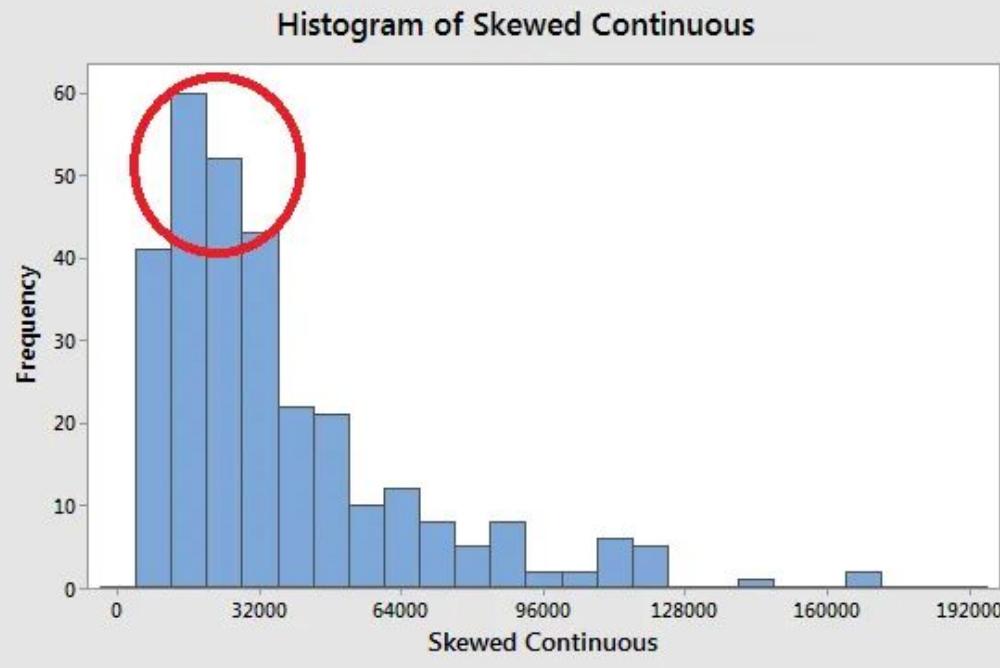
## Locating the Center of Your Data



## Locating the Center of Your Data



## Locating the Center of Your Data



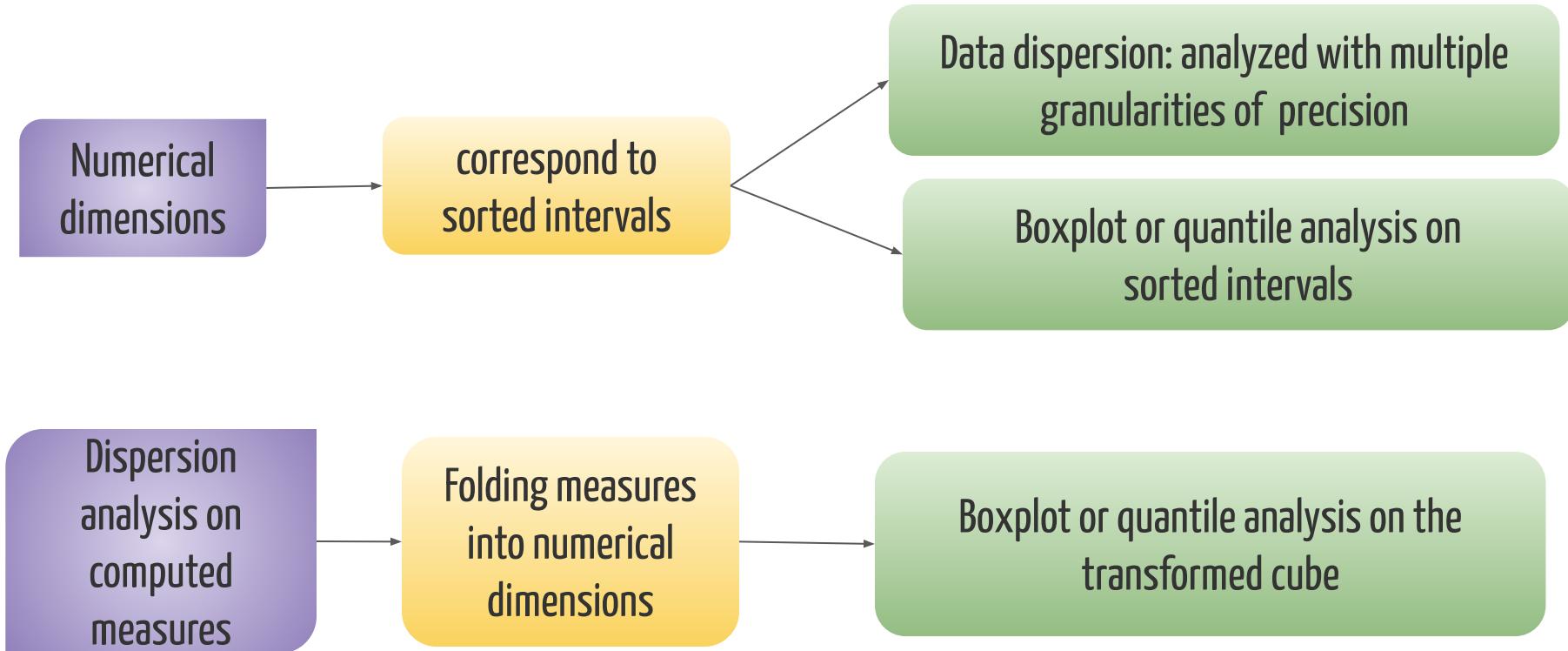
Motivation

To better understand  
the data

central tendency, variation  
and spread

Data dispersion  
characteristics

median, max, min, quantiles,  
outliers, variance, etc.



# Need of Statistics

## Measures of Central Tendency



## Types of Data Attribute

Central Tendency

Mean

Median

Mode

## Measures of Central Tendency



1

Mean

$$2 + 2 + 5 + 6 + 7 + 8 = 30$$

$$30 \div 6 = 5$$

The mean  
number is

5

# Need of Statistics

## Measures of Central Tendency



1

### Mean

- The mean represents the average value of the dataset.
- $n$  is sample size and  $N$  is population size.
- $$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

$$\mu = \frac{\sum x}{n}$$

# Need of Statistics

## Measures of Central Tendency



1

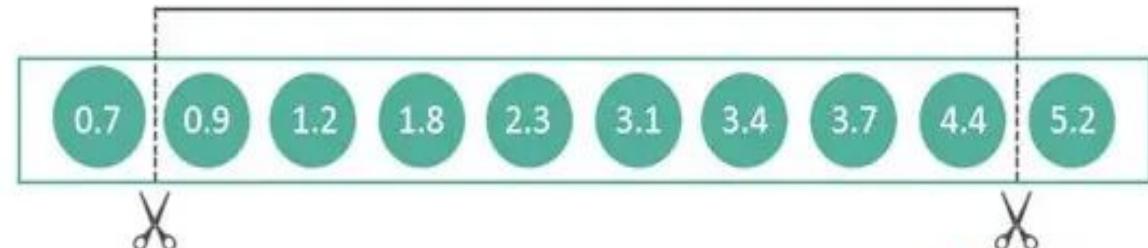
### Mean

**Loss of Information**

## Trimmed Mean

"Trimmed mean is a central tendency measure that cuts down the smallest and highest values before applying the standard averaging formula for greater accuracy."

10% Trimmed Mean = 2.6



- 10% samples will be cut down from each side
- Three commonly applied trim percentages, i.e., 5%, 10%, and 20%

# Need of Statistics

## Measures of Central Tendency



1

### Mean

## Trimmed Mean

| Staff  | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  |
|--------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Salary | 15k | 18k | 16k | 14k | 15k | 15k | 12k | 17k | 90k | 95k |

### Arithmetic Mean

$$= (15+18+16+14+15+12+17+90+95) / 10$$

$$= 307 / 10$$

$$= 30.7$$

- Average salary by observation is 12k to 18 k
- It is not the best way to accurately reflect the typical salary of a worker

# Need of Statistics

## Measures of Central Tendency



1

**Mean**

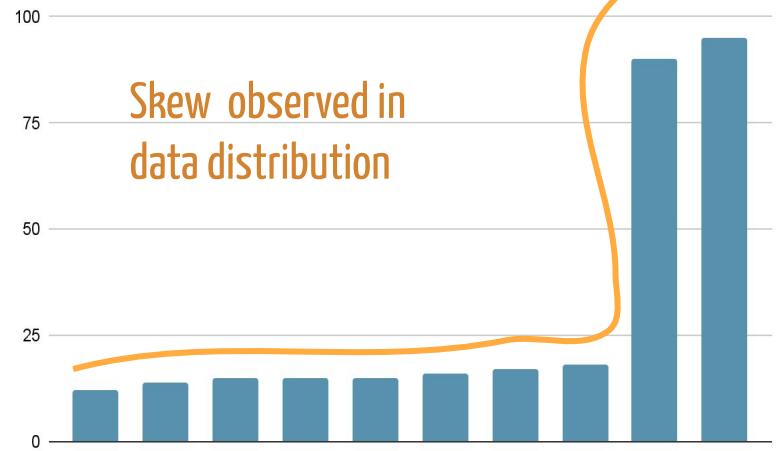
- Sort in descending order

12    14    15    15    16    17    18    90    95

## When mean fails...

| Staff  | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  |
|--------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Salary | 15k | 18k | 16k | 14k | 15k | 15k | 12k | 17k | 90k | 95k |

Points scored



# Need of Statistics

## Measures of Central Tendency

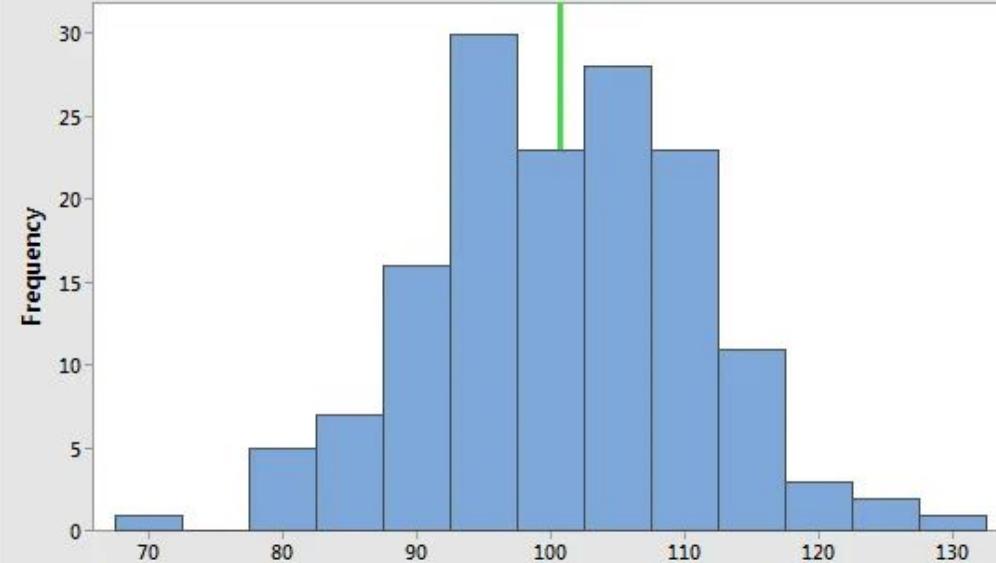


1

### Mean

Histogram of Symmetric Continuous

Mean 100.67



# Need of Statistics

## Measures of Central Tendency

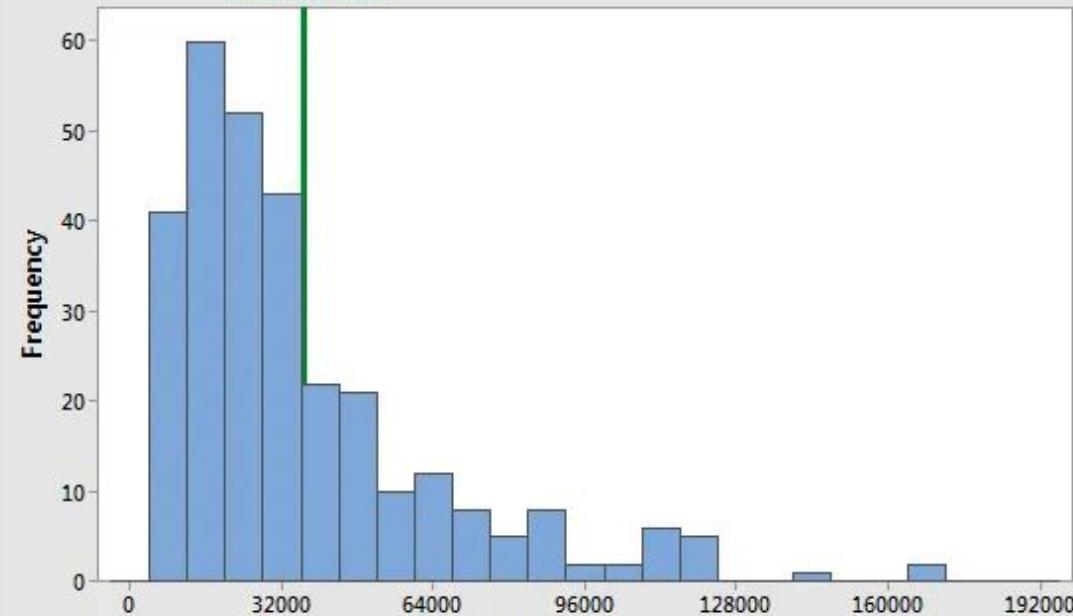


1

### Mean

Histogram of Skewed Continuous

Mean 36624



## Measures of Central Tendency



2

### Median

middle  
value

of

the dataset

Arranged in

ascending order or  
in descending order

| Median odd |
|------------|
| 23         |
| 21         |
| 18         |
| 16         |
| 15         |
| 13         |
| 12         |
| 10         |
| 9          |
| 7          |
| 6          |
| 5          |
| 2          |

| Median even |
|-------------|
| 40          |
| 38          |
| 35          |
| 33          |
| 32          |
| 30          |
| 29          |
| 27          |
| 26          |
| 24          |
| 23          |
| 22          |
| 19          |
| 17          |

## Measures of Central Tendency



2

### Median

Odd number of Samples:

$$\text{Median} = \text{value of } (n+1/2)^{\text{th}} \text{ observation}$$

Even number of Samples:

$$\text{Median} = \frac{\text{value of } \left(\frac{n}{2}\right)^{\text{th}} \text{ observation} + \text{value of } \left(\frac{n}{2}+1\right)^{\text{th}} \text{ observation}}{2}$$

[https://www.brainkart.com/article/Median\\_35083/](https://www.brainkart.com/article/Median_35083/)

<https://statisticsbyjim.com/basics/measures-central-tendency-mean-median-mode/>

# Need of Statistics

## Measures of Central Tendency



2

### Median

The number of rooms in the seven five stars hotel in Chennai city is 71, 30, 61, 59, 31, 40 and 29. Find the median number of rooms:

**Step 1**

Arrange the data in ascending order : 29, 30, 31, 40, 59, 61, 71

**Step 2**

$n = 7$  (odd)

**Step 3**

$\text{Median} = 7+1 / 2 = 4\text{th positional value}$

**Step 4**

Median = 40 rooms

## Measures of Central Tendency



2

## Median

Median for Discrete grouped data:

- i. Calculate the cumulative frequencies
- ii. Find  $(N+1)/2$ , where  $N=\sum f$ =total frequencies
- iii. Identify the cumulative frequency just greater than  $(N+1)/2$
- iv. The value of  $x$  corresponding to that cumulative frequency is the  $(N+1)/2$  median

# Need of Statistics

## Measures of Central Tendency



2

### Median

The following data are the weights of students in a class. Find the median weights of the students

| Weight(kg)         | 10 | 20 | 30 | 40 | 50 | 60 | 70 |
|--------------------|----|----|----|----|----|----|----|
| Number of Students | 4  | 7  | 12 | 15 | 13 | 5  | 4  |

| Weight (kg)<br>$x$ | Frequency<br>$f$ | Cumulative Frequency<br>$c.f$ |
|--------------------|------------------|-------------------------------|
| 10                 | 4                | 4                             |
| 20                 | 7                | 11                            |
| 30                 | 12               | 23                            |
| 40                 | 15               | 38                            |
| 50                 | 13               | 51                            |
| 60                 | 5                | 56                            |
| 70                 | 4                | 60                            |
| Total              | N = 60           |                               |

# Need of Statistics

| Weight (kg)<br><i>x</i> | Frequency<br><i>f</i> | Cumulative Frequency<br><i>c.f</i> |
|-------------------------|-----------------------|------------------------------------|
| 10                      | 4                     | 4                                  |
| 20                      | 7                     | 11                                 |
| 30                      | 12                    | 23                                 |
| 40                      | 15                    | 38                                 |
| 50                      | 13                    | 51                                 |
| 60                      | 5                     | 56                                 |
| 70                      | 4                     | 60                                 |
| Total                   | N = 60                |                                    |

**Step 1**

N= 60

**Step 2**

$(N+1)/2 = (60+1)/2 = 30.5$

**Step 3**

Cumulative frequency >30.5 is 38

**Step 4**

Value of x corresponding to 38 is 40

**Step 5**

The median weight of students is 40

## Measures of Central Tendency



2

### Median

Median for Continuous grouped data:

$$\text{Median} = l + \frac{\frac{N}{2} - m}{f} \times c$$

$l$  = Lower limit of the median class

$N$  = Total Numbers of frequencies

$f$  = Frequency of the median class

$m$  = Cumulative frequency of the class preceding the median class

$c$  = the class interval of the median class

Note: one has to find the median class first. Median class is, that class which correspond to the cumulative frequency just greater than  $N/2$ .

# Need of Statistics

## Measures of Central Tendency



2

### Median

The following data obtained from a garden records of certain period  
Calculate the median weight of the apple

| Weight in grams  | 410 – 420 | 420 – 430 | 430 – 440 | 440 – 450 | 450 – 460 | 460 – 470 | 470 – 480 |
|------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| Number of apples | 14        | 20        | 42        | 54        | 45        | 18        | 7         |

| Weight in grams | Number of apples | Cumulative Frequency |
|-----------------|------------------|----------------------|
| 410 – 420       | 14               | 14                   |
| 420 – 430       | 20               | 34                   |
| 430 – 440       | 42               | 76                   |
| 440 – 450       | 54               | 130                  |
| 450 – 460       | 45               | 175                  |
| 460 – 470       | 18               | 193                  |
| 470 – 480       | 7                | 200                  |
| Total           | N = 200          |                      |

# Need of Statistics

| Weight in grams | Number of apples | Cumulative Frequency |
|-----------------|------------------|----------------------|
| 410 – 420       | 14               | 14                   |
| 420 – 430       | 20               | 34                   |
| 430 – 440       | 42               | 76                   |
| 440 – 450       | 54               | 130                  |
| 450 – 460       | 45               | 175                  |
| 460 – 470       | 18               | 193                  |
| 470 – 480       | 7                | 200                  |
| <b>Total</b>    | <b>N = 200</b>   |                      |

**Step 1**

$$N/2 = 200/2 = 100$$

**Step 2**

Median class id 440-450  
As Frequency > 100

**Step 3**

$l$  = lower boundary of 440-450 = 440

**Step 4**

$m$  = cumulative frequency of 430-440  
 $m = 76$

**Step 5**

$c$  = Interval of 440-450

**Step 6**

$f$  = frequency of 440-450 = 54

| Weight in grams | Number of apples | Cumulative Frequency |
|-----------------|------------------|----------------------|
| 410 – 420       | 14               | 14                   |
| 420 – 430       | 20               | 34                   |
| 430 – 440       | 42               | 76                   |
| 440 – 450       | 54               | 130                  |
| 450 – 460       | 45               | 175                  |
| 460 – 470       | 18               | 193                  |
| 470 – 480       | 7                | 200                  |
| <b>Total</b>    | <b>N = 200</b>   |                      |

**Step 7**

$$\begin{aligned}\text{Mode} &= 440 + ((100-76)/54) * 10 \\ &= 440 + 4.44 \\ &= 444.44\end{aligned}$$

## Measures of Central Tendency



3

**Mode**

Most frequent  
value

of

the dataset

Around which

Most items tend to  
be most heavily  
concentrated

**Mode**

5

5

5

4

4

3

2

2

1

# Need of Statistics

## Measures of Central Tendency



3

**Mode**

Two wheelers are more than cars.

Because of higher frequency the modal value of this survey is

**'two wheelers'**

# Need of Statistics

## Measures of Central Tendency



3

### Mode

The following are the marks scored by 20 students in the class. Find the mode

90, 70, 50, 30, 40, 86, 65, 73, 68, 90, 90, 10, 73, 25, 35, 88, 67, 80, 74, 46

The marks 90 occurs the maximum number of times

**Mode=90**

# Need of Statistics

## Measures of Central Tendency



3

**Mode**

A doctor who checked 9 patients' sugar level is given below. Find the mode value of the sugar levels

80, 112, 110, 115, 124, 130, 100, 90, 150, 180

Each values occurs only once

**there is no mode**

## Measures of Central Tendency



3

Mode

Compute mode value for the following observations.

7, 10, 12, 10, 19, 2, 11, 3, 12

the observations 10 and 12 occurs twice in the data set

the modes are 10 and 12

# Need of Statistics

## Measures of Central Tendency



3

**Mode**

Calculate the mode from the following data

| Days of Confinement | 6 | 7 | 8 | 9 | 10 |
|---------------------|---|---|---|---|----|
| Number of patients  | 4 | 6 | 7 | 5 | 3  |

7 is the maximum frequency

the value of x corresponding to 7 is 8

**Mode=8**

# Need of Statistics

## Measures of Central Tendency



3

### Mode

#### Mode for Continuous data

$$\text{Mode} = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times c$$

Modal class is the class which has maximum frequency.

$f_1$  = frequency of the modal class

$f_0$  = frequency of the class preceding the modal class

$f_2$  = frequency of the class succeeding the modal class

$c$  = width of the class limits

# Need of Statistics

## Measures of Central Tendency



3

### Mode

The given data relates to the daily income of families in an urban area. Find the modal income of the families.

| Income (`)    | 0-100 | 100-200 | 200-300 | 300-400 | 400-500 | 500-600 | 600-700 |
|---------------|-------|---------|---------|---------|---------|---------|---------|
| No.of persons | 5     | 7       | 12      | 18      | 16      | 10      | 5       |

| Income (`) | No.of persons ( $f$ ) |
|------------|-----------------------|
| 0-100      | 5                     |
| 100-200    | 7                     |
| 200-300    | 12 $f_0$              |
| 300-400    | 18 $f_1$              |
| 400-500    | 16 $f_2$              |
| 500-600    | 10                    |
| 600-700    | 5                     |

# Need of Statistics

| Income (`) | No.of persons ( $f$ ) |
|------------|-----------------------|
| 0-100      | 5                     |
| 100-200    | 7                     |
| 200-300    | 12 $f_0$              |
| 300-400    | 18 $f_1$              |
| 400-500    | 16 $f_2$              |
| 500-600    | 10                    |
| 600-700    | 5                     |

**Step 1**

Highest Frequency is 18. Modal class is 300-400

**Step 2**

$l$  = lower boundary of 300-400 = 300

**Step 3**

$f_1$  = frequency of 300-400 = 18

**Step 4**

$f_0$  = frequency of 200-300 = 12

**Step 5**

$f_2$  = frequency of 400-500 = 16

**Step 6**

$$\begin{aligned} \text{Mode} &= 300 + (18-12)/(2*18-12-16)*100 \\ &= 300 + 6/(36-28)*100 \\ &= 300 + 600/8 = 300 + 75 = 375 \end{aligned}$$

**Data Attributes and****Measure of Central Tendency****Levels of measurement**

| Scale    | Mode | Median | Mean |
|----------|------|--------|------|
| Nominal  | √    |        |      |
| Ordinal  | √    | √      |      |
| Interval | √    | √      | √    |
| Ratio    | √    | √      | √    |

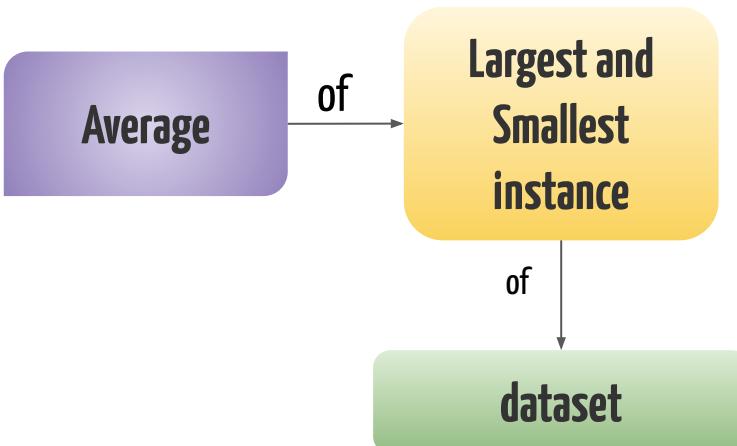
# Need of Statistics

## Measures of Central Tendency



4

### Mid Range



#### Example

##### Problem

**Find the range and midrange for the following set of numbers: 2, 4, 7, 10, 14, 35.**

range:  $35 - 2 = 33$  Subtract the least value from the greatest value to find the range.

midrange: Add together the greatest value and the least value and divide by 2.  
$$\frac{35 + 2}{2} = \frac{37}{2} = 18.5$$

##### Answer

The range is 33.  
The midrange is 18.5.

# Relationship among mean, median and mode

| No of days spend in training |    |
|------------------------------|----|
| Team 1                       | 4  |
| Team 2                       | 5  |
| Team 3                       | 6  |
| Team 4                       | 6  |
| Team 5                       | 6  |
| Team 6                       | 7  |
| Team 7                       | 7  |
| Team 8                       | 7  |
| Team 9                       | 7  |
| Team 10                      | 7  |
| Team 11                      | 7  |
| Team 12                      | 8  |
| Team 13                      | 8  |
| Team 14                      | 8  |
| Team 15                      | 9  |
| Team 16                      | 10 |

| No of days spend in training |   |
|------------------------------|---|
| Team 1                       | 4 |
| Team 2                       | 5 |
| Team 3                       | 6 |
| Team 4                       | 6 |
| Team 5                       | 6 |
| Team 6                       | 7 |
| Team 7                       | 7 |
| Team 8                       | 7 |
| Team 9                       | 7 |
| Team 10                      | 8 |

| No of days spend in training |    |
|------------------------------|----|
| Team 1                       | 6  |
| Team 2                       | 7  |
| Team 3                       | 7  |
| Team 4                       | 7  |
| Team 5                       | 7  |
| Team 6                       | 8  |
| Team 7                       | 8  |
| Team 8                       | 8  |
| Team 9                       | 9  |
| Team 10                      | 10 |

# Relationship among mean, median and mode

## No of days spend in training

Team 1

4

Team 2

5

Team 3

6

Team 4

6

Team 5

6

Team 6

7

Team 7

7

Team 8

7

Team 9

7

Team 10

7

Team 11

7

Team 12

8

Team 13

8

Team 14

8

Team 15

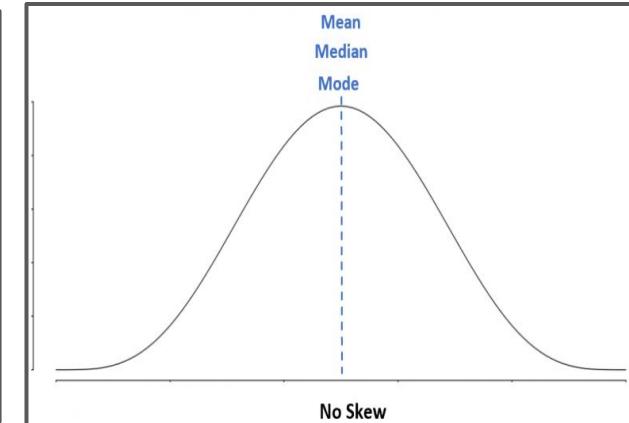
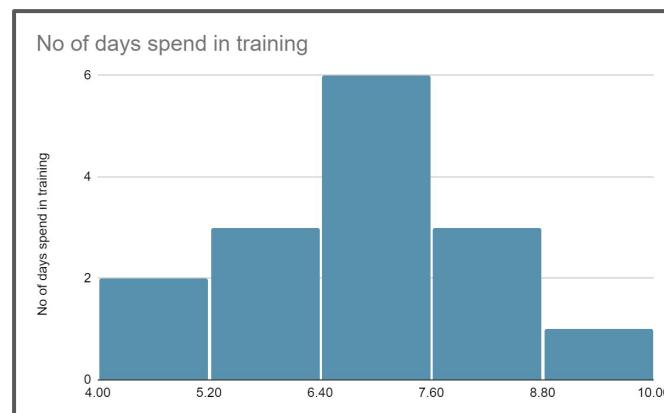
9

Team 16

10

|        |   |
|--------|---|
| Mean   | 7 |
| Median | 7 |
| Mode   | 7 |

Mean= Median=Mode



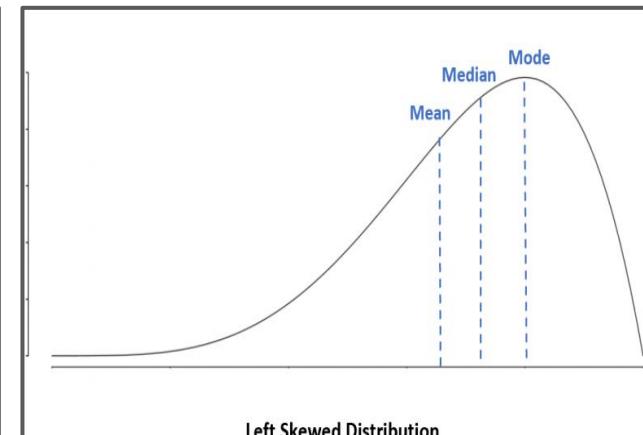
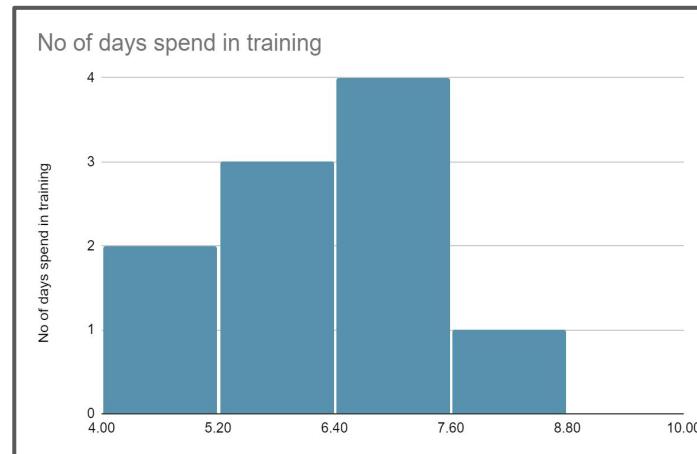
Symmetric Data Distribution

# Relationship among mean, median and mode

| No of days spend in training |   |
|------------------------------|---|
| Team 1                       | 4 |
| Team 2                       | 5 |
| Team 3                       | 6 |
| Team 4                       | 6 |
| Team 5                       | 6 |
| Team 6                       | 7 |
| Team 7                       | 7 |
| Team 8                       | 7 |
| Team 9                       | 7 |
| Team 10                      | 8 |

|        |     |
|--------|-----|
| Mean   | 6.3 |
| Median | 6.5 |
| Mode   | 7   |

Mean < Median < Mode



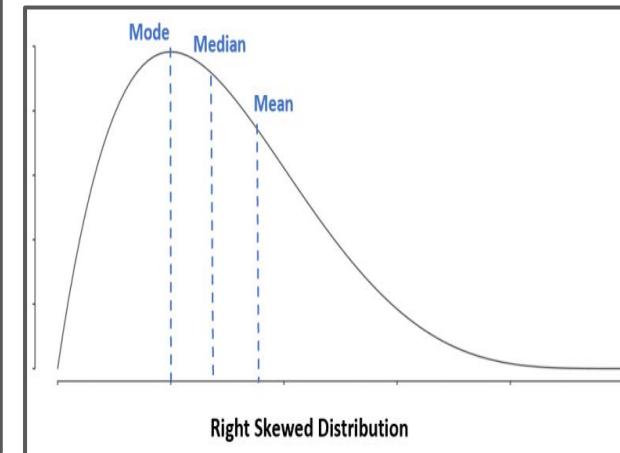
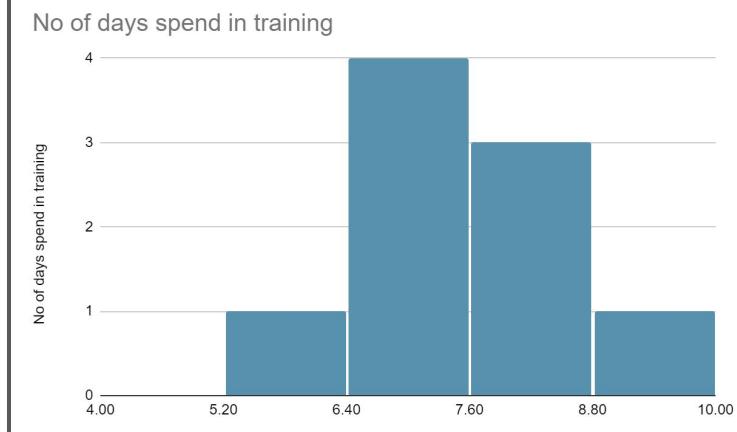
Left Skew data distribution

# Relationship among mean, median and mode

| No of days spend in training |    |
|------------------------------|----|
| Team 1                       | 6  |
| Team 2                       | 7  |
| Team 3                       | 7  |
| Team 4                       | 7  |
| Team 5                       | 7  |
| Team 6                       | 8  |
| Team 7                       | 8  |
| Team 8                       | 8  |
| Team 9                       | 9  |
| Team 10                      | 10 |

|        |     |
|--------|-----|
| Mean   | 7.7 |
| Median | 7.5 |
| Mode   | 7   |

Mode < Median < Mean



Right Skew data distribution

## Measures of Central Tendency



Empirical Relationship among mean, median and mode

$$\text{Mean} - \text{Median} = \frac{1}{3}(\text{Mean} - \text{Mode})$$

$$7.7 - 7.5 = \frac{1}{3}(7.7 - 7)$$

$$0.2 \Rightarrow 0.26$$

For Right Skew Data

|        |     |
|--------|-----|
| Mean   | 7.7 |
| Median | 7.5 |
| Mode   | 7   |

## Measures of Central Tendency



Empirical Relationship among mean, median and mode

$$\text{Mean} - \text{Median} = \frac{1}{3}(\text{Mean} - \text{Mode})$$

$$6.3 - 6.5 = \frac{1}{3}(6.3 - 7)$$

$$(-0.2) \approx (-0.26)$$

For Left Skew Data

|        |     |
|--------|-----|
| Mean   | 6.3 |
| Median | 6.5 |
| Mode   | 7   |

# Need of Statistics

## Measures of Central Tendency



Empirical Relationship among mean, median and mode

$$\text{Mean} - \text{Median} = \frac{1}{3}(\text{Mean} - \text{Mode})$$

$$\text{mean} - \text{mode} = 3 (\text{mean} - \text{median})$$

## Measures of Dispersion



01

RANGE



The range can measure by subtracting the lowest value from the massive Number

The wide range indicates high variability, and the small range specifies low variability in the distribution.

Range = Highest\_value – Lowest\_value

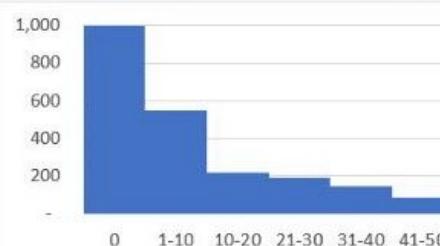
01

## RANGE

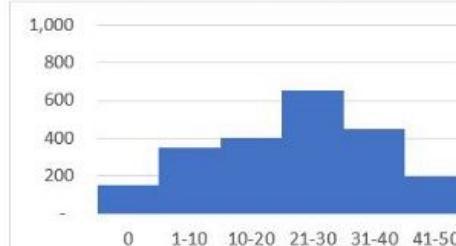


The wide range indicates high variability, and the small range specifies low variability in the distribution.

| Distribution 1 |              |
|----------------|--------------|
| Value of X     | Frequency    |
| 0              | 1,000        |
| 1-10           | 550          |
| 10-20          | 220          |
| 21-30          | 190          |
| 31-40          | 150          |
| 41-50          | 90           |
| <b>Total</b>   | <b>2,200</b> |



| Distribution 2 |              |
|----------------|--------------|
| Value of X     | Frequency    |
| 0              | 150          |
| 1-10           | 350          |
| 10-20          | 400          |
| 21-30          | 650          |
| 31-40          | 450          |
| 41-50          | 200          |
| <b>Total</b>   | <b>2,200</b> |

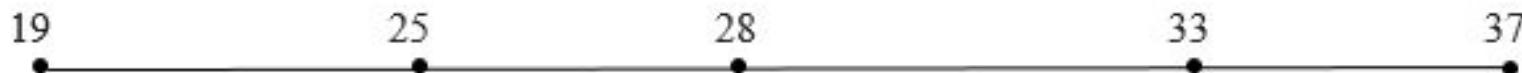


01

RANGE



| Student_id | 1  | 2  | 3  | 4  | 5  |
|------------|----|----|----|----|----|
| Marks      | 37 | 33 | 19 | 25 | 28 |



$$\begin{aligned}\text{Range} &= H - L \\ &= 37 - 19 \longrightarrow 18\end{aligned}$$

RANGE OF SEQUENCE IS 18

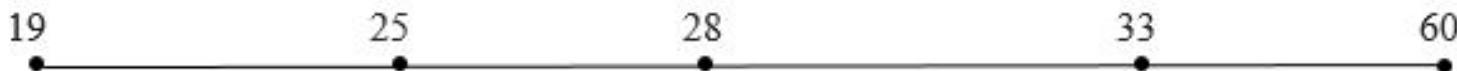
# Measures of Dispersion

01

RANGE

**RANGE CAN INFLUENCE BY OUTLIERS**

| Student_id | 1  | 2  | 3  | 4  | 5  |
|------------|----|----|----|----|----|
| Marks      | 60 | 33 | 19 | 25 | 28 |



$$\text{Range} = H - L$$

$$= 60 - 19 \implies 41 \quad \text{Now range of marks is 41.}$$

**RANGE OF SEQUENCE IS 18**

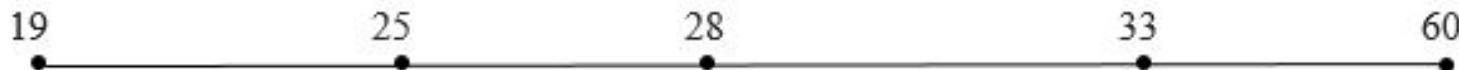
# Measures of Dispersion

01

RANGE

**RANGE CAN INFLUENCE BY OUTLIERS**

| Student_id | 1  | 2  | 3  | 4  | 5  |
|------------|----|----|----|----|----|
| Marks      | 60 | 33 | 19 | 25 | 28 |



$$\text{Range} = H - L$$

$$= 60 - 19 \implies 41 \quad \text{Now range of marks is 41.}$$

**RANGE OF SEQUENCE IS 18**

01

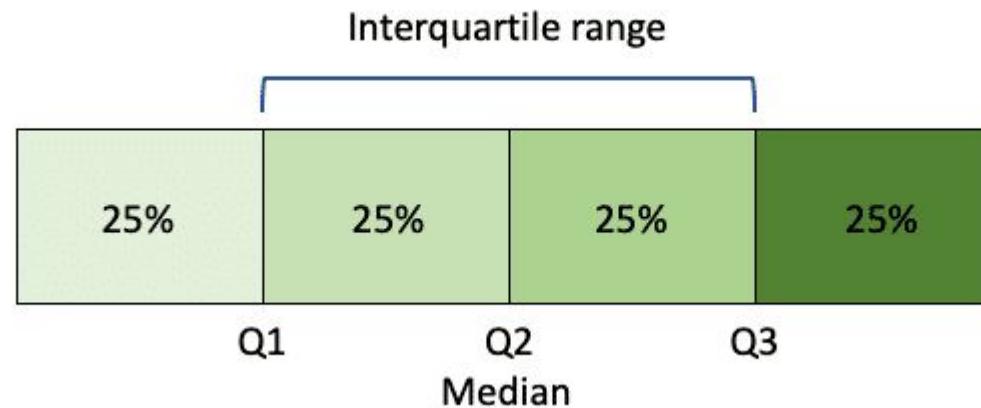
## INTER-QUARTILE RANGE



The spread of the middle half of your distribution

Quartile : each of four equal groups

Quartiles segment any distribution that's ordered from low to high into four equal parts



01

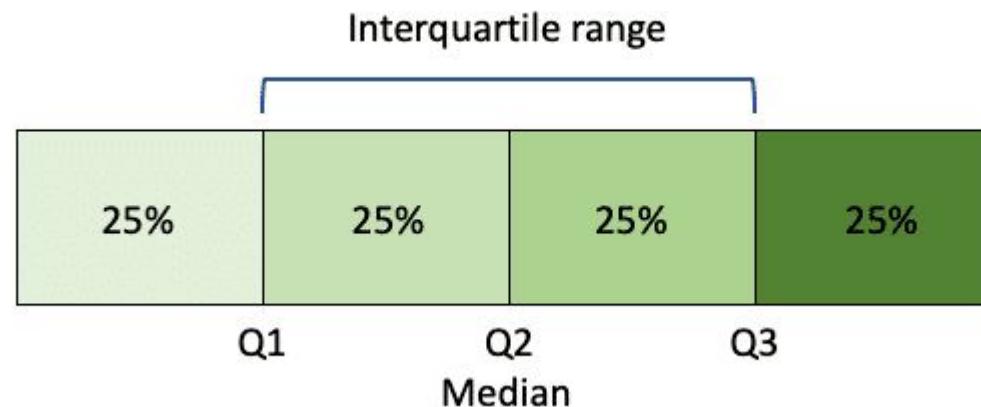
## INTER-QUARTILE RANGE



The spread of the middle half of your distribution

Quartile : each of four equal groups

Quartiles segment any distribution that's ordered from low to high into four equal parts



01

## INTER-QUARTILE RANGE

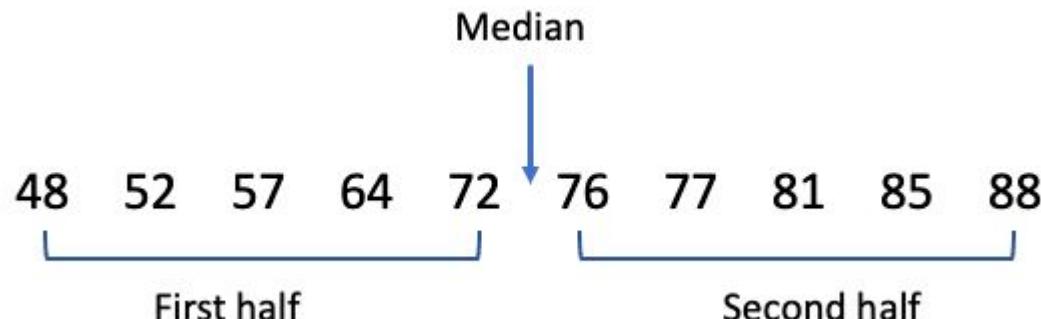


## EVEN NUMBER OF ELEMENTS

Ascending Order of Sequence:

48 52 57 64 72 76 77 81 85 88

Locate the median, and then separate the values below it from the values above it.



# Measures of Dispersion

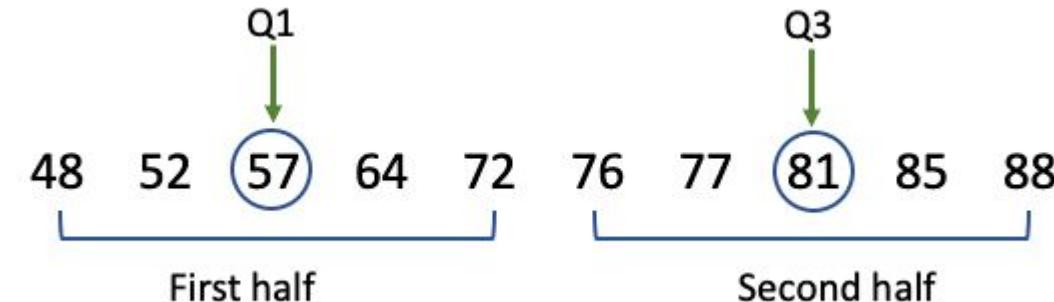
01

## INTER-QUARTILE RANGE



## EVEN NUMBER OF ELEMENTS

Find Q<sub>1</sub> and Q<sub>3</sub>.



IQR

$$\begin{aligned} IQR &= Q_3 - Q_1 \\ IQR &= 81 - 57 = 24 \end{aligned}$$

01

## INTER-QUARTILE RANGE

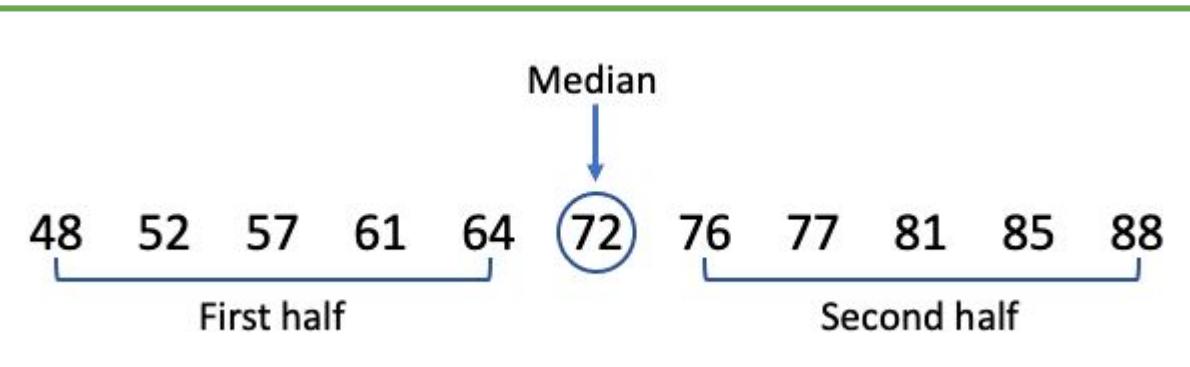


## ODD NUMBER OF ELEMENTS

Ascending Order of Sequence:

48 52 57 61 64 72 76 77 81 85 88

Locate the median, and then separate the values below it from the values above it.



# Measures of Dispersion

01

## INTER-QUARTILE RANGE

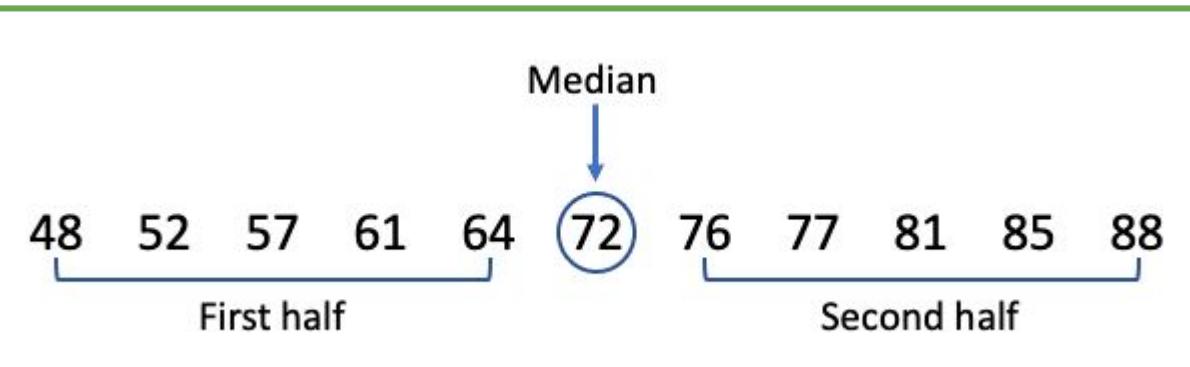


## ODD NUMBER OF ELEMENTS

Ascending Order of Sequence:

48 52 57 61 64 72 76 77 81 85 88

Locate the median, and then separate the values below it from the values above it.



01

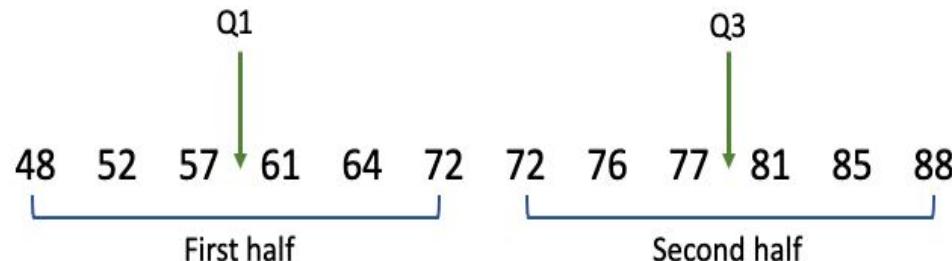
# INTER-QUARTILE RANGE



# EVEN NUMBER OF ELEMENTS

## With Inclusion

## Find Q1 and Q3.



$$Q1 = \frac{57 + 61}{2} = 59$$

$$Q3 = \frac{77 + 81}{2} = 79$$

01

## INTER-QUARTILE RANGE



EVEN NUMBER OF ELEMENTS

With Inclusion

IQR

$$\begin{aligned} IQR &= Q_3 - Q_1 \\ IQR &= 79 - 59 = 20 \end{aligned}$$

# Measures of Dispersion

01

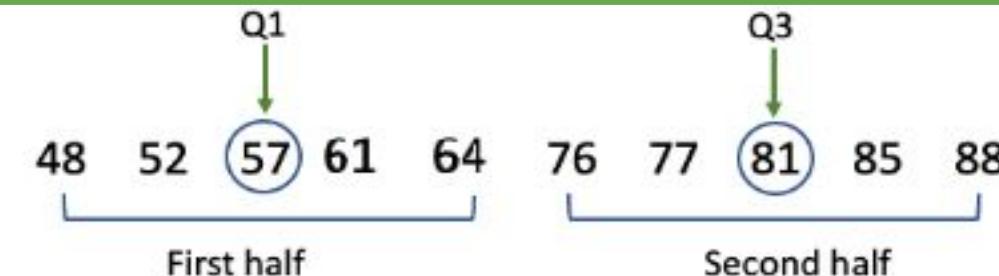
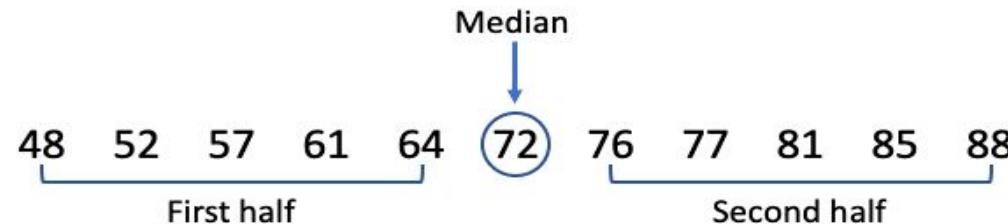
## INTER-QUARTILE RANGE



## EVEN NUMBER OF ELEMENTS

### Exclusive Method

Find Q<sub>1</sub> and Q<sub>3</sub>.



01

## INTER-QUARTILE RANGE



## EVEN NUMBER OF ELEMENTS

### Exclusive Method

IQR

$$\begin{aligned} IQR &= Q_3 - Q_1 \\ IQR &= 81 - 57 = 24 \end{aligned}$$

01

## INTER-QUARTILE RANGE



Useful measure of variability for skewed distributions

IQR can give you an overview of where most of your values lie

Detection of Outlier using IQR

DSBADL Assignment 2

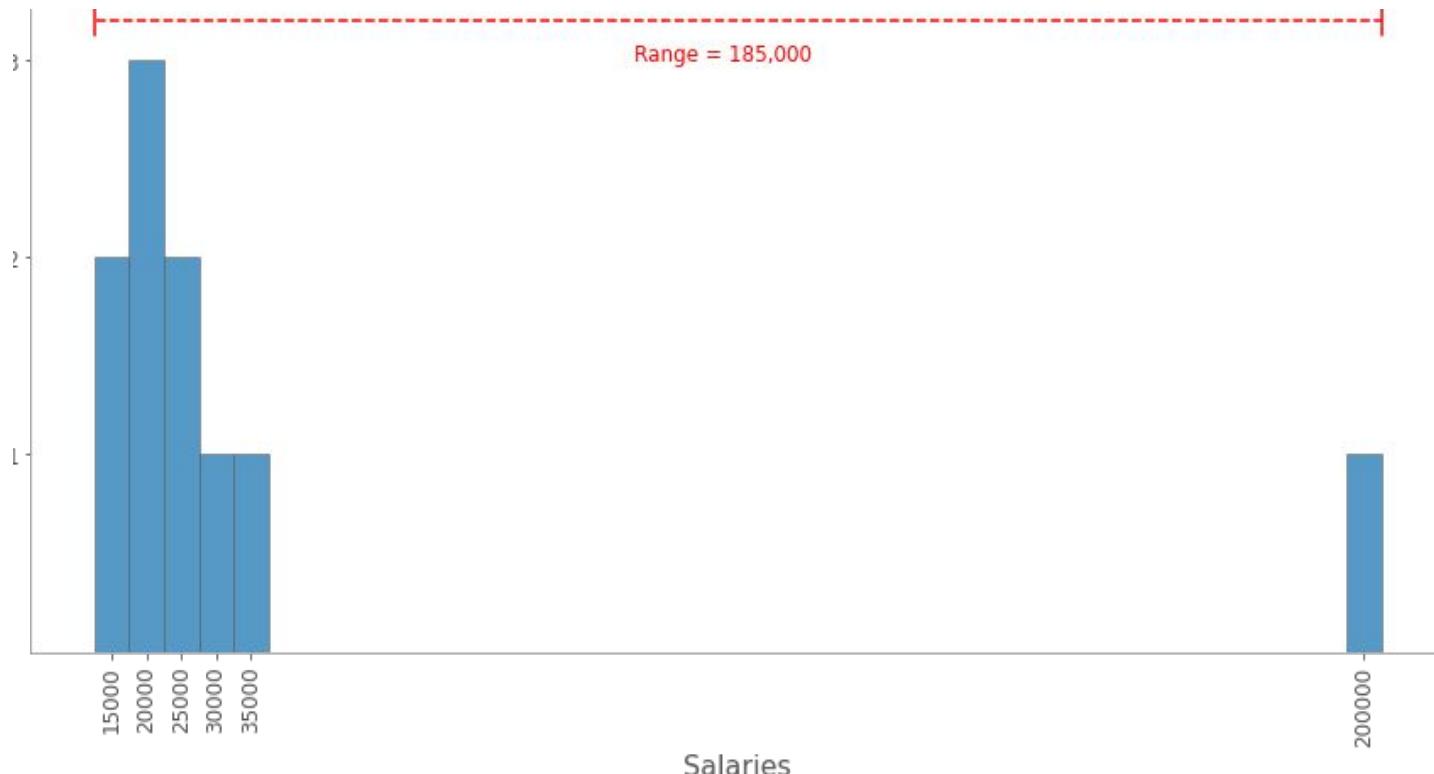
# Measures of Dispersion

\$15,000 | \$15,000 | \$20,000 | \$20,000 | \$20,000 | \$25,000 | \$25,000 | \$30,000 | \$35,000 | \$200,000



# Measures of Dispersion

\$15,000 | \$15,000 | \$20,000 | \$20,000 | \$20,000 | \$25,000 | \$25,000 | \$30,000 | \$35,000 | \$200,000



\$15,000 | \$15,000 | \$20,000 | \$20,000 | \$20,000 | \$25,000 | \$25,000 | \$30,000 | \$35,000 | \$200,000

Most of the values are concentrated around 15,000 and 35,000,

there is an **extreme value (an outlier)** of 200,000 that pushes up the **mean** to 40,500 and  
**dilates the range to 185,000**

02

## VARIANCE



variance is the average of the squared differences from the mean

variance is the average of the squared differences from the mean

Marks of Student A : 30, 50, 70, 100, 100

Mean : 70

Marks of Student B: 70,70,70,70,70

Mean : 70

two data sets are not identical! The variance shows how they are different

02

## VARIANCE



## Formula to Compute Variance

$$\sigma^2 = \frac{\sum(x - \bar{X})^2}{N}$$

# Measures of Dispersion

02

## VARIANCE



|              | Score<br>A | $X - \bar{X}$ | $(X - \bar{X})^2$ |
|--------------|------------|---------------|-------------------|
| 1            | 30         |               |                   |
| 2            | 50         |               |                   |
| 3            | 70         |               |                   |
| 4            | 100        |               |                   |
| 5            | 100        |               |                   |
| <b>Total</b> | <b>350</b> |               |                   |

Mean : 70

02

## VARIANCE



|              | Score<br>X | $X - \bar{X}$ | $(X - \bar{X})^2$ |
|--------------|------------|---------------|-------------------|
| 1            | 30         | $30-70=-40$   |                   |
| 2            | 50         | $50-70=-20$   |                   |
| 3            | 70         | $70-70=0$     |                   |
| 4            | 100        | $100-70=30$   |                   |
| 5            | 100        | $100-70=30$   |                   |
| <b>Total</b> | <b>350</b> |               |                   |

02

## VARIANCE



|              | Score X    | $X - \bar{X}$ | $(X - \bar{X})^2$ |
|--------------|------------|---------------|-------------------|
| 1            | 30         | $30-70=-40$   | 1600              |
| 2            | 50         | $50-70=-20$   | 400               |
| 3            | 70         | $70-70=0$     | 00                |
| 4            | 100        | $100-70=30$   | 900               |
| 5            | 100        | $100-70=30$   | 900               |
| <b>Total</b> | <b>350</b> |               | <b>3800</b>       |

# Measures of Dispersion

02

## VARIANCE



|              | Score<br>X | X - $\bar{X}$ | $(X - \bar{X})^2$ |
|--------------|------------|---------------|-------------------|
| 1            | 30         | 30-70=-40     | 1600              |
| 2            | 50         | 50-70=-20     | 400               |
| 3            | 70         | 70-70=0       | 00                |
| 4            | 100        | 100-70=30     | 900               |
| 5            | 100        | 100-70=30     | 900               |
| <b>Total</b> | <b>350</b> |               | <b>3800</b>       |

$$\sigma^2 = \frac{\sum(x - \bar{X})^2}{N}$$

Variance  
 $= 3800/5$   
**=760**

02

## VARIANCE



|               | Score<br>B | $X - \bar{X}$ | $(X - \bar{X})^2$ |
|---------------|------------|---------------|-------------------|
| 1             | 70         | $70-70=0$     | 0                 |
| 2             | 70         | $70-70=0$     | 0                 |
| 3             | 70         | $70-70=0$     | 0                 |
| 4             | 70         | $70-70=0$     | 0                 |
| 5             | 70         | $70-70=0$     | 0                 |
| <b>Totals</b> | 350        |               | 0                 |

$$\sigma^2 = \frac{\sum(x - \bar{X})^2}{N}$$

Variance  
 $= 0/5$   
 $= 0$

# Measures of Dispersion

02

VARIANCE



| Drive | Mark | Myrna |
|-------|------|-------|
| 1     | 28   | 27    |
| 2     | 22   | 27    |
| 3     | 21   | 28    |
| 4     | 26   | 6     |
| 5     | 18   | 27    |

Which diver was more consistent?

02

## VARIANCE



| Dive          | Mark's Score X | $X - \bar{X}$ | $(X - \bar{X})^2$ |
|---------------|----------------|---------------|-------------------|
| 1             | 28             | 5             | 25                |
| 2             | 22             | -1            | 1                 |
| 3             | 21             | -2            | 4                 |
| 4             | 26             | 3             | 9                 |
| 5             | 18             | -5            | 25                |
| <b>Totals</b> | <b>115</b>     | <b>0</b>      | <b>64</b>         |

# Measures of Dispersion

02

## VARIANCE



| Dive          | Myrna's Score X | $X - \bar{X}$ | $(X - \bar{X})^2$ |
|---------------|-----------------|---------------|-------------------|
| 1             | 27              | 4             | 16                |
| 2             | 27              | 4             | 16                |
| 3             | 28              | 5             | 25                |
| 4             | 06              | -17           | 289               |
| 5             | 27              | 4             | 16                |
| <b>Totals</b> | <b>115</b>      | <b>0</b>      | <b>362</b>        |

02

## VARIANCE



Mark's Variance =  $64 / 5 = 12.8$

Myrna's Variance =  $362 / 5 = 72.4$

Mark has a lower variance therefore he is more consistent.

# Measures of Dispersion

03

## MEAN DEVIATION



- Mean deviation is used to compute how far the values in a data set are from the center point
- Given Instances 5,7,9,3
- $\text{Mean} = (5+7+9+3)/4=6$

$$\begin{aligned}\text{Mean Deviation} &= \frac{(5 - 6) + (7 - 6) + (9 - 6) + (3 - 6)}{4} \\ &= \frac{(-1) + (1) + (3) + (-3)}{4} \Rightarrow 0\end{aligned}$$

$$\begin{aligned}\text{Mean Absolute Deviation} &= \frac{|5 - 6| + |7 - 6| + |9 - 6| + |3 - 6|}{4} \\ &= \frac{(1) + (1) + (3) + (3)}{4} \Rightarrow 2\end{aligned}$$

03

MEAN  
DEVIATION

$$\text{mean absolute deviation} = \frac{\sum |X - \mu|}{N}$$

Where  $\mu$  = mean,  $X$  = score,  $\sum$  = the sum of,  $N$  = number of scores,  $\sum X$  = "add up all the scores",  
 $||$  = take the absolute value (i.e. ignore the minus sign).

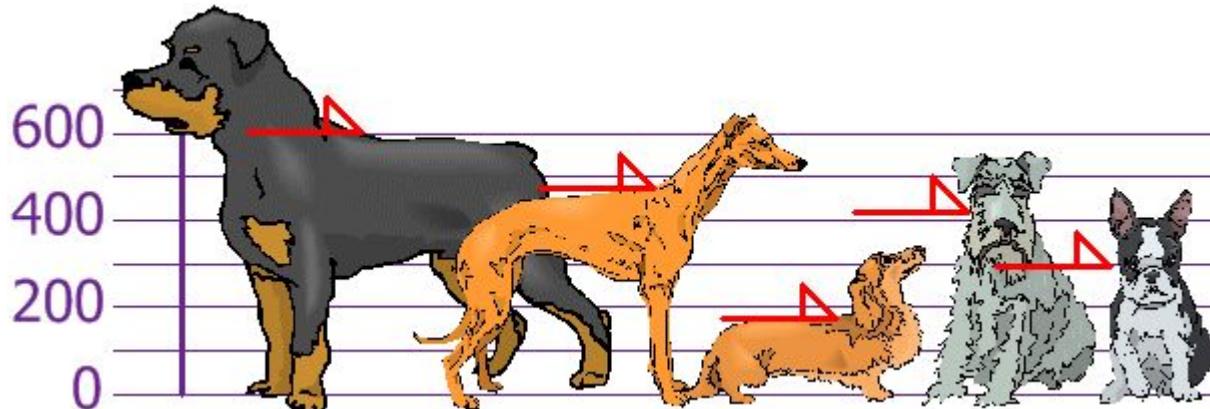
# Measures of Dispersion

03

## MEAN DEVIATION



- You and your friends have just measured the heights of your dogs (in millimeters):
- The heights (at the shoulders) are: 600mm, 470mm, 170mm, 430mm and 300mm.



# Measures of Dispersion

03

## MEAN DEVIATION



Step 1: Find the **mean**:

$$\mu = \frac{600 + 470 + 170 + 430 + 300}{5} = \frac{1970}{5} = 394$$

Step 2: Find the **Absolute Deviations**:

| x                       | x - μ |
|-------------------------|-------|
| 600                     | 206   |
| 470                     | 76    |
| 170                     | 224   |
| 430                     | 36    |
| 300                     | 94    |
| $\Sigma x - \mu  = 636$ |       |

03

MEAN  
DEVIATION

Step 3. Find the **Mean Deviation**:

$$\text{Mean Deviation} = \frac{\sum |x - \mu|}{N} = \frac{636}{5} = 127.2$$

So, on average, the dogs' heights are **127.2 mm from the mean**.

# Measures of Dispersion

04

## STANDARD DEVIATION



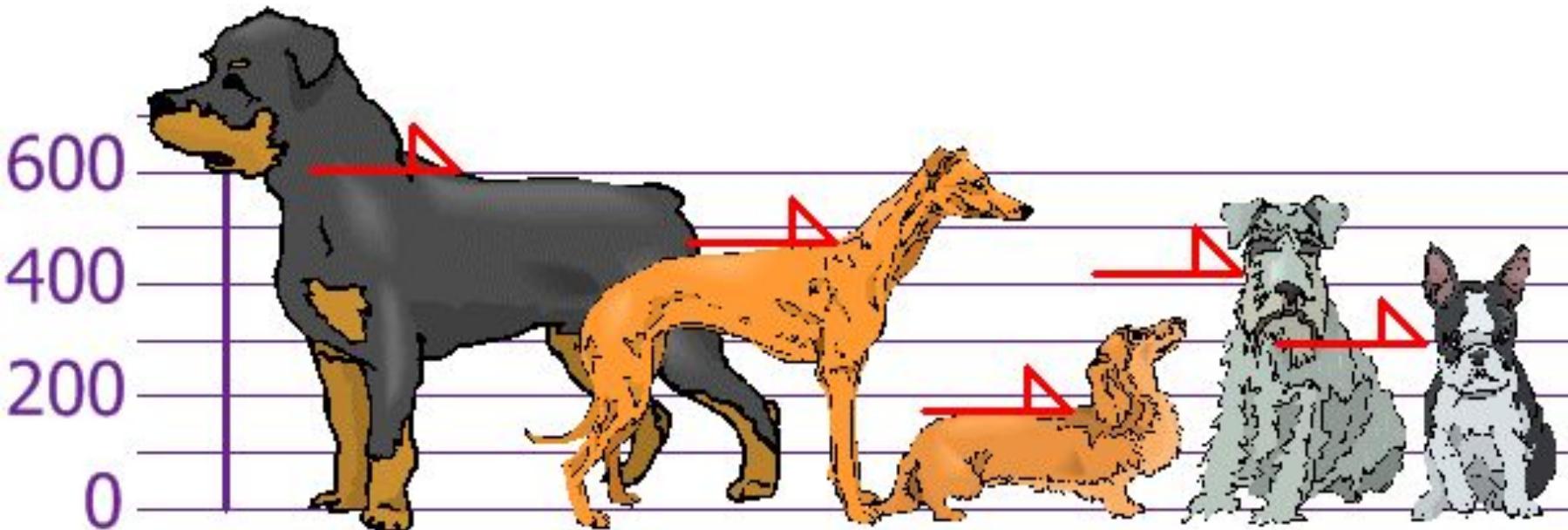
- The Standard Deviation is a measure of how spread out numbers are
- Its symbol is  $\sigma$  (the greek letter sigma)
- It is the square root of the Variance

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$$

# Measures of Dispersion

## VARIANCE AND STANDARD DEVIATION

- You and your friends have just measured the heights of your dogs (in millimeters):
- The heights (at the shoulders) are: 600mm, 470mm, 170mm, 430mm and 300mm.



# Measures of Dispersion

## MEAN, VARIANCE AND STANDARD DEVIATION

- You and your friends have just measured the heights of your dogs (in millimeters):
- The heights (at the shoulders) are: 600mm, 470mm, 170mm, 430mm and 300mm.

$$\text{Mean} = 600 + 470 + 170 + 430 + 300 / 5$$

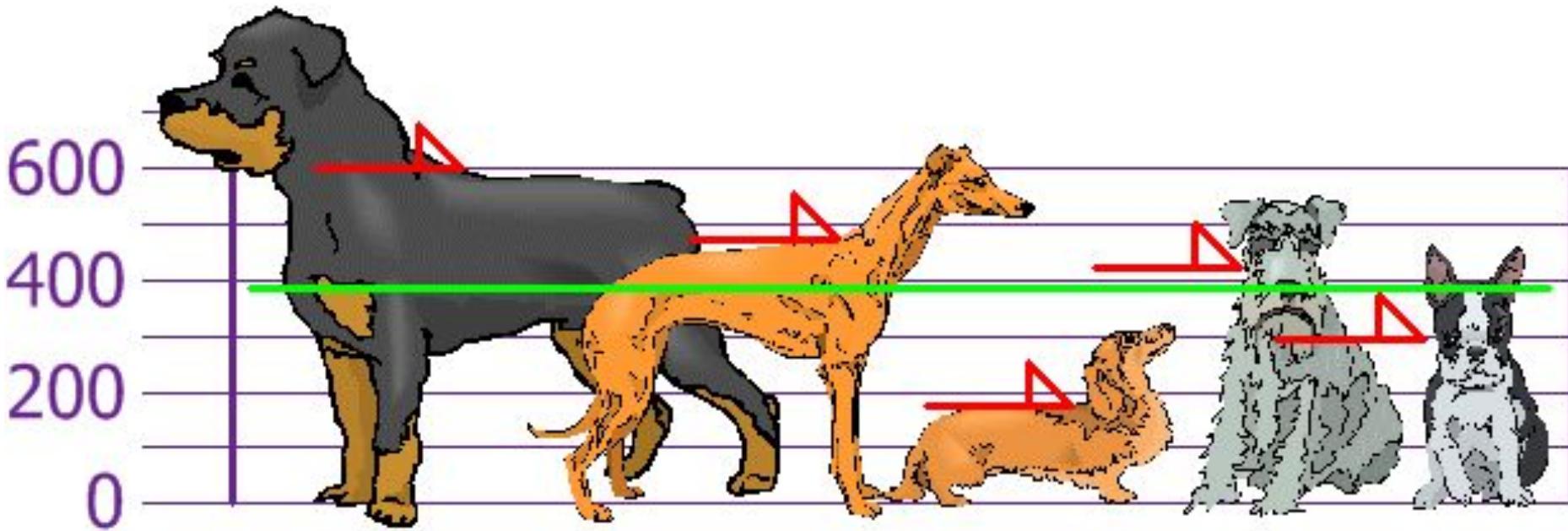
$$= 1970 / 5$$

$$= 394$$

# Measures of Dispersion

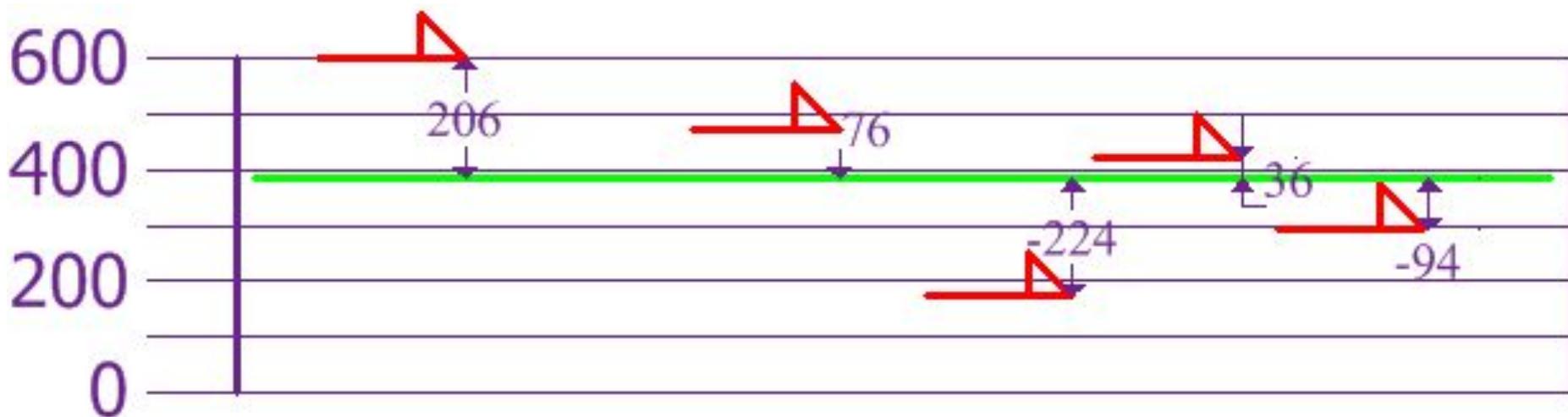
## MEAN, VARIANCE AND STANDARD DEVIATION

The mean (average) height is 394 mm. Let's plot this on the chart:



## MEAN, VARIANCE AND STANDARD DEVIATION

Now we calculate each dog's difference from the Mean(394):



# Measures of Dispersion

## MEAN, VARIANCE AND STANDARD DEVIATION

### Variance

$$\begin{aligned}\sigma^2 &= \frac{206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2}{5} \\&= \frac{42436 + 5776 + 50176 + 1296 + 8836}{5} \\&= \frac{108520}{5} \\&= 21704\end{aligned}$$

So the Variance is **21,704**

## MEAN, VARIANCE AND STANDARD DEVIATION

**Standard Deviation**

$$\sigma = \sqrt{21704}$$

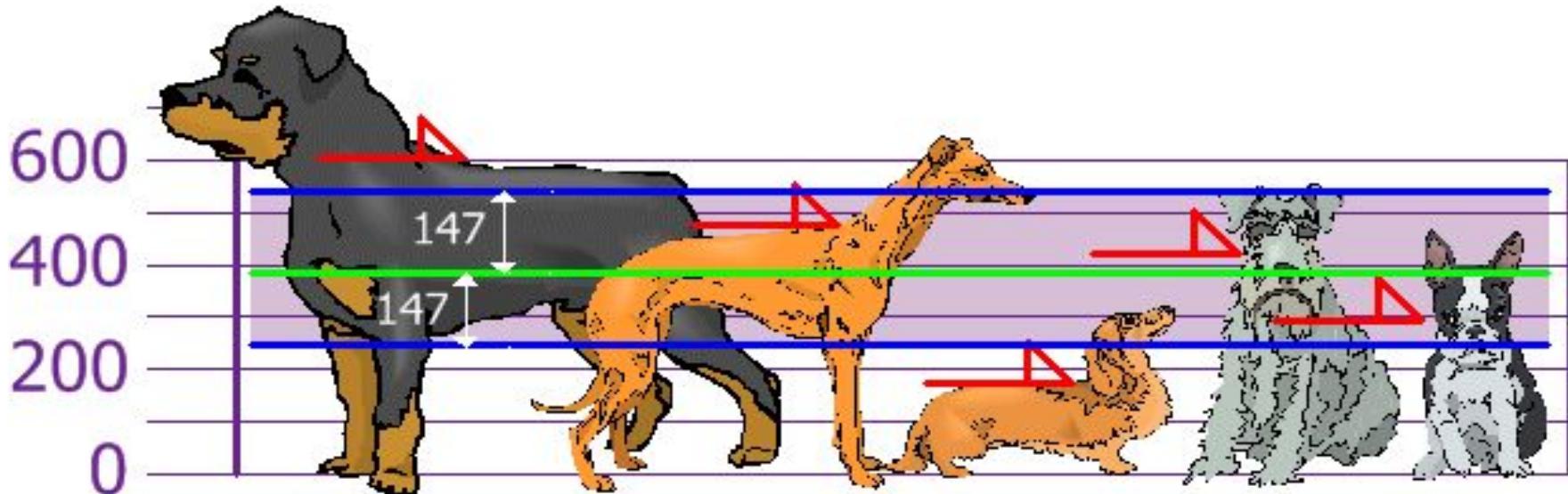
$$= 147.32\dots$$

= **147** (*to the nearest mm*)

# Measures of Dispersion

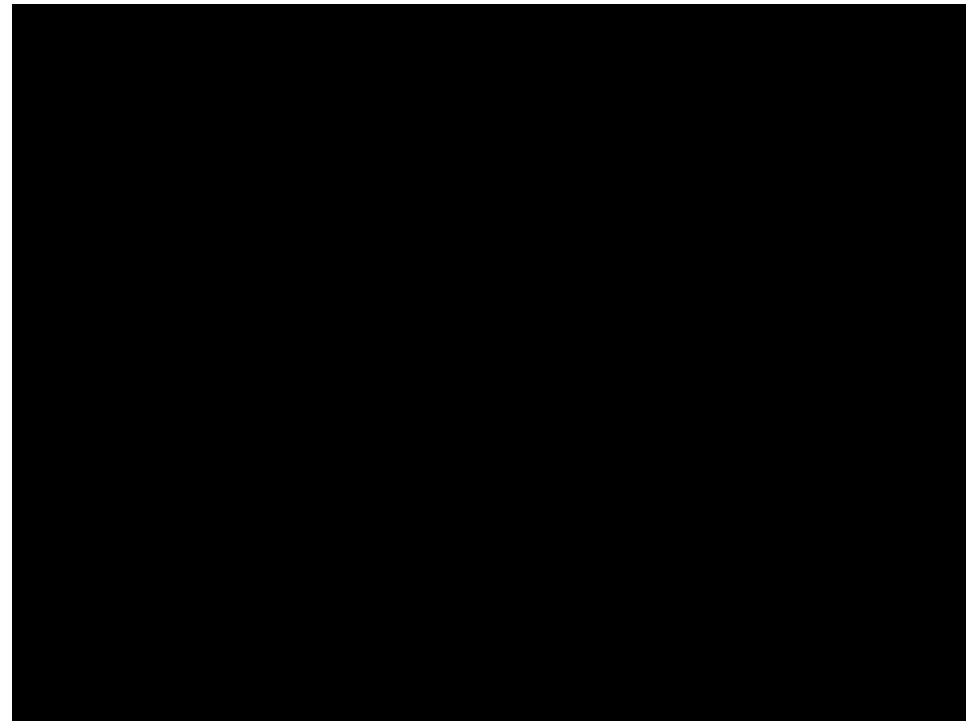
## MEAN, VARIANCE AND STANDARD DEVIATION

we can show which heights are within one Standard Deviation (147mm) of the Mean:



So, using the Standard Deviation we have a "standard" way of knowing what is normal

| Data type | Mathematical operations   | Measures of central tendency   | Measures of variability   |
|-----------|---|--|---|
| Nominal   | <ul style="list-style-type: none"> <li>Equality (<math>=, =</math>)</li> </ul>  | <ul style="list-style-type: none"> <li>Mode</li> </ul>   | <ul style="list-style-type: none"> <li>None</li> </ul>  |
| Ordinal   | <ul style="list-style-type: none"> <li>Equality (<math>=, =</math>)</li> <li>Comparison (<math>&gt;, &lt;</math>)</li> </ul>  | <ul style="list-style-type: none"> <li>Mode</li> <li>Median</li> </ul>   | <ul style="list-style-type: none"> <li>Range</li> <li>Interquartile range</li> </ul>  |
| Interval  | <ul style="list-style-type: none"> <li>Equality (<math>=, =</math>)</li> <li>Comparison (<math>&gt;, &lt;</math>)</li> <li>Addition, subtraction<br/><math>(+, -)</math></li> </ul>   | <ul style="list-style-type: none"> <li>Mode</li> <li>Median</li> <li>Arithmetic mean</li> </ul>                          | <ul style="list-style-type: none"> <li>Range</li> <li>Interquartile range</li> <li>Standard deviation</li> <li>Variance</li> </ul>  |
| Ratio     | <ul style="list-style-type: none"> <li>Equality (<math>=, =</math>)</li> <li>Comparison (<math>&gt;, &lt;</math>)</li> <li>Addition, subtraction<br/><math>(+, -)</math></li> <li>Multiplication, division<br/><math>(\times, \div)</math></li> </ul> | <ul style="list-style-type: none"> <li>Mode</li> <li>Median</li> <li>Arithmetic mean</li> <li>*Geometric mean</li> </ul> | <ul style="list-style-type: none"> <li>Range</li> <li>Interquartile range</li> <li>Standard deviation</li> <li>Variance</li> <li>**Relative standard deviation</li> </ul> |



$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

**LIKELIHOOD**  
the probability of "B" being TRUE given that "A" is TRUE

**PRIOR**  
the probability of "A" being TRUE

**POSTERIOR**  
the probability of "A" being TRUE given that "B" is TRUE

The probability of "B" being TRUE

# Bayes theorem



Probability  
of King



## Marginal Probability:

The probability of an event irrespective of the outcomes of other random variables, e.g.  $P(A)$ .

$P(\text{King})$

$$= P(\text{King and Red}) + P(\text{King and Black}) = \frac{2}{52} + \frac{2}{52} = \frac{4}{52}$$

| Type     | Color |       | Total |
|----------|-------|-------|-------|
|          | Red   | Black |       |
| King     | 2     | 2     | 4     |
| Non-King | 24    | 24    | 48    |
| Total    | 26    | 26    | 52    |

# Bayes theorem



Red And  
King



# Bayes theorem

## Join Probability:

Probability of two (or more) simultaneous events, e.g.  $P(A \text{ and } B)$  or  $P(A, B)$

$P(\text{Red and King})$

$$= \frac{\text{number of cards that are red and king}}{\text{total number of cards}} = \frac{2}{52}$$

| Type     | Color |       | Total |
|----------|-------|-------|-------|
|          | Red   | Black |       |
| King     | 2     | 2     | 4     |
| Non-King | 24    | 24    | 48    |
| Total    | 26    | 26    | 52    |

# Bayes theorem



Jack of Hearts  
from Given  
Face Cards



## Conditional Probability:

Conditional Probability: Probability of one (or more) event given the occurrence of another event,  
e.g.  $P(A \text{ given } B)$  or  $P(A | B)$

$$P(A, B) = P(A | B) * P(B)$$

$$P(A | B) = P(A \cap B) / P(B)$$

$$P(\text{Jack of Hearts} | \text{face card}) = \frac{P(\text{Jack of Hearts} \cap \text{face card})}{P(\text{face card})}$$

$$P(\text{Jack of Hearts} | \text{face card}) = \frac{\left(\frac{1}{52}\right)}{\frac{12}{52}}$$

$$P(\text{Jack of Hearts} | \text{face card}) = \frac{1}{52} \times \frac{52}{12} = \frac{1}{12}$$

$$P(\text{Jack of Hearts} | \text{face card}) = \frac{1}{12} \text{ or } 8.33\%$$

# Bayes theorem

A person speak the truth 3/4 times

Claims to have a king

probability that it is actually a king



E be the event that the man reports that king is drawn from the pack of cards

AK be the event that the king is drawn

$P(AK) = \text{probability that king is drawn} = 4/52 = 1/13$

NK be the event that the king is not drawn.

$P(NK) = \text{probability that king is drawn} = 1 - 1/13 = 12/13$

$P(RK|AK)$  = Probability that the man says the truth  
that king is drawn when actually king is drawn

$P(\text{truth}) = 3/4$

$P(RK/NK)$  = Probability that the man lies that king is  
drawn when actually king is drawn

$P(\text{lie}) = 1/4$

The probability that it is actually a king =  $P(\text{Actual King} | \text{Reported King})$   
=  $P(AK|RK)$

$$P(AK|RK) = P(RK|AK) * P(AK) / P(RK)$$

$$\begin{aligned}P(RK) &= P(RK|AK)P(AK) + P(RK|NK)*P(NK) \\&= 3/4 * 1/13 + 1/4 * 12/13 \\&= 3+12 / 4 * 13\end{aligned}$$

$$P(AK|RK) = \frac{\frac{3}{4} * \frac{1}{13}}{\frac{15}{4} * 13} = \frac{3}{15} = \frac{1}{5} = 0.2$$

The probability that it is actually a king is 20%



# Bayes theorem



# Bayes theorem



# Bayes theorem

Correctly identified each one of  
the two colors **80%**

Failed in identifying each one of  
the two colors **20%**



What is the probability that the cab involved in the accident was  
**blue** rather **green**?

# Bayes theorem



Most people answer **80%**, because the witness is **80%** reliable

How could the probability be so low when the witness is 80% reliable?

But the right answer is **12/29**, or about **41%**

**100** cabs in a town

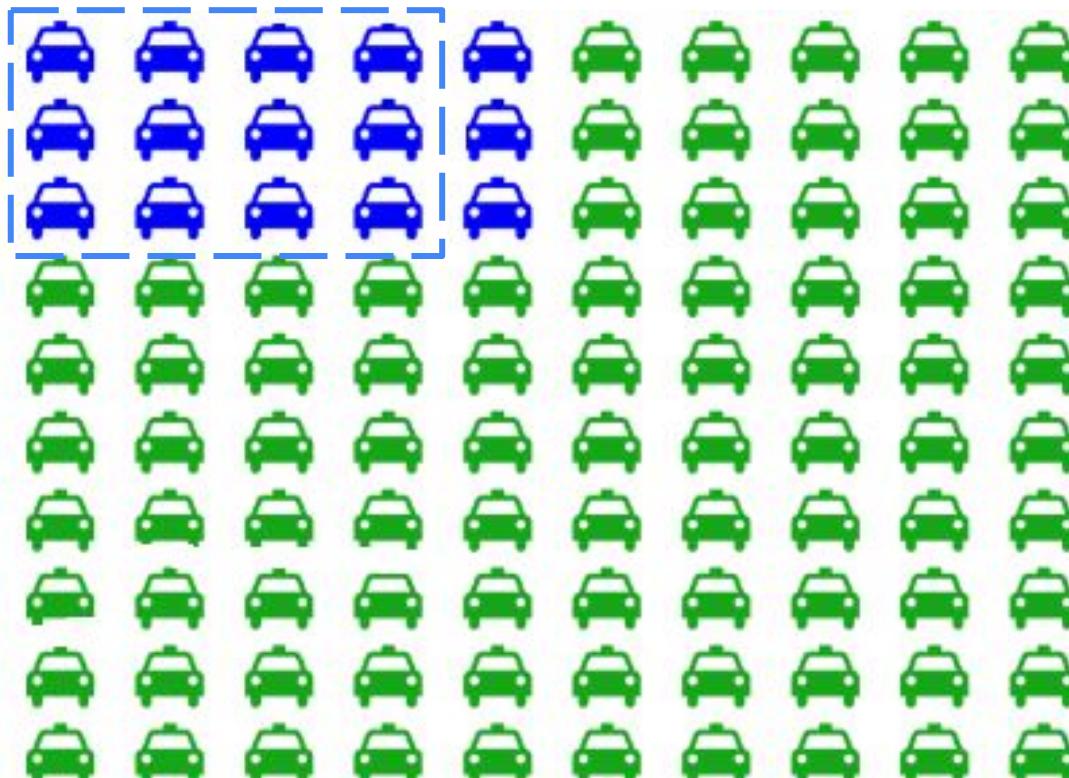
**85**



**15**

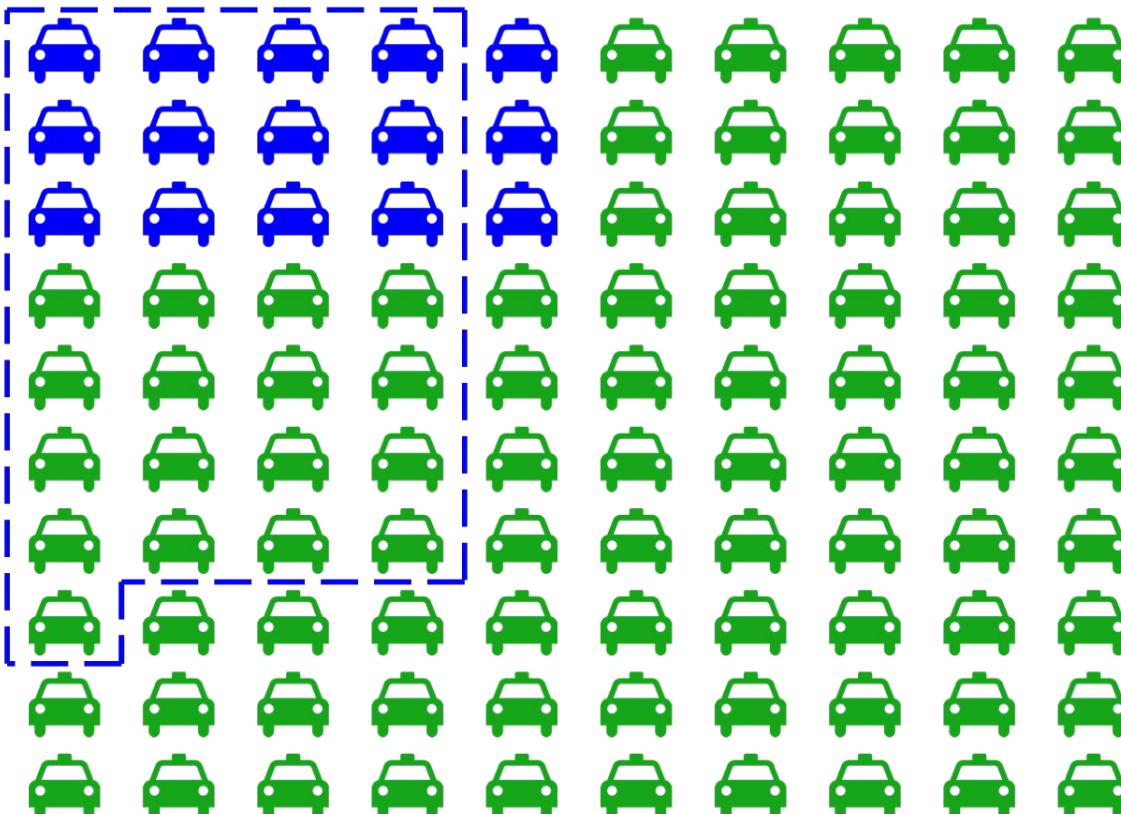


# Bayes theorem



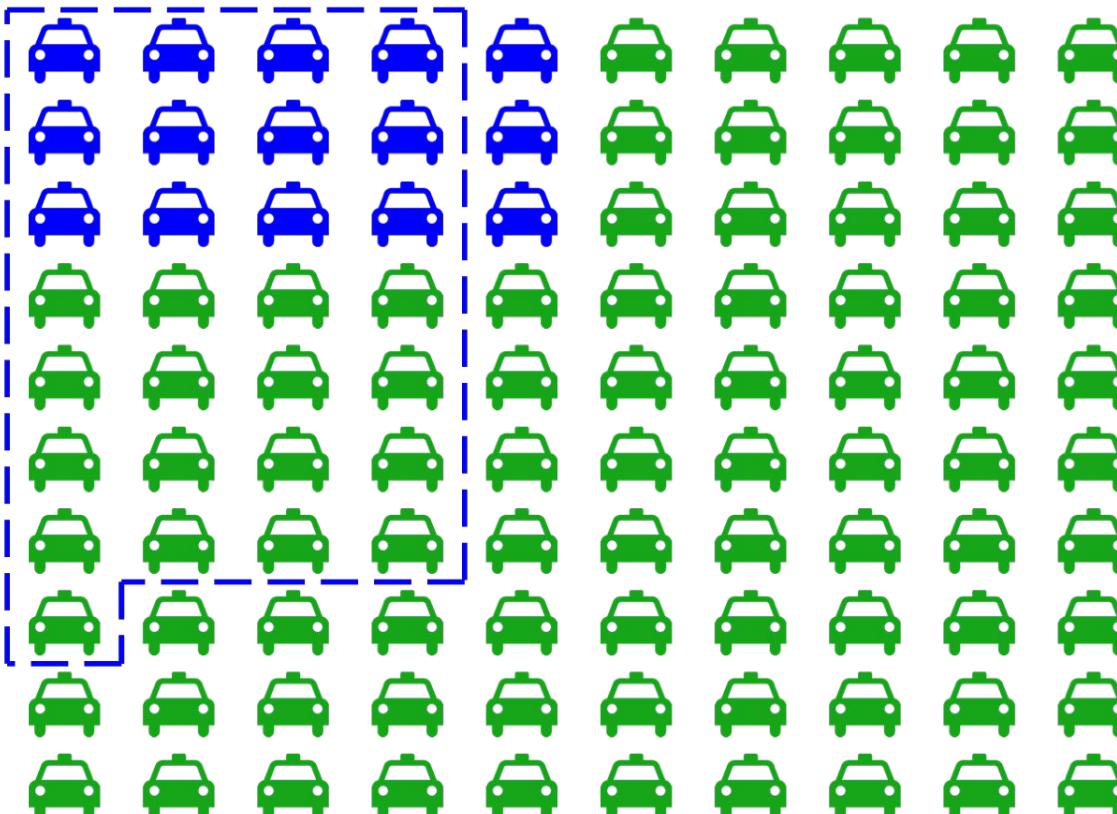
the witness is 80% accurate, that line encompasses 80% of the blue cabs(15), which is 12 cabs.

# Bayes theorem

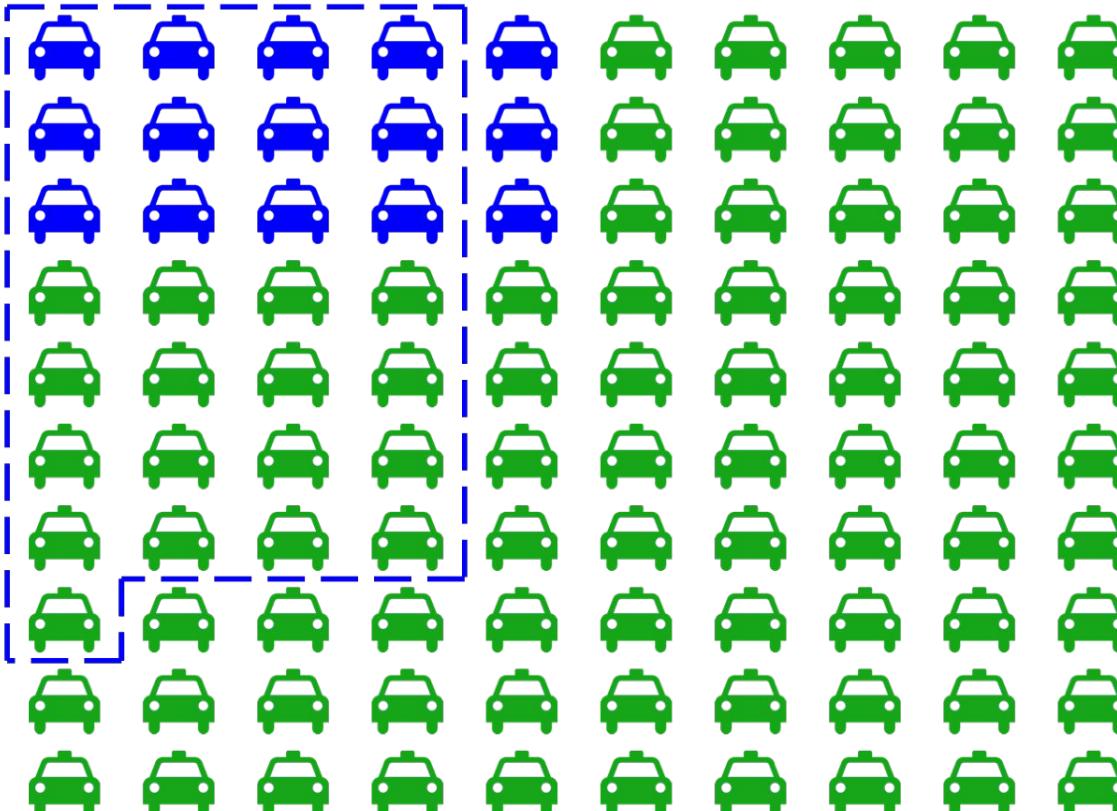


it also encompasses  
20% of the  
green cabs(85),  
which is 17.

# Bayes theorem

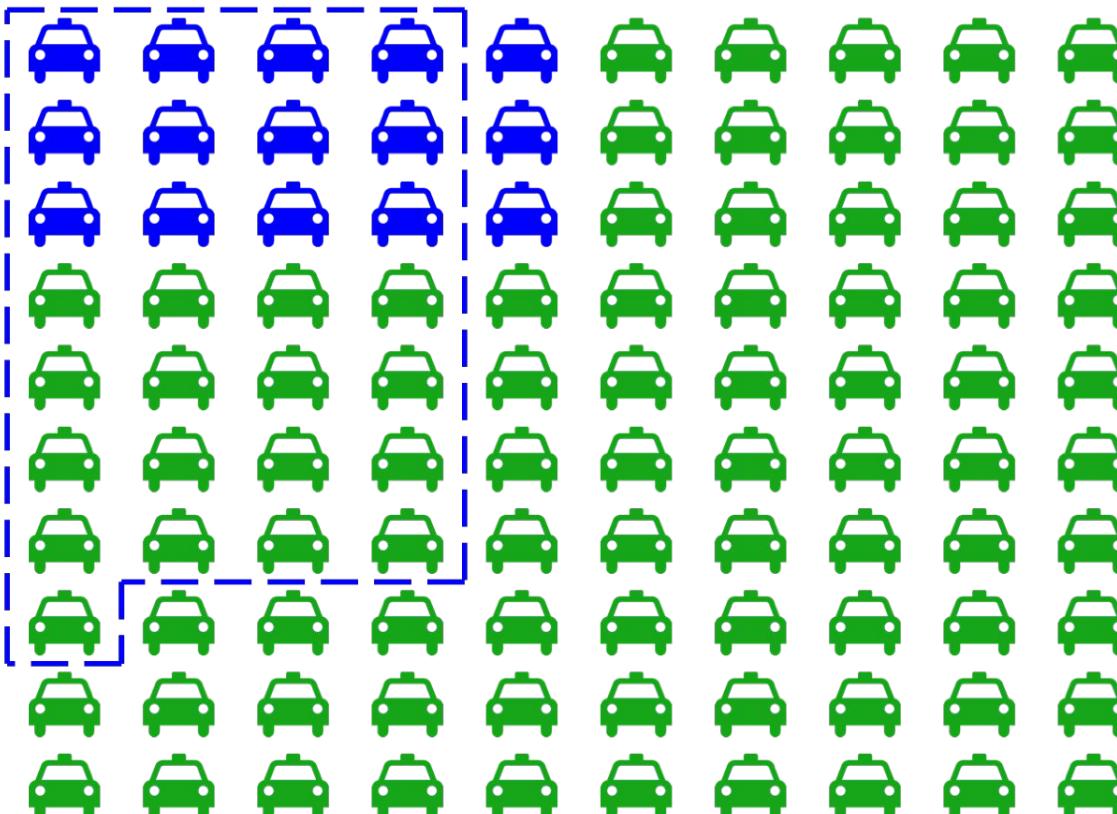


The dashed blue line  
represents the cabs  
the  
**witness identifies as**  
**“blue,”**  
**both right or wrong**



That's a total of  
**29 cabs identified as  
“blue,” only 12 of  
which  
actually are blue.**

# Bayes theorem



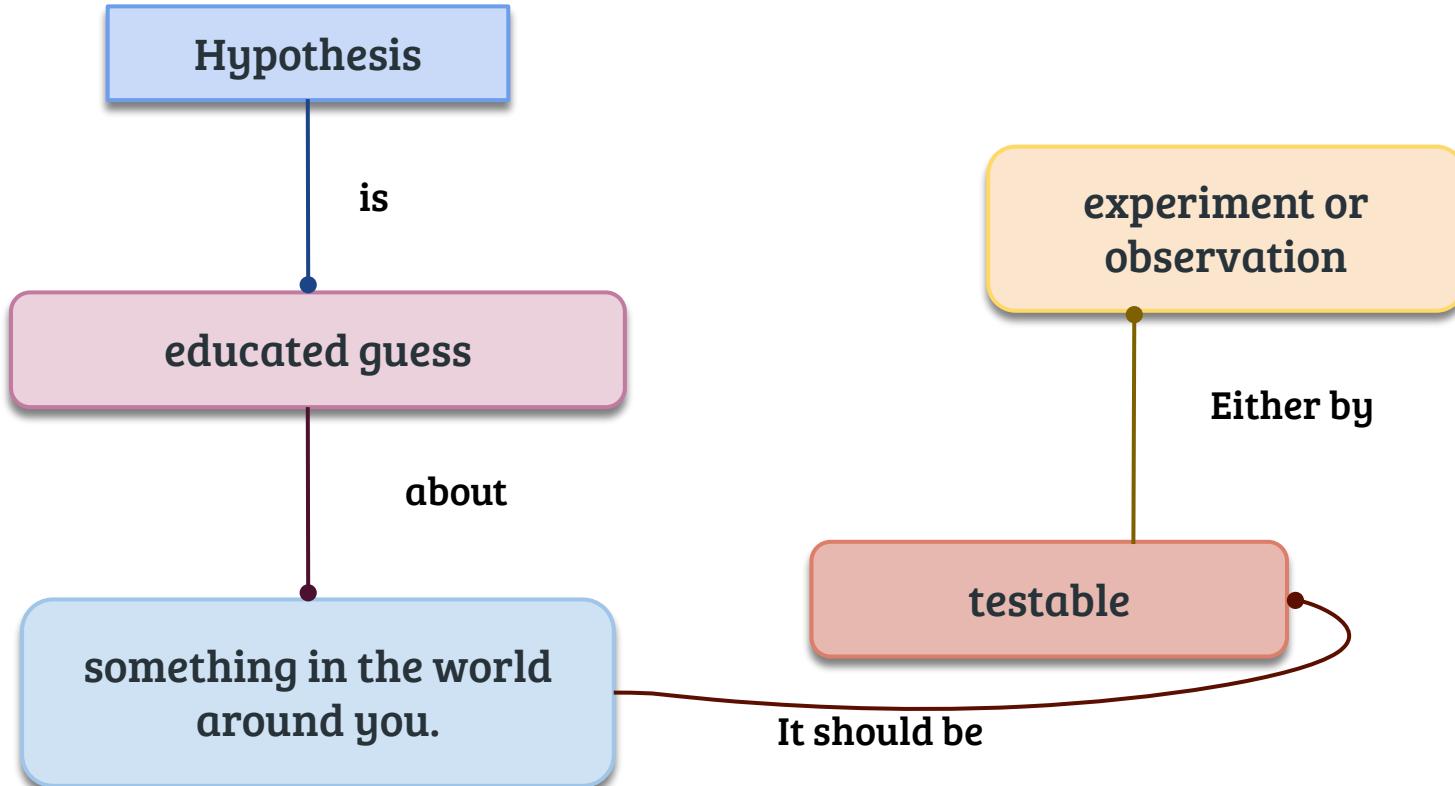
The probability a cab  
really is blue  
given the witness says  
so is  
only **12/29**, about  
**41%**.

- Three companies A, B and C supply 25%, 35% and 40% of the notebooks to a school.
- Past experience shows that 5%, 4% and 2% of the notebooks produced by these companies are defective.
- If a notebook was found to be defective, what is the probability that the notebook was supplied by A?

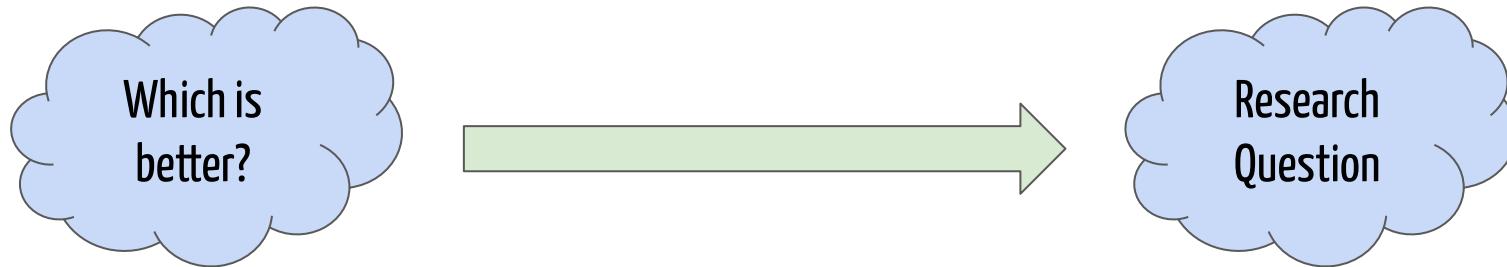
- Let A, B and C be the events that notebooks are provided by A, B and C respectively.
- Let D be the event that notebooks are defective
- Then,
- $P(A) = 0.25, P(B) = 0.35, P(C) = 0.4$
- $P(D|A) = 0.05, P(D|B) = 0.04, P(D|C) = 0.02$
- $$\begin{aligned}P(A | D) &= (P(D | A) * P(A)) / (P(D | A) * P(A) + P(D | B) * P(B) + P(D | C) * P(C)) \\&= (0.05 * 0.25) / ((0.05 * 0.25) + (0.04 * 0.35) + (0.02 * 0.4)) \\&= 2000 / (80 * 69) \\&= \frac{2}{69}\end{aligned}$$

- At a certain university, **4% of men are over 6 feet tall and 1% of women are over 6 feet tall.**
- The total student population is divided in **the ratio 3:2 in favour of women.**
- If a student is selected at random from among **all those over six feet tall, what is the probability that the student is a woman?**

- Let M be the event that student is male and F be the event that the student is female. Let T be the event that student is taller than 6 ft.
- $P(M) = 2/5$   $P(F) = 3/5$   $P(T|M) = 4/100$   $P(T|F) = 1/100$
- $$P(F | T) = (P(T | F) * P(F)) / (P(T | F) * P(F) + P(T | M) * P(M))$$
$$= ((1/100) * (3/5)) / ((1/100) * (3/5) + (4/100) * (2/5))$$
$$= 3/11.$$



# Basic of hypothesis



# Basic of hypothesis

Both  
vaccination  
has same  
efficiency



Hypothesis  
Statement



# Basic of hypothesis

Who can solve  
Sudoku  
Faster?  
Girls/Boys

Research  
Question



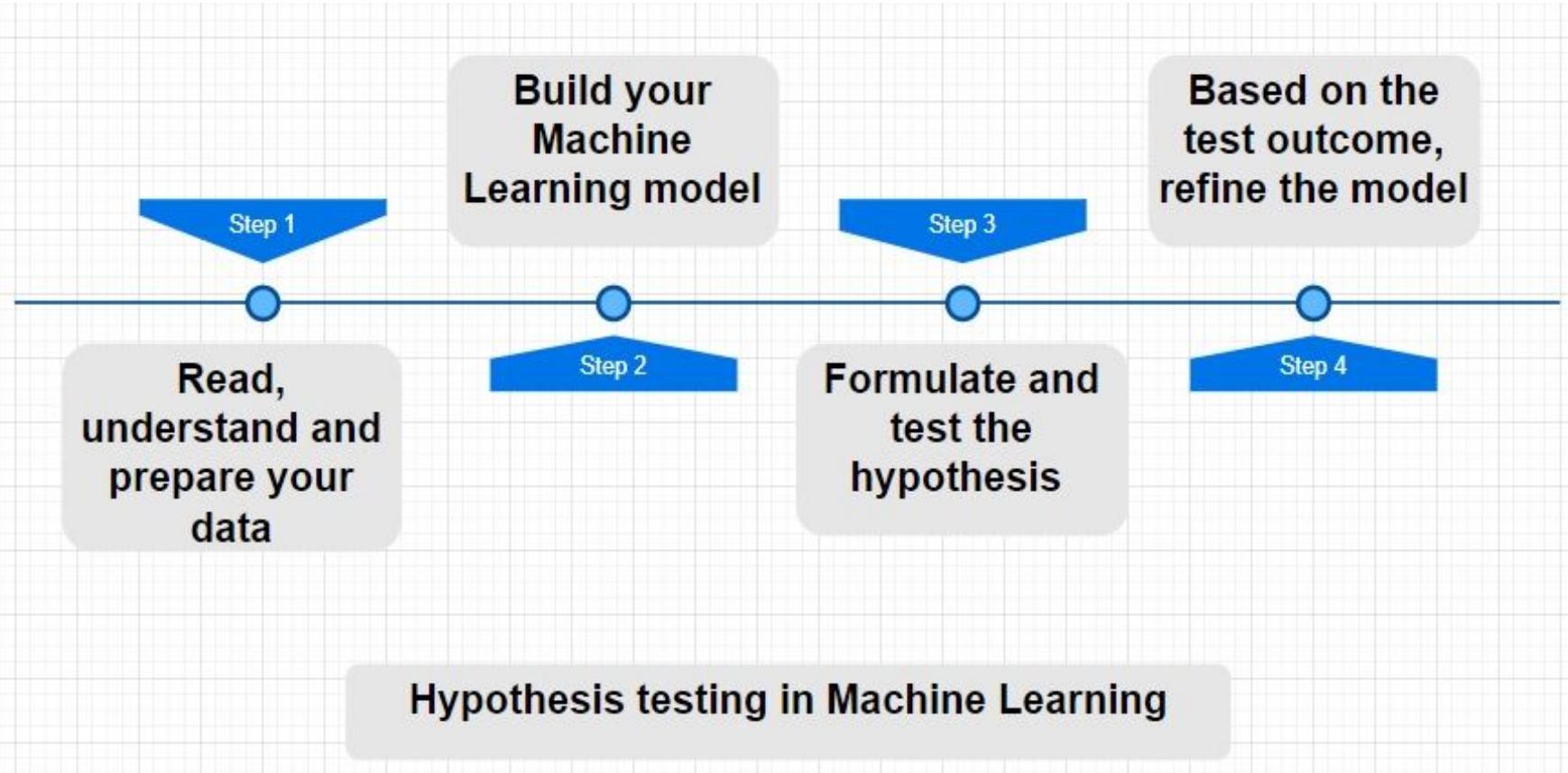
# Basic of hypothesis

The time in seconds  
to solve the SUDOKU  
significantly same for  
Girls and Boys



Hypothesis

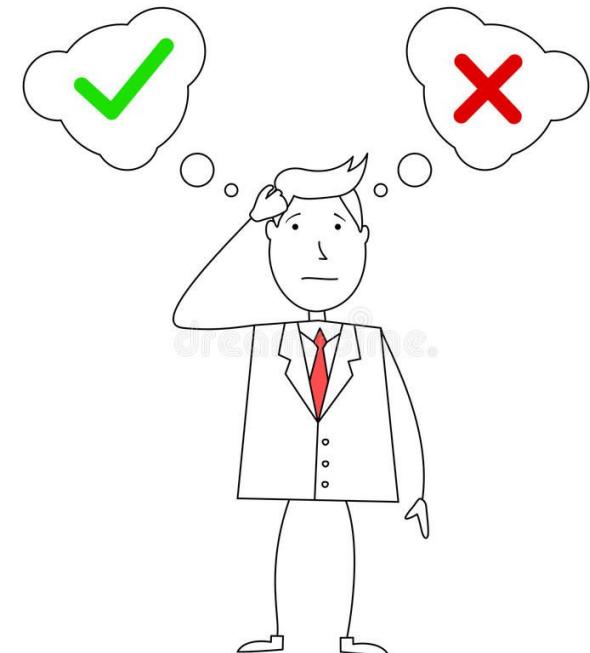
# Need of hypothesis

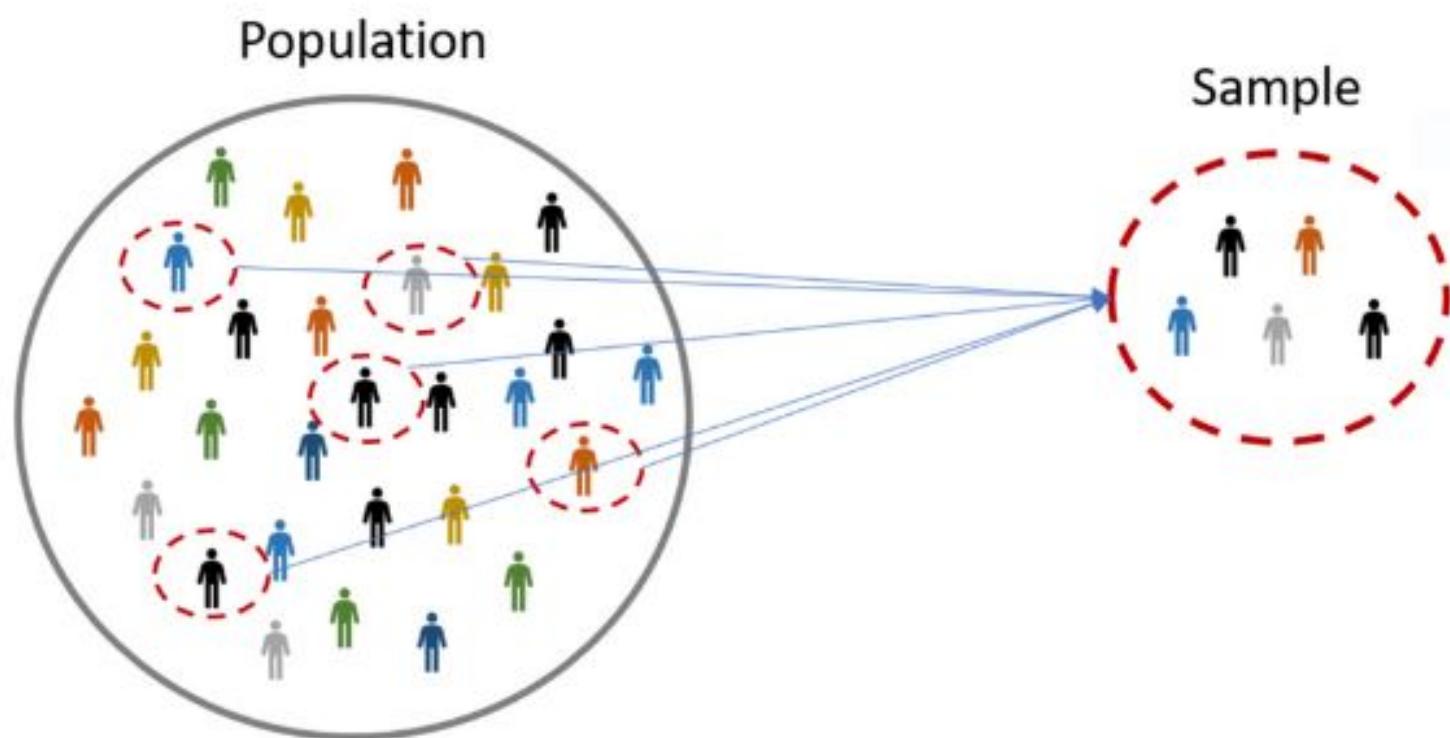


- A statement about the population that **may or may not be true.**
- Hypothesis testing aims to make a statistical conclusion about accepting or not accepting the hypothesis

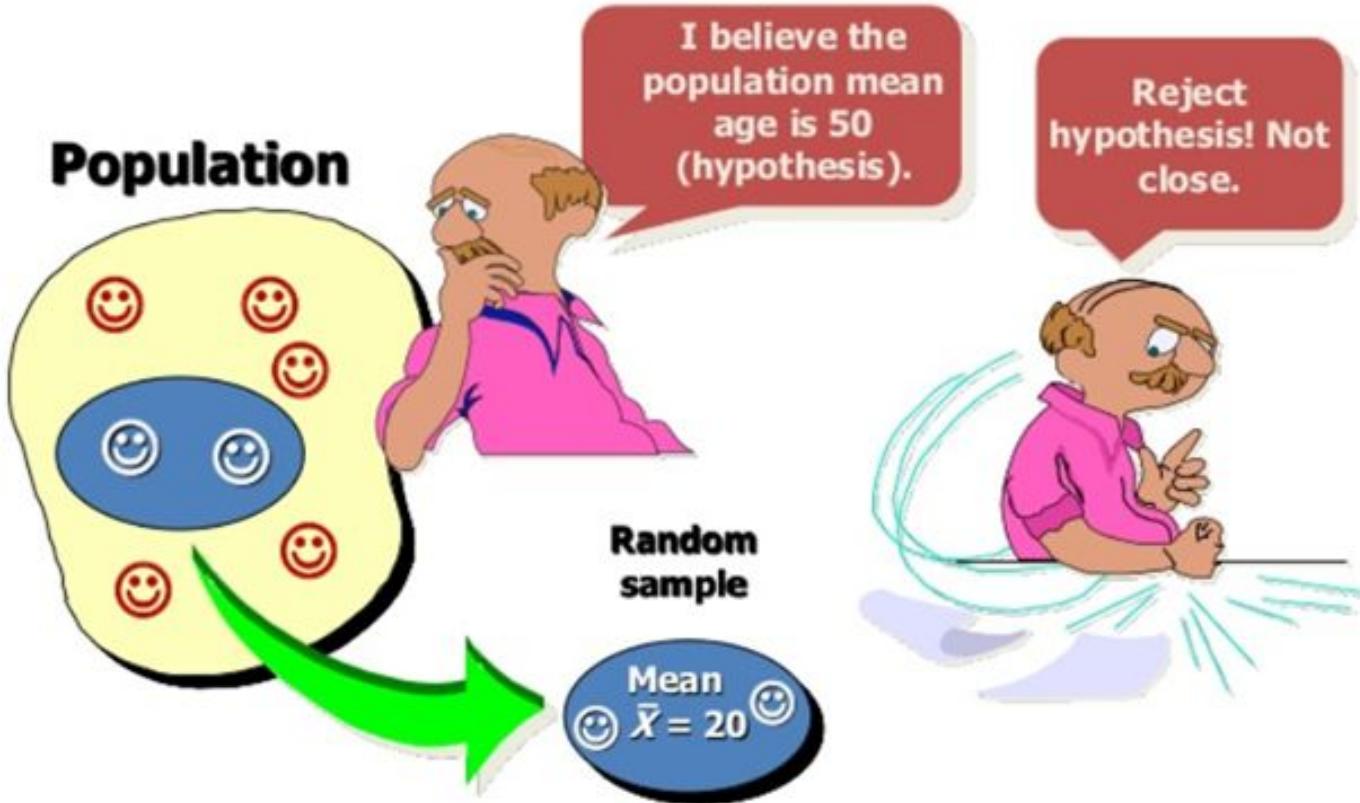


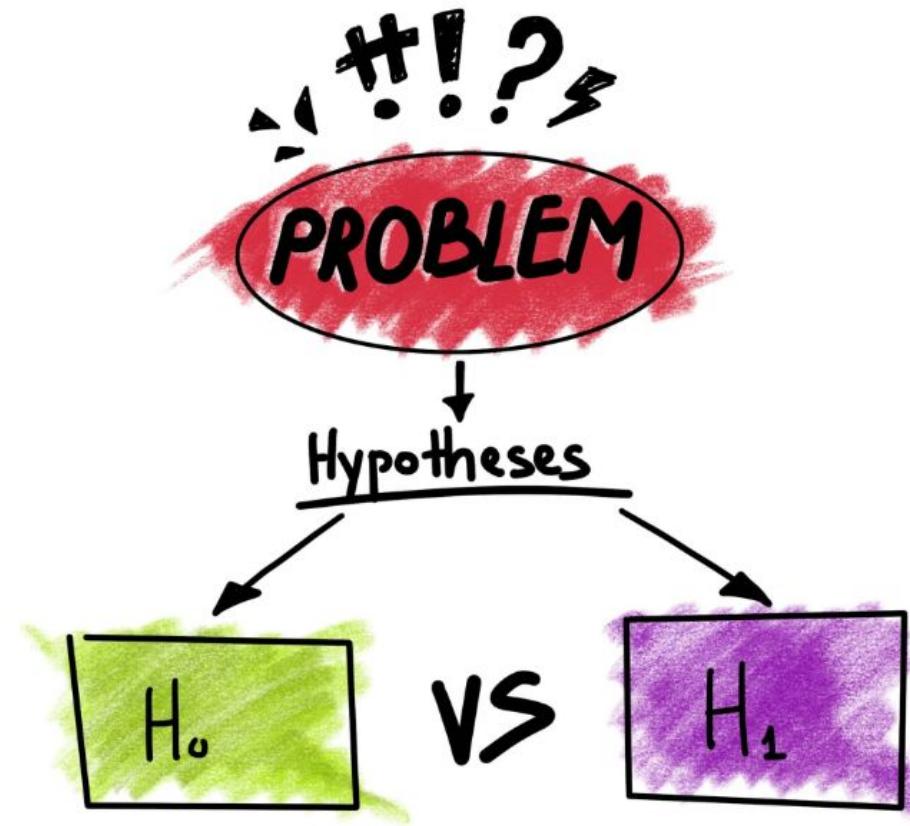
- The best way to determine if a hypothesis was true would be **to examine the entire population**
- Usually **impractical (time, money, resources)**
- Examine random samples from population
- **If sample data are not consistent with hypothesis – reject**

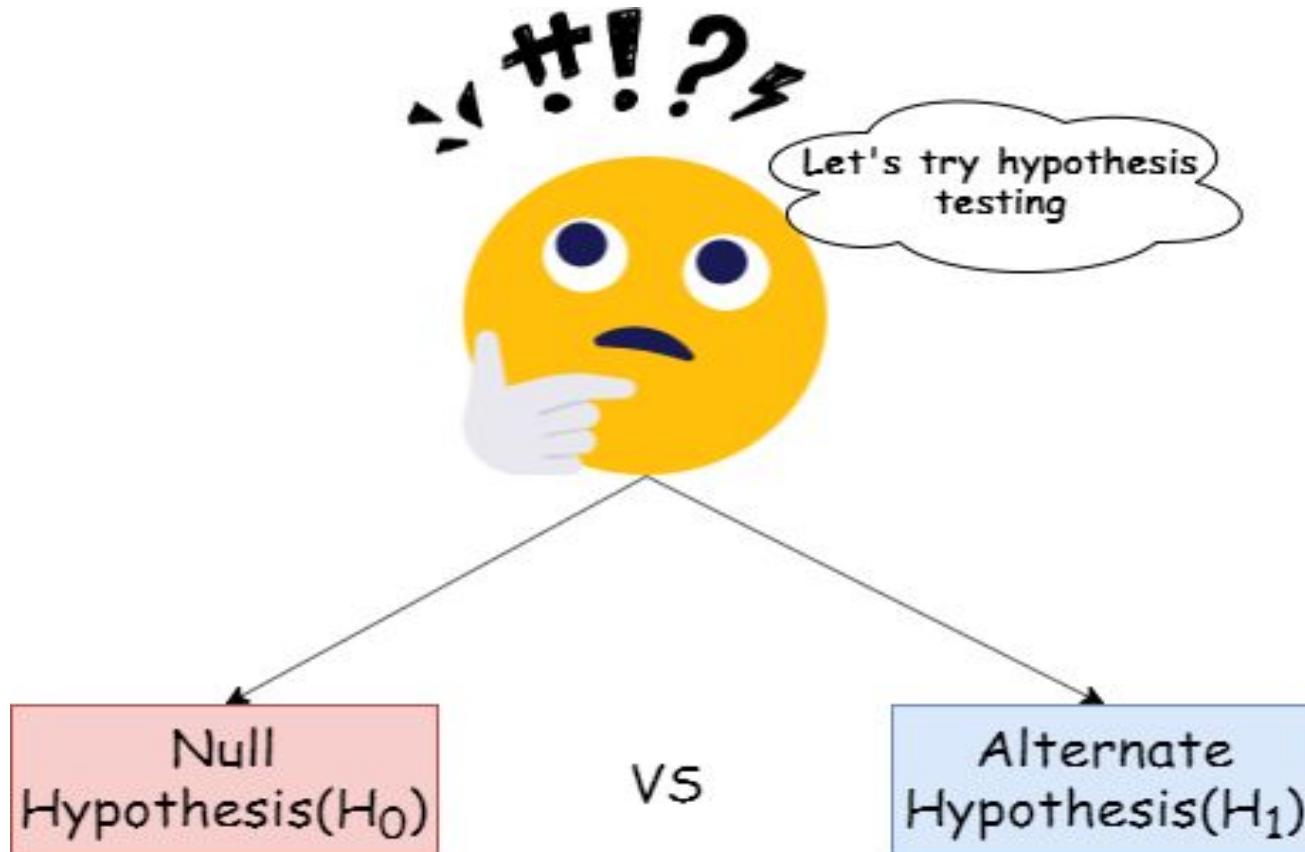




# Hypothesis Testing



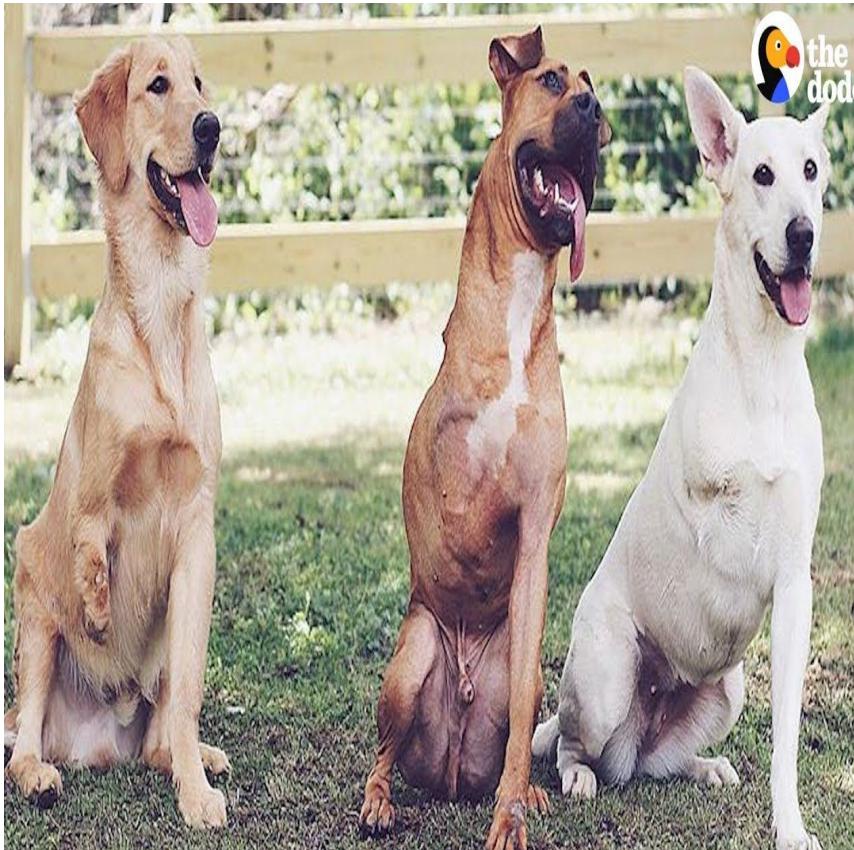






Null  
Hypothesis( $H_0$ )

All dogs have Four  
Lags



Alternate  
Hypothesis( $H_1$ )

5% dogs have Three  
Lags

Student's Claim (Null)

**“I did not cheat”**



Instructor's Claim (Alternative)

**“The student did cheat”**





## Statistical Hypothesis



Null hypothesis  $H_0$

Alternative hypothesis  $H_1$  or  $H_a$





## Statistical Hypothesis



Null hypothesis  $H_0$

- Represents the status quo
- The hypothesis that states there is **no statistical significance between two variables in the hypothesis**
- Believed to be true unless there is overwhelming evidence to the contrary
- It is the hypothesis the researcher is trying to disprove



## Statistical Hypothesis



### Null hypothesis $H_0$

#### Example:

- It is hypothesised that flowers watered with lemonade will grow faster than flowers watered with plain water.

#### Null Hypothesis:

- There is no statistically significant relationship between the type of water used and the growth of the flowers.





## Statistical Hypothesis



## Alternative Hypothesis $H_1$

- Inverse of the null hypothesis
- States that there is a statistical significance between two variables
- Holds true if the null hypothesis is rejected
- Usually what the researcher thinks is true and is testing

## Statistical Hypothesis



## Alternative hypothesis $H_1$

### Null Hypothesis:

If one plant is fed lemonade for one month and another is fed plain water, there will be no difference in growth between the two plants

### Alternative Hypothesis

If one plant is fed lemonade for one month and another is fed plain water, the plant that is fed lemonade will grow more than the plant that is fed plain water



| Null Hypothesis ( $H_o$ )   | Alternate Hypothesis ( $H_a$ )                             |
|---|--|
| Usually describes a status quo, it's a neutral statement, without researcher's study bias | Usually describes a difference, an alternative proposition |
| The one we assume to be true, unless proven otherwise                                     | The one we accept, if we reject the null hypothesis        |
| The one we reject or fail to reject based upon statistical evidence                       | Signs used in Minitab:<br>≠ or < or >                      |
| Signs used in Minitab:<br>= or ≥ or ≤   |  |



Assignment is due for my subject



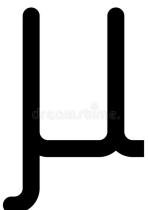
**Hypothesis:**  
an average of 6 days for me to complete the  
assignment.

## Hypothesis Testing :

If the purpose is to test that the population mean is equal to a specific value



gather a sample of people who have completed the assignment in the past



calculate the average number of days it took them to complete it.

hypothesis test states that whether **6.1 days is significantly different from 6.0 days.**

Suppose the sample mean is 6.1 days



## Hypothesis:

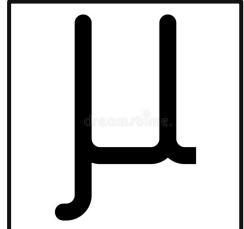
increase the distance that **the ball is hit off the tee by more than 20 meters.**

## Hypothesis Testing :

Improvement over current products, processes or procedures



gather a sample of golfers and



calculate the mean increase in distance hit when using the  
golf balls



## Stating the Null and Alternative Hypothesis



If the purpose is to test that the population mean is equal to a specific value  
(assignment example)

$$H_0 : \mu = 6.0 \text{ days}$$

$$H_1 : \mu \neq 6.0 \text{ days}$$



## Stating the Null and Alternative Hypothesis



Improvement over current products, processes or procedures (golf example)

$$H_0 : \mu \leq 20 \text{ m}$$

$$H_1 : \mu > 20 \text{ m}$$



## Two -Tail Hypothesis Test



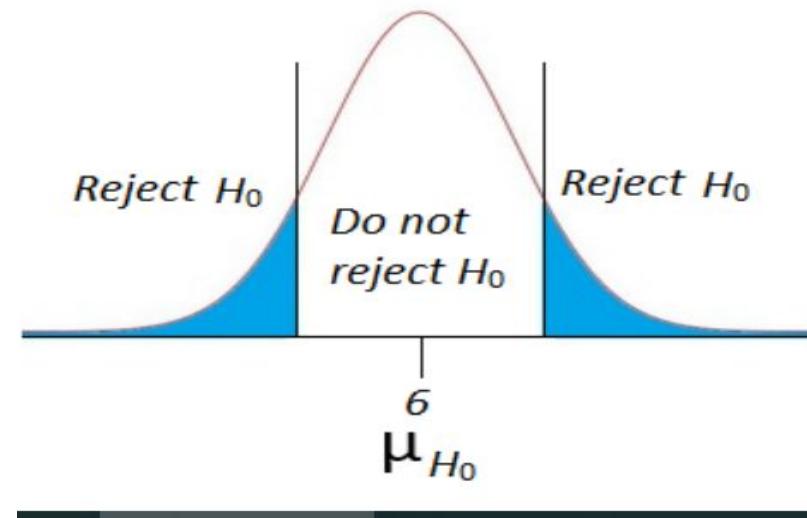
- **Two-tail hypothesis test** is used whenever the alternative **hypothesis is stated as ≠**
- The assignment example would require a two-tail test because the alternative hypothesis is stated as:

$$H_1 : \mu \neq 6.0 \text{ days}$$



## Two -Tail Hypothesis Test

- The curve represents the sampling distribution of the mean for the number of days it takes to complete the assignment



## Two -Tail Hypothesis Test -Procedure



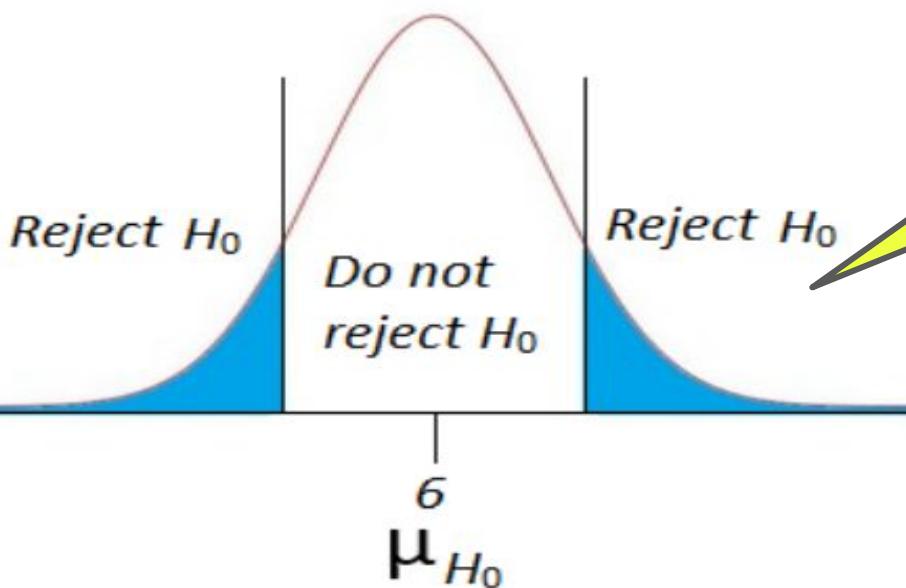
Collect a sample size of  $n$ , and calculate the test statistic – in this case sample mean.

Plot the sample mean on x-axis of the sampling distribution curve

If sample mean falls within white region – we do not reject null hypothesis

If sample mean falls in either shaded region – reject null hypothesis

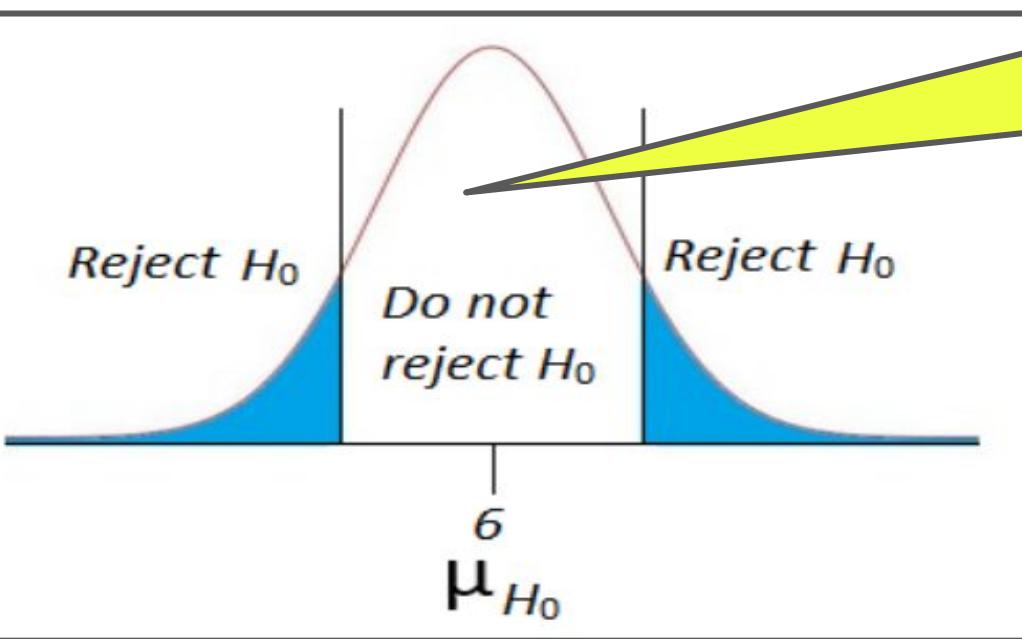
## Two -Tail Hypothesis Test -Procedure



We have enough evidence to support the alternative hypothesis – true population mean is not equal to 6 days

$$H_1 : \mu \neq 6.0 \text{ days}$$

## Two -Tail Hypothesis Test -Procedure



We do not have enough evidence to support the alternative hypothesis – which states that the true population mean is not equal to 6 days

$$H_0 : \mu = 6.0 \text{ days}$$



## Two -Tail Hypothesis Test



There are only two statements we can make about the null Hypothesis:

- **Reject the null hypothesis**
- **Do not reject the null hypothesis**

As conclusions are based on a sample, we do not have enough evidence to ever accept the null hypothesis.

## One Tail Hypothesis Test



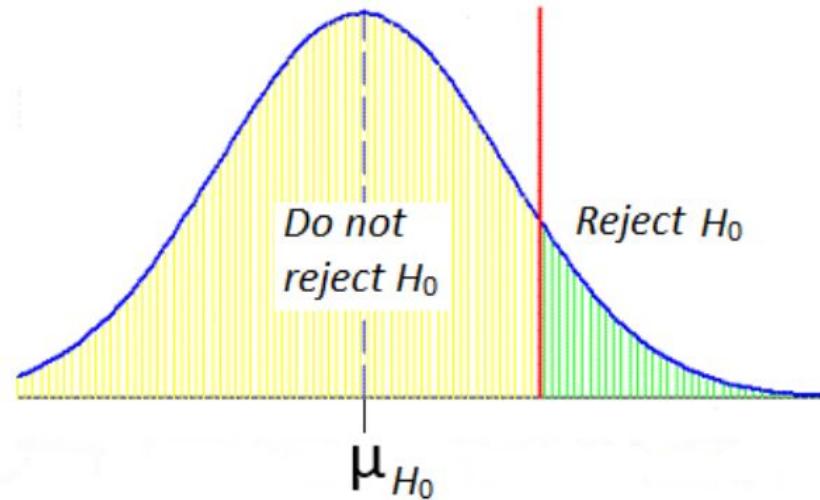
- **One -tail hypothesis test** is used whenever the alternative hypothesis is stated as **< or >**
- The golf example would require a one-tail test because the alternative hypothesis is expressed as:

$$H_1 : \mu > 20 \text{ m}$$



## One Tail Hypothesis Test

Test and plot the sample mean, which represents the average increase in distance from the tee using the golf ball



## One -Tail Hypothesis Test -Procedure



Collect a sample size of  $n$ , and calculate the test statistic – in this case sample mean.

Plot the sample mean on x-axis of the sampling distribution curve

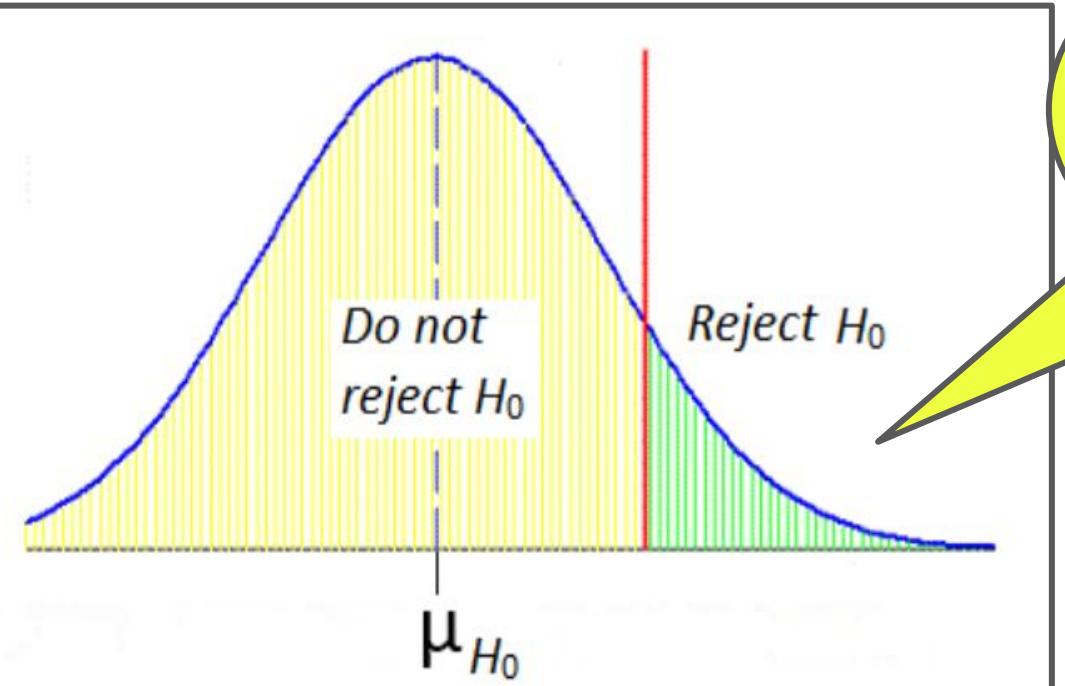
If sample mean falls within white region – we do not reject null hypothesis

If sample mean falls in either shaded region – reject null hypothesis

# Hypothesis Testing

| BASIS OF COMPARISON            | ONE-TAILED TEST  | TWO-TAILED TEST  |
|--------------------------------|--|--|
| Meaning                        | A statistical hypothesis test in which alternative hypothesis has only one end, is known as one tailed test. | A significance test in which alternative hypothesis has two ends, is called two-tailed test. |
| Hypothesis                     | Directional  | Non-directional  |
| Region of rejection            | Either left or right   | Both left and right  |
| Determines                     | If there is a relationship between variables in single direction.  | If there is a relationship between variables in either direction.                            |
| Result                         | Greater or less than certain value.  | Greater or less than certain range of values.  |
| Sign in alternative hypothesis | > or <   | =  |

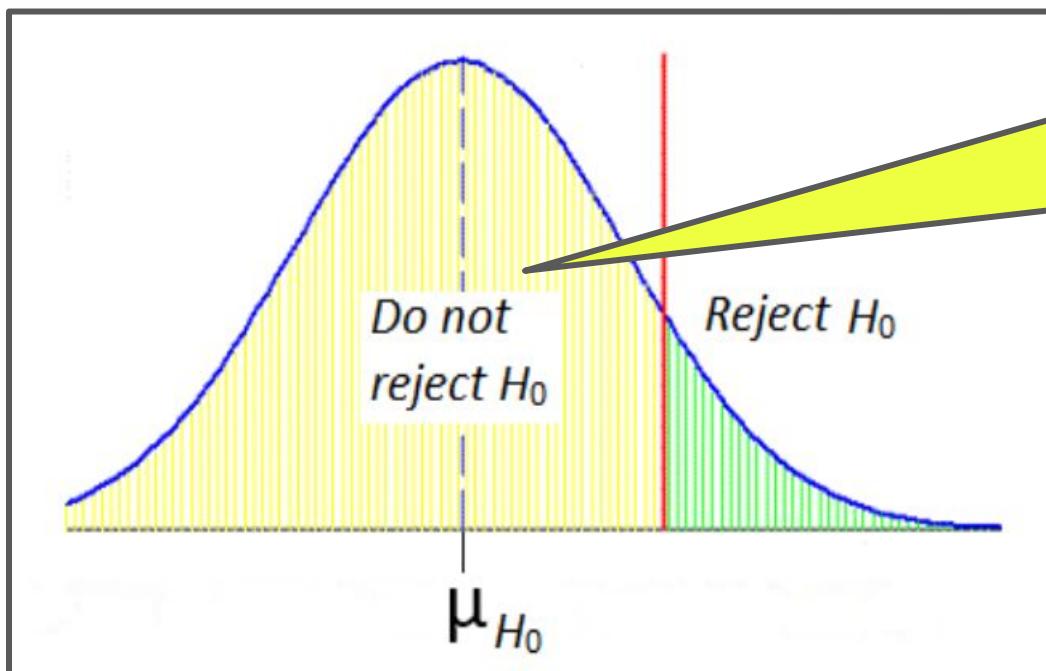
## One-Tail Hypothesis Test -Procedure



We have enough evidence to support the alternative hypothesis – golf ball will increase distance off the tee by more than 20 m

$$H_1 : \mu > 20 \text{ m}$$

## One-Tail Hypothesis Test -Procedure



We do not have enough evidence to support the alternative hypothesis – which states that the golf ball increased distance off the tee by more than 20 m

$$H_0 : \mu \leq 20 \text{ m}$$

## Type I & Type II Errors in Hypothesis Testing: Example



|         |                           | Measured                           |                                    |
|---------|---------------------------|------------------------------------|------------------------------------|
|         |                           | New version is NOT better          | New version is better              |
| Reality | New version is NOT better | Correct decision 😊                 | Type I False Positive ( $\alpha$ ) |
|         | New version is better     | Type II False Negative ( $\beta$ ) | Correct decision 😊                 |

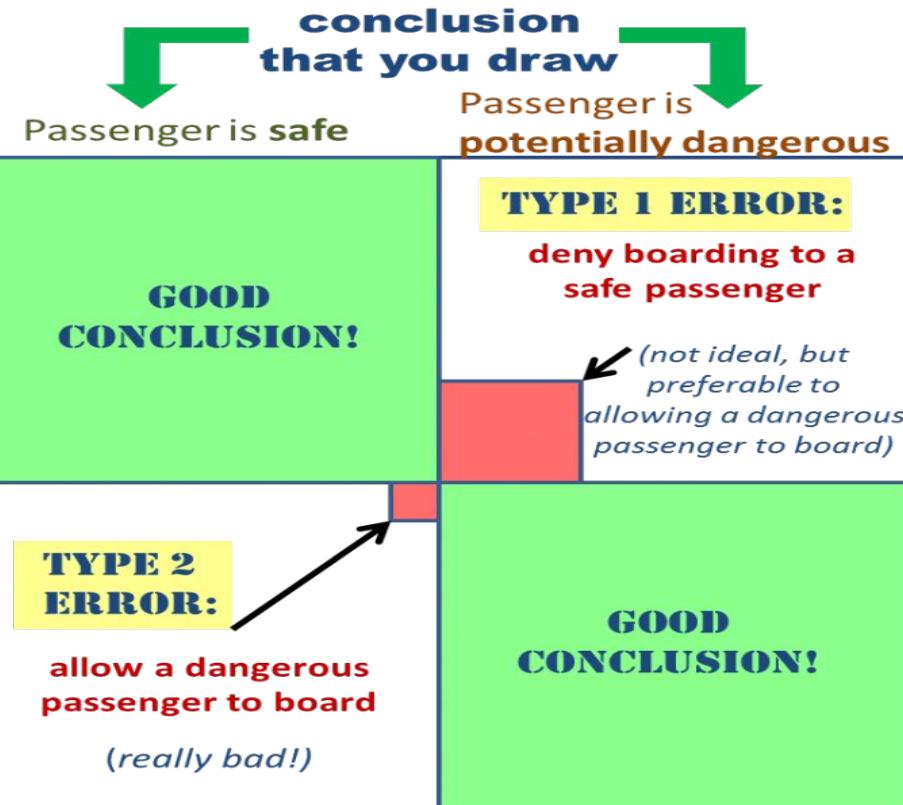
## Type I & Type II Errors in Hypothesis Testing



Passenger is innocent, meaning he is not armed

↑  
**the true state of nature**

↓  
Passenger is not innocent, meaning he is armed



## Type I & Type II Errors in Hypothesis Testing



|                                   | Null hypothesis<br>is TRUE          | Null hypothesis<br>is FALSE         |
|-----------------------------------|-------------------------------------|-------------------------------------|
| Reject null<br>hypothesis         | Type I Error<br>(False positive)    | Correct outcome!<br>(True positive) |
| Fail to reject<br>null hypothesis | Correct outcome!<br>(True negative) | Type II Error<br>(False negative)   |

## Type I & Type II Errors in Hypothesis Testing



### Type I and Type II Error

| Null hypothesis is ... | True  | False  |
|------------------------|---|--|
| Rejected               | Type I error<br>False positive<br>Probability = $\alpha$        | Correct decision<br>True positive<br>Probability = $1 - \beta$ |
| Not rejected           | Correct decision<br>True negative<br>Probability = $1 - \alpha$ | Type II error<br>False negative<br>Probability = $\beta$       |



## Type I Errors in Hypothesis Testing - Alpha Error



This is when you **reject a true Null Hypothesis**. Meaning you find **something significant when it's really not significant..**

The probability of making a **Type I Error** is determined in the decision making process because it is **the level of significance (or alpha level)**

When you make a **correct rejection by rejecting a false Null Hypothesis**,  $p = (1-\beta)$  which is the **probability (p) of being correct (also known as Power)**.

It is only possible to make a **Type I Error if the Null Hypothesis is rejected.**

## Type II Errors in Hypothesis Testing - Beta Error



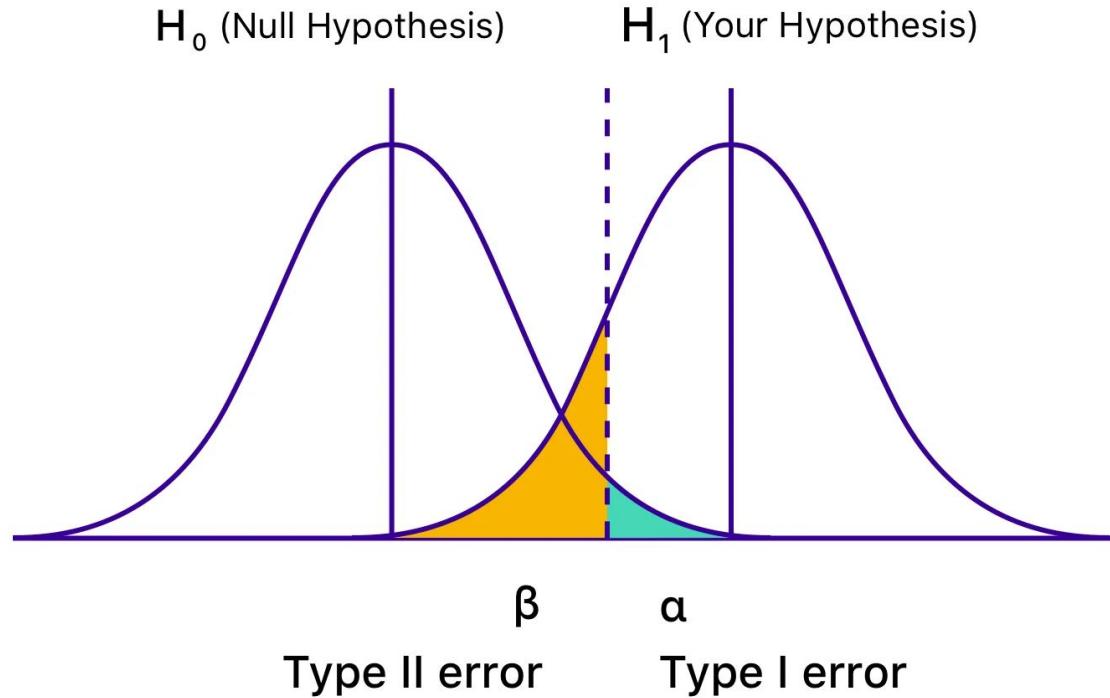
This is when you fail to reject a false Null Hypothesis. You don't find significance when it should be significant.

The probability of making a Type II Error is the value of beta which is a function of two factors: the actual difference between the samples and the sample size.

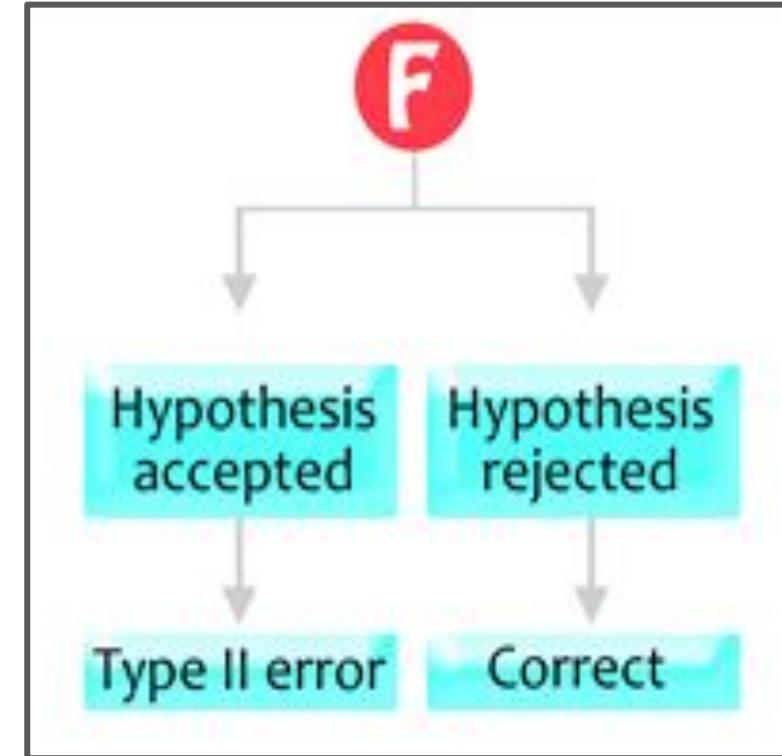
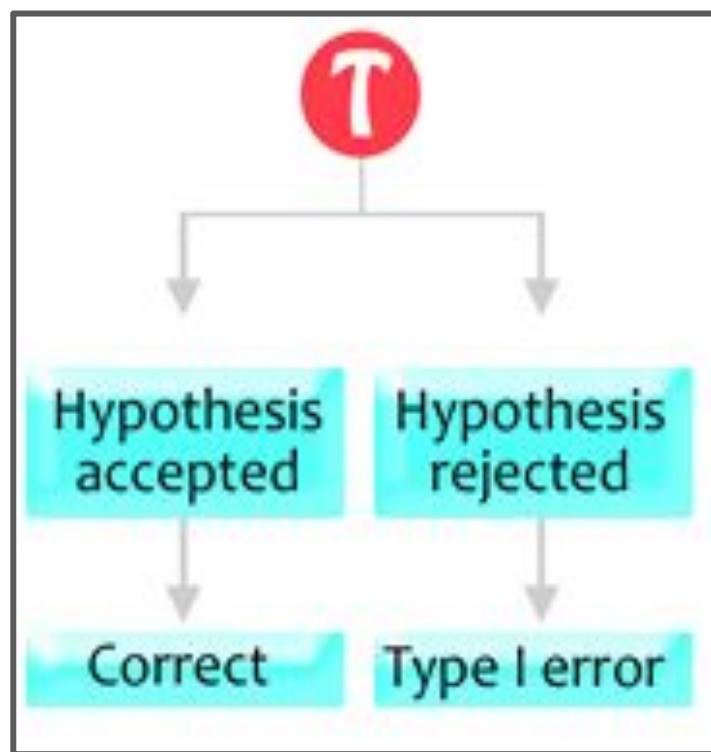
When you make a correct decision by failing to reject a true Null Hypothesis,  $p = (1-\alpha)$  which is the probability ( $p$ ) of being correct (this is not Power).

It is only possible to make a Type II Error if the Null Hypothesis is not rejected.

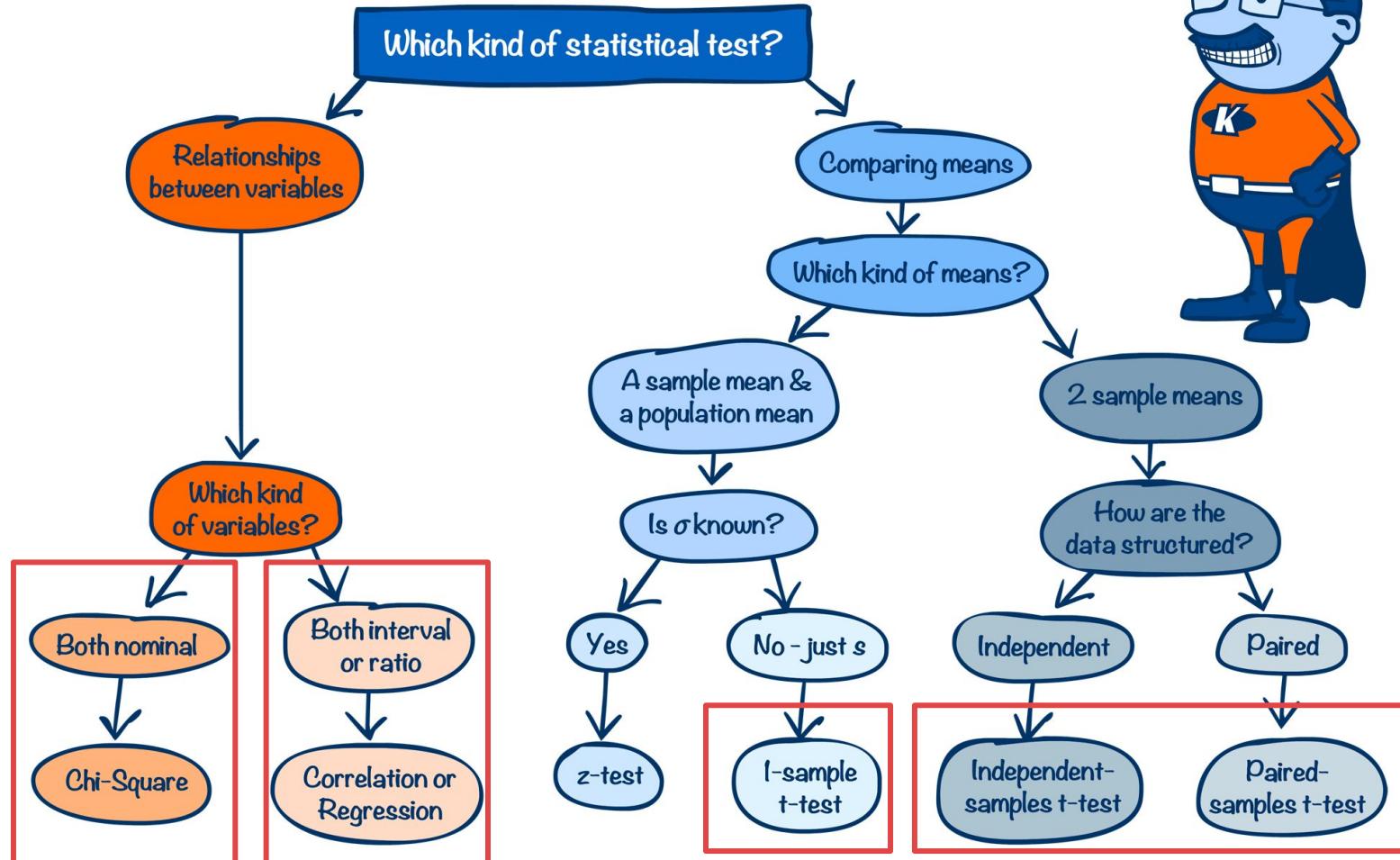
## Type I & Type II Errors in Hypothesis Testing



## Type I & Type II Errors in Hypothesis Testing



# STATISTICS DECISION TREE



- useful for analyzing such differences in categorical variables, especially those nominal in nature
- If observed frequencies in one or more categories match expected frequencies.
- depends on the size of the difference between actual and observed values, the degrees of freedom, and the samples size
- can be used to test whether two variables are related or independent from one another
- Most Common Two Types of Chi Square
  - Chi-square goodness of fit test
  - Chi-square test of independence.

1. Define your null and alternative hypotheses and collect your data.
2. Decide on the alpha value. This involves deciding the risk you are willing to take of drawing the wrong conclusion. Most common value is  $\alpha=0.05$ .
3. Check the data for errors.
4. Check the assumptions for the test
5. Perform the test and draw your conclusion.

- A statistical hypothesis test used to determine whether a variable is likely to come from a specified distribution or not.
- It is often used to evaluate whether sample data is representative of the full population.
- You can use the test when you have counts of values for a categorical variable.
- We have a set of data values, and an idea about how the data values are distributed.
- The test gives us a way to decide if the data values have a “good enough” fit to our idea, or if our idea is questionable.

- Test statistic:

$$\chi^2 = \sum \frac{(O_i - E)^2}{E}$$

$O_i$  = Frequency of Outcome (Original Frequency)

$E$  = Expected Frequency



a random sample of **10 bags**



**100 pieces** of candy and  
**five flavors**

1. Define your null and alternative hypotheses before collecting your data.

Null hypothesis  $H_0$  :

The proportions of the five flavors in each bag are the same.

Alternative hypothesis  $H_1$  :

The proportions of the five flavors in each bag are different.

1. Define your null and alternative hypotheses and collect your data.

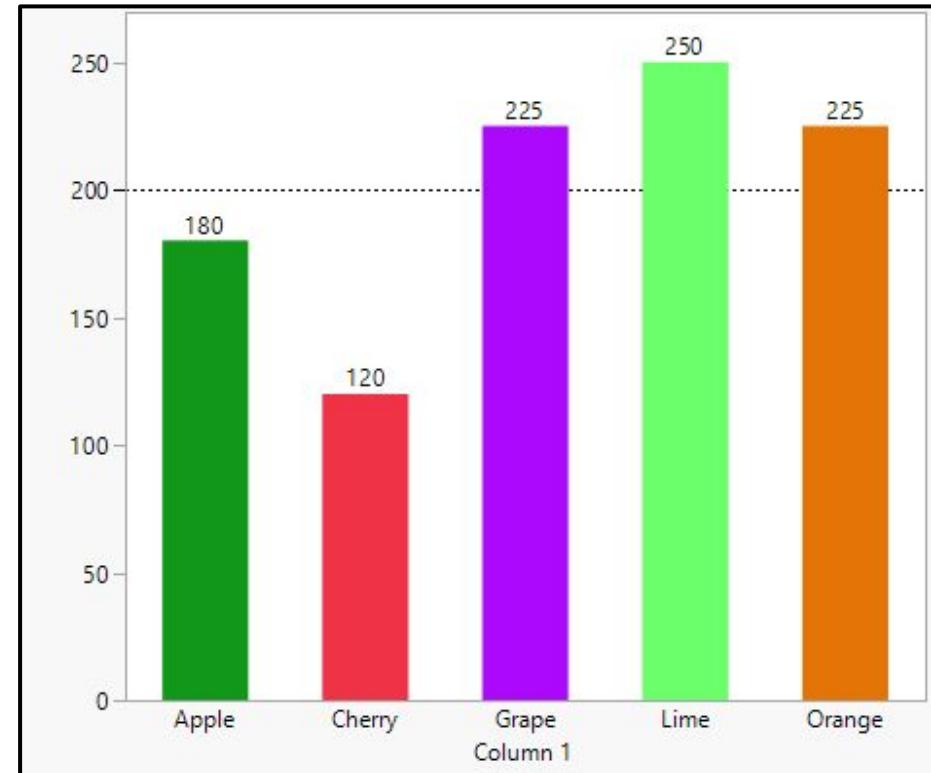
## Expected Frequency

- Each bag has 100 pieces of candy.
- Each bag has five flavors of candy.
- We expect to have equal numbers for each flavor.
- This means we expect  $100 / 5 = 20$  pieces of candy in each flavor from each bag.
- For 10 bags in our sample, we expect  $10 \times 20 = 200$  pieces of candy in each flavour

1. Define your null and alternative hypotheses and collect your data.

Actual Frequency

*Bar chart of counts of candy flavors from all 10 bags*



2. Decide on the **alpha value**. This involves deciding the risk you are willing to take of drawing the wrong conclusion. Most common value is  $\alpha=0.05$ .

For the candy data, we decide prior to collecting data that we are willing to take a 5% risk of concluding that the flavor counts in each bag across the full population are not equal when they really are. In statistics-speak, we set the significance level,  $\alpha$ , to 0.05.

### 3. Check the data for errors.

**Table 1: Comparison of actual vs expected number of pieces of each flavor of candy**

| Flavor | Number of Pieces of Candy<br>(10 bags) | Expected Number of<br>Pieces of Candy |
|--------|--|---------------------------------------|
| Apple  | 180                                    | 200                                   |
| Lime   | 250                                    | 200                                   |
| Cherry | 120                                    | 200                                   |
| Cherry | 225                                    | 200                                   |
| Grape  | 225                                    | 200                                   |

### 3. Check the data for errors.

**Table 2: Difference between observed and expected pieces of candy by flavor**

| Flavor | Number of Pieces of Candy<br>(10 bags) | Expected Number of<br>Pieces of Candy | Observed-Expected |
|--------|--|---------------------------------------|-------------------|
| Apple  | 180                                    | 200                                   | 180-200 = -20     |
| Lime   | 250                                    | 200                                   | 250-200 = 50      |
| Cherry | 120                                    | 200                                   | 120-200 = -80     |
| Orange | 225                                    | 200                                   | 225-200 = 25      |
| Grape  | 225                                    | 200                                   | 225-200 = 25      |

### 3. Check the data for errors.

**Table 3: Calculation of the squared difference between Observed and Expected for each flavor of candy**

| Flavor | Number of Pieces of Candy (10 bags) | Expected Number of Pieces of Candy | Observed-Expected | Squared Difference |
|--------|-------------------------------------|------------------------------------|-------------------|--------------------|
| Apple  | 180                                 | 200                                | 180-200 = -20     | 400                |
| Lime   | 250                                 | 200                                | 250-200 = 50      | 2500               |
| Cherry | 120                                 | 200                                | 120-200 = -80     | 1600               |
| Orange | 225                                 | 200                                | 225-200 = 25      | 625                |
| Grape  | 225                                 | 200                                | 225-200 = 25      | 625                |

### 3. Check the data for errors.

**Table 4: Calculation of the squared difference/expected number of pieces of candy per flavor**

| Flavor | Number of Pieces of Candy (10 bags) | Expected Number of Pieces of Candy | Observed-Expected | Squared Difference | Squared Difference/Expected Number |
|--------|-------------------------------------|------------------------------------|-------------------|--------------------|------------------------------------|
| Apple  | 180                                 | 200                                | 180-200 = -20     | 400                | 400/200=2                          |
| Lime   | 250                                 | 200                                | 250-200 = 50      | 2500               | 2500/200=12.5                      |
| Cherry | 120                                 | 200                                | 120-200 = -80     | 1600               | 1600/200=32                        |
| Orange | 225                                 | 200                                | 225-200 = 25      | 625                | 625/200=3.125                      |
| Grape  | 225                                 | 200                                | 225-200 = 25      | 625                | 625/200=3.125                      |

### 3. Check the data for errors.

Finally, we add the numbers in the final column to calculate our test statistic:

$$2 + 12.5 + 32 + 3.125 + 3.125 = 52.75$$

#### 4. Check the assumptions for the test

|                   |  |  |
|-------------------|--|--|
| Based on $\chi^2$ | $\chi^2_{\text{calculated}} < \chi^2_{\text{table}}$ | no statistically significant difference,<br><b>can not be rejected,</b><br>$H_0$         |
|                   | $\chi^2_{\text{calculated}} > \chi^2_{\text{table}}$ | statistically significant difference,<br><b><math>H_0</math> is rejected</b>             |
| Based on P value  | $P \text{ value}_{\text{table}} > \alpha = 0.05$     | no statistically significant difference,<br><b><math>H_0</math> can not be rejected.</b> |
|                   | $P \text{ value}_{\text{table}} < \alpha = 0.05$     | statistically significant difference<br><b><math>H_0</math> is rejected</b>              |

## 4. Check the assumptions for the test

 $\chi^2$  Table

| Right-tail area | df = 1 | df = 2 | df = 3 | df = 4 | df = 5 | Right-tail area | df = 6 | df = 7 | df = 8 | df = 9 | df = 10 |
|-----------------|--------|--------|--------|--------|--------|-----------------|--------|--------|--------|--------|---------|
| >0.100          | < 2.70 | < 4.60 | < 6.25 | < 7.77 | < 9.23 | >0.100          | <10.64 | <12.01 | <13.36 | <14.68 | <15.98  |
| 0.100           | 2.70   | 4.60   | 6.25   | 7.77   | 9.23   | 0.100           | 10.64  | 12.01  | 13.36  | 14.68  | 15.98   |
| 0.095           | 2.78   | 4.70   | 6.36   | 7.90   | 9.37   | 0.095           | 10.79  | 12.17  | 13.52  | 14.85  | 16.16   |
| 0.090           | 2.87   | 4.81   | 6.49   | 8.04   | 9.52   | 0.090           | 10.94  | 12.33  | 13.69  | 15.03  | 16.35   |
| 0.085           | 2.96   | 4.93   | 6.62   | 8.18   | 9.67   | 0.085           | 11.11  | 12.50  | 13.87  | 15.22  | 16.54   |
| 0.080           | 3.06   | 5.05   | 6.75   | 8.33   | 9.83   | 0.080           | 11.28  | 12.69  | 14.06  | 15.42  | 16.75   |
| 0.075           | 3.17   | 5.18   | 6.90   | 8.49   | 10.00  | 0.075           | 11.46  | 12.88  | 14.26  | 15.63  | 16.97   |
| 0.070           | 3.28   | 5.31   | 7.06   | 8.66   | 10.19  | 0.070           | 11.65  | 13.08  | 14.48  | 15.85  | 17.20   |
| 0.065           | 3.40   | 5.46   | 7.22   | 8.84   | 10.38  | 0.065           | 11.86  | 13.30  | 14.71  | 16.09  | 17.44   |
| 0.060           | 3.53   | 5.62   | 7.40   | 9.04   | 10.59  | 0.060           | 12.08  | 13.53  | 14.95  | 16.34  | 17.71   |
| 0.055           | 3.68   | 5.80   | 7.60   | 9.25   | 10.82  | 0.055           | 12.33  | 13.79  | 15.22  | 16.62  | 17.99   |
| 0.050           | 3.84   | 5.99   | 7.81   | 9.48   | 11.07  | 0.050           | 12.59  | 14.06  | 15.50  | 16.91  | 18.30   |
| 0.045           | 4.01   | 6.20   | 8.04   | 9.74   | 11.34  | 0.045           | 12.87  | 14.36  | 15.82  | 17.24  | 18.64   |
| 0.040           | 4.21   | 6.43   | 8.31   | 10.02  | 11.64  | 0.040           | 13.19  | 14.70  | 16.17  | 17.60  | 19.02   |
| 0.035           | 4.44   | 6.70   | 8.60   | 10.34  | 11.98  | 0.035           | 13.55  | 15.07  | 16.56  | 18.01  | 19.44   |
| 0.030           | 4.70   | 7.01   | 8.94   | 10.71  | 12.37  | 0.030           | 13.96  | 15.50  | 17.01  | 18.47  | 19.92   |
| 0.025           | 5.02   | 7.37   | 9.34   | 11.14  | 12.83  | 0.025           | 14.44  | 16.01  | 17.53  | 19.02  | 20.48   |
| 0.020           | 5.41   | 7.82   | 9.83   | 11.66  | 13.38  | 0.020           | 15.03  | 16.62  | 18.16  | 19.67  | 21.16   |
| 0.015           | 5.91   | 8.39   | 10.46  | 12.33  | 14.09  | 0.015           | 15.77  | 17.39  | 18.97  | 20.51  | 22.02   |
| 0.010           | 6.63   | 9.21   | 11.34  | 13.27  | 15.08  | 0.010           | 16.81  | 18.47  | 20.09  | 21.66  | 23.20   |
| 0.005           | 7.87   | 10.59  | 12.83  | 14.86  | 16.74  | 0.005           | 18.54  | 20.27  | 21.95  | 23.58  | 25.18   |
| 0.001           | 10.82  | 13.81  | 16.26  | 18.46  | 20.51  | 0.001           | 22.45  | 24.32  | 26.12  | 27.87  | 29.58   |
| <0.001          | >10.82 | >13.81 | >16.26 | >18.46 | >20.51 | <0.001          | >22.45 | >24.32 | >26.12 | >27.87 | >29.58  |

#### 4. Check the assumptions for the test

#### Degree of Freedom and P value

- For the goodness of fit test, Degree of freedom is one fewer than the number of categories.
- We have five flavors of candy, so we have  $5 - 1 = 4$  degrees of freedom.

- P value Calculator: [Click Here](#)
- $\chi^2 = 52.75$
- df= 4
- P Value: 0.00

## 4. Check the assumptions for the test

## Degree of Freedom and P value

$\alpha = 0.05$  and 4 degrees of freedom is 9.488.

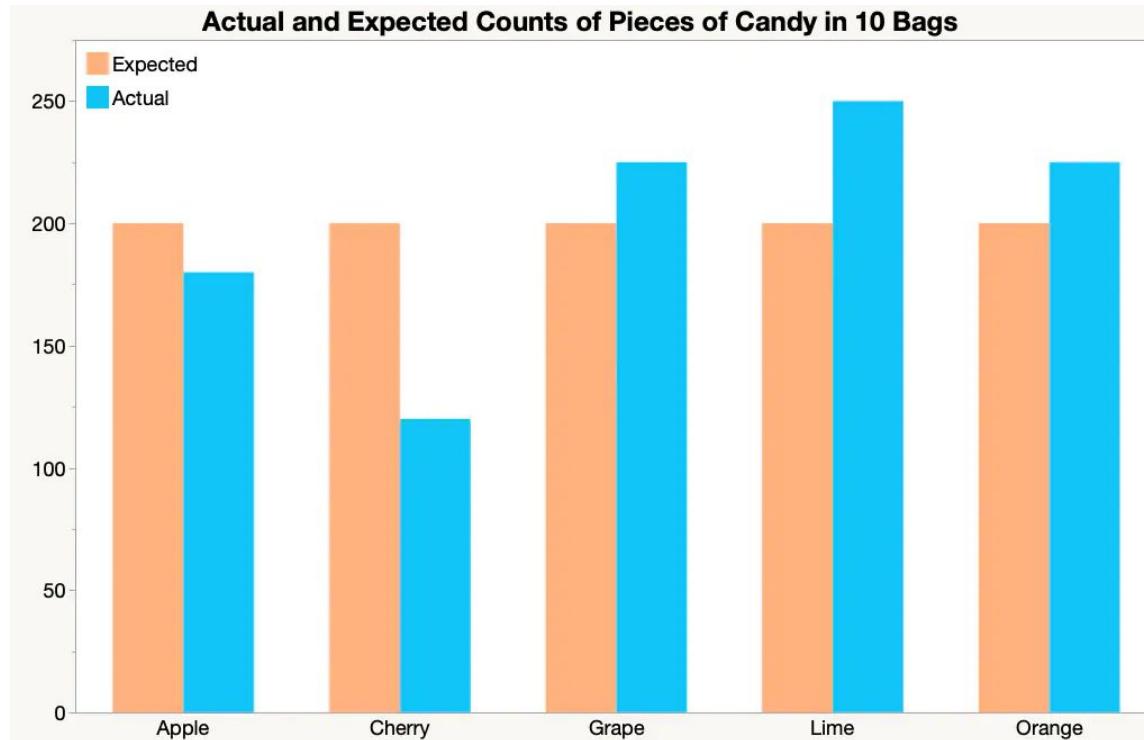
| Right-tail area | df = 1 | df = 2 | df = 3 | df = 4 | df = 5 |
|-----------------|--------|--------|--------|--------|--------|
| >0.100          | < 2.70 | < 4.60 | < 6.25 | < 7.77 | < 9.23 |
| 0.100           | 2.70   | 4.60   | 6.25   | 7.77   | 9.23   |
| 0.095           | 2.78   | 4.70   | 6.36   | 7.90   | 9.37   |
| 0.090           | 2.87   | 4.81   | 6.49   | 8.04   | 9.52   |
| 0.085           | 2.96   | 4.93   | 6.62   | 8.18   | 9.67   |
| 0.080           | 3.06   | 5.05   | 6.75   | 8.33   | 9.83   |
| 0.075           | 3.17   | 5.18   | 6.90   | 8.49   | 10.00  |
| 0.070           | 3.28   | 5.31   | 7.06   | 8.66   | 10.19  |
| 0.065           | 3.40   | 5.46   | 7.22   | 8.84   | 10.38  |
| 0.060           | 3.53   | 5.62   | 7.40   | 9.04   | 10.59  |
| 0.055           | 3.68   | 5.80   | 7.60   | 9.25   | 10.82  |
| 0.050           | 3.84   | 5.99   | 7.81   | 9.48   | 11.07  |
| 0.045           | 4.01   | 6.20   | 8.04   | 9.74   | 11.34  |

## 5. Perform the test and draw your conclusion.

- The value of our test statistic (52.75) to the Chi-square value.
- Since  $52.75 > 9.488$  ( $X^2_{\text{calculated}} > X^2_{\text{table}}$ )
- we reject the null hypothesis that the proportions of flavors of candy are equal

- The value of P Value is 0.000 < 0.05 ( $P \text{ value}_{\text{table}} > \alpha$ )
- we reject the null hypothesis that the proportions of flavors of candy are equal

## Interpretation of Results



## Practise Example

A cubical die is thrown 300 times and results are obtained as follows.  
Is the die unbiased. (Univariate)

| Point on die | 1  | 2  | 3  | 4  | 5  | 6  |
|--------------|----|----|----|----|----|----|
| Frequency    | 41 | 44 | 49 | 53 | 57 | 56 |

## Practise Example

| Point x | Frequency $f_o$ | Expected $f_e$ | $f_o - f_e$ | $(f_o - f_e)^2$ | $(f_o - f_e)^2 / f_e$ |
|---------|-----------------|----------------|-------------|-----------------|-----------------------|
| 1       | 41              | 50             | -9          | 81              | 1.62                  |
| 2       | 44              | 50             | -6          | 36              | 0.72                  |
| 3       | 49              | 50             | -1          | 1               | 0.02                  |
| 4       | 53              | 50             | 3           | 9               | 0.18                  |
| 5       | 57              | 50             | 7           | 49              | 0.98                  |
| 6       | 56              | 50             | 6           | 36              | 0.72                  |
|         | N=300           |                |             |                 | $\chi^2 = 4.24$       |

- The Chi-square test of independence is a **statistical hypothesis test used to determine whether two categorical or nominal variables are likely to be related or not.**
- You can use the test when you have **counts of values for two categorical variables.**

- Test statistic:

$$E = \frac{\text{Row Total} \times \text{Column Total}}{\text{Total no of experiments}}$$

1. Define your null and alternative hypotheses and collect your data.

## Expected Frequency

$H_0$  Movie type and snacks purchase are independent.

$H_1$  Movie type and snacks purchase are not independent.

2. Decide on the **alpha value**. This involves deciding the risk you are willing to take of drawing the wrong conclusion. Most common value is  $\alpha=0.05$ .

we had decided prior to our data collection that we are **willing to take a 5% risk** of saying that the two variables – **Movie Type and Snack Purchase** – are **not independent** when they really are independent. In statistics-speak, we set the significance level,  $\alpha$ , to 0.05.

### 3. Check the data for errors.

Table 1: Contingency table for movie snacks data

| Type of Movie | Snacks | No Snacks |
|---------------|--------|-----------|
| Action        | 50     | 75        |
| Comedy        | 125    | 175       |
| Family        | 90     | 30        |
| Horror        | 45     | 10        |

### 3. Check the data for errors.

**Table 2: Contingency table for movie snacks data with row and column totals**

| Type of Movie        | Snacks     | No Snacks  | Row totals               |
|----------------------|------------|------------|--------------------------|
| Action               | 50         | 75         | 125                      |
| Comedy               | 125        | 175        | 300                      |
| Family               | 90         | 30         | 120                      |
| Horror               | 45         | 10         | 55                       |
| <b>Column totals</b> | <b>310</b> | <b>290</b> | <b>GRAND TOTAL = 600</b> |

### 3. Check the data for errors.

$$E_{\text{Action, Snacks}} = \frac{125 \times 310}{600} = \frac{38,750}{600} = 65$$

$$E_{\text{Comedy, Snacks}} = 155$$

$$E_{\text{Family, Snacks}} = 62$$

$$E_{\text{Horror, Snacks}} = 28.42$$

Table 3: Contingency table for movie snacks data showing actual count vs. expected count

| Type of Movie | Snacks | No Snacks | Row totals        |
|---------------|--------|-----------|-------------------|
| Action        | 50     | 75        | 125               |
|               | 65     | 60        |                   |
| Comedy        | 125    | 175       | 300               |
|               | 155    | 145       |                   |
| Family        | 90     | 30        | 120               |
|               | 62     | 58        |                   |
| Horror        | 45     | 10        | 55                |
|               | 28     | 27        |                   |
| Column totals | 310    | 290       | GRAND TOTAL = 600 |

Table 4: Preparing to calculate our test statistic

| Type of Movie | Snack  | No Snacks   |
|---------------|--|---|
| Action        | Actual: 50<br><br><b>Expected: 64.58</b><br><br>Difference: $50 - 64.58 = -14.58$<br><br>Squared Difference: 212.67<br><br>Divide by Expected: $212.67 / 64.58 = 3.29$ | Actual: 75<br><br><b>Expected: 60.42</b><br><br>Difference: $75 - 60.42 = 14.58$<br><br>Squared Difference: 212.67<br><br>Divide by Expected: $212.67 / 60.42 = 3.52$ |

Table 4: Preparing to calculate our test statistic

| Type of Movie | Snack  | No Snacks   |
|---------------|--|---|
| <b>Comedy</b> | Actual:125<br><br><b>Expected: 155</b><br><br>Difference: $125 - 155 = -30$<br><br>Squared Difference: 900<br><br>Divide by Expected: $900/155 = 5.81$ | Actual: 175<br><br><b>Expected:145</b><br><br>Difference: $175 - 145 = 30$<br><br>Squared Difference: 900<br><br>Divide by Expected: $900/145 = 6.21$ |

Table 4: Preparing to calculate our test statistic

| Type of Movie | Snack   | No Snacks  |
|---------------|---|--|
| <b>Family</b> | Actual:90<br><br><b>Expected: 62</b><br><br>Difference: $90 - 62 = 28$<br><br>Squared Difference: 784<br><br>Divide by Expected: $784/62 = 12.65$ | Actual: 30<br><br><b>Expected:58</b><br><br>Difference: $30 - 58 = -28$<br><br>Squared Difference: 784<br><br>Divide by Expected: $784/58 = 13.52$ |

Table 4: Preparing to calculate our test statistic

| Type of Movie | Snack   | No Snacks  |
|---------------|---|--|
| <b>Horror</b> | Actual:45<br><br><b>Expected:28.42</b><br><br>Difference: $45 - 28.42 = 16.58$<br><br>Squared Difference: 275.01<br><br>Divide by Expected: $275.01/28.42 = 9.68$ | Actual: 10<br><br><b>Expected:26.58</b><br><br>Difference: $10 - 26.58 = -16.58$<br><br>Squared Difference: 275.01<br><br>Divide by Expected: $275.01/26.58 = 10.35$ |

### 3. Check the data for errors.

Finally, we add the numbers in the final column to calculate our test statistic:

$$3.29 + 3.52 + 5.81 + 6.21 + 12.65 + 13.52 + 9.68 + 10.35 = \mathbf{65.03}$$

## 4. Check the assumptions for the test

## Degree of Freedom and P value

- For the goodness of fit test, Degree of freedom is depend on how many rows and how many columns

$$df = (r - 1) \times (c - 1) \quad df = (4 - 1) \times (2 - 1) = 3 \times 1 = 3$$

- P value Calculator: [Click Here](#)
- $\chi^2 = 65.03$
- $df = 3$
- P Value: 0.00

## 4. Check the assumptions for the test

## Degree of Freedom and P value

$\alpha = 0.05$  and 3 degrees of freedom is 9.488.

| Right-tail area | df = 1 | df = 2 | df = 3 | df = 4 | df = 5 |
|-----------------|--------|--------|--------|--------|--------|
| >0.100          | < 2.70 | < 4.60 | < 6.25 | < 7.77 | < 9.23 |
| 0.100           | 2.70   | 4.60   | 6.25   | 7.77   | 9.23   |
| 0.095           | 2.78   | 4.70   | 6.36   | 7.90   | 9.37   |
| 0.090           | 2.87   | 4.81   | 6.49   | 8.04   | 9.52   |
| 0.085           | 2.96   | 4.93   | 6.62   | 8.18   | 9.67   |
| 0.080           | 3.06   | 5.05   | 6.75   | 8.33   | 9.83   |
| 0.075           | 3.17   | 5.18   | 6.90   | 8.49   | 10.00  |
| 0.070           | 3.28   | 5.31   | 7.06   | 8.66   | 10.19  |
| 0.065           | 3.40   | 5.46   | 7.22   | 8.84   | 10.38  |
| 0.060           | 3.53   | 5.62   | 7.40   | 9.04   | 10.59  |
| 0.055           | 3.68   | 5.80   | 7.60   | 9.25   | 10.82  |
| 0.050           | 3.84   | 5.99   | 7.81   | 9.48   | 11.07  |
| 0.045           | 4.01   | 6.20   | 8.04   | 9.74   | 11.34  |

## 5. Perform the test and draw your conclusion.

- The value of our test statistic (65.03) to the Chi-square value.
- Since  $65.03 > 7.81$  ( $X^2_{\text{calculated}} > X^2_{\text{table}}$  )
- we reject the null hypothesis that Movie Type and Snack purchases are independent

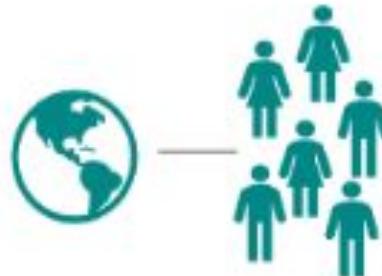
- The value of P Value is  $0.00 < 0.05$  ( $P \text{ value}_{\text{table}} < \alpha$ )
- we reject the null hypothesis that Movie Type and Snack purchases are independent

## Practise Example

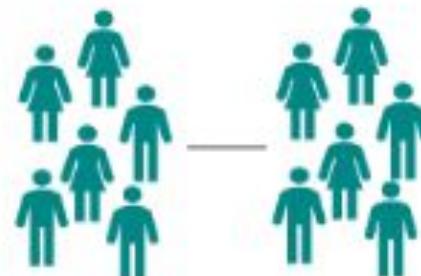
| Point x | Frequency $f_o$ | Expected $f_e$ | $f_o - f_e$ | $(f_o - f_e)^2$ | $(f_o - f_e)^2 / f_e$ |
|---------|-----------------|----------------|-------------|-----------------|-----------------------|
| 1       | 41              | 50             | -9          | 81              | 1.62                  |
| 2       | 44              | 50             | -6          | 36              | 0.72                  |
| 3       | 49              | 50             | -1          | 1               | 0.02                  |
| 4       | 53              | 50             | 3           | 9               | 0.18                  |
| 5       | 57              | 50             | 7           | 49              | 0.98                  |
| 6       | 56              | 50             | 6           | 36              | 0.72                  |
|         | N=300           |                |             |                 | $\chi^2 = 4.24$       |

- A t-test (**also known as Student's t-test**)
- a tool for evaluating the means of **one or two populations using hypothesis testing**.
- A t-test may be used to evaluate whether
  - a single group differs from a known value (a one-sample t-test),
  - two groups differ from each other (an independent two-sample t-test),
  - there is a significant difference in paired measurements (a paired, or dependent samples t-test).

- A t-test (**also known as Student's t-test**)
- a tool for evaluating the means of **one or two populations using hypothesis testing**.
- A t-test may be used to evaluate whether
  - a single group differs from a known value (a one-sample t-test),
  - two groups differ from each other (an independent two-sample t-test),
  - there is a significant difference in paired measurements (a paired, or dependent samples t-test).

**One sample t-test**

Is there a **difference** between a **group** and the **population**

**Independent samples t-test**

Is there a **difference** between **two groups**

**Paired samples t-test**

Is there a **difference** in a **group** between **two points in time**

|                               | One-sample <i>t</i> -test  | Two-sample <i>t</i> -test   | Paired <i>t</i> -test   |
|-------------------------------|--|---|---|
| Synonyms                      | Student's <i>t</i> -test   | <ul style="list-style-type: none"> <li>• Independent groups <i>t</i>-test</li> <li>• Independent samples <i>t</i>-test</li> <li>• Equal variances <i>t</i>-test</li> <li>• Pooled <i>t</i>-test</li> <li>• Unequal variances <i>t</i>-test</li> </ul> | <ul style="list-style-type: none"> <li>• Paired groups <i>t</i>-test</li> <li>• Dependent samples <i>t</i>-test</li> </ul>                  |
| Number of variables           | One  | Two   | Two   |
| Type of variable              | <ul style="list-style-type: none"> <li>• Continuous measurement</li> </ul> | <ul style="list-style-type: none"> <li>• Continuous measurement</li> <li>• Categorical or Nominal to define groups</li> </ul>   | <ul style="list-style-type: none"> <li>• Continuous measurement</li> <li>• Categorical or Nominal to define pairing within group</li> </ul> |
| Purpose of test               | Decide if the population mean is equal to a specific value or not          | Decide if the population means for two different groups are equal or not  | Decide if the difference between paired measurements for a population is zero or not  |
| Example: test if...           | Mean heart rate of a group of people is equal to 65 or not                 | Mean heart rates for two groups of people are the same or not   | Mean difference in heart rate for a group of people before and after exercise is zero or not  |
| Estimate of population mean   | Sample average   | Sample average for each group   | Sample average of the differences in paired measurements  |
| Population standard deviation | Unknown, use sample standard deviation                                     | Unknown, use sample standard deviations for each group  | Unknown, use sample standard deviation of differences in paired measurements  |
| Degrees of freedom            | Number of observations in sample minus 1, or: $n - 1$                      | Sum of observations in each sample minus 2, or: $n_1 + n_2 - 2$   | Number of paired observations in sample minus 1, or: $n - 1$  |

1. Define your null and alternative hypotheses and collect your data.
2. Decide on the alpha value. This involves deciding the risk you are willing to take of drawing the wrong conclusion. Most common value is  $\alpha=0.05$ .
3. Check the data for errors.
4. Check the assumptions for the test
5. Perform the test and draw your conclusion. t-tests for means involve calculating a test statistic. You compare the test statistic to a theoretical value from the t-distribution. The theoretical value involves both the  $\alpha$  value and the degrees of freedom for your data.

## One Sample T- Test

- To compare a sample mean with the population mean.
- For a valid test, we need data values that are:
  - Independent (values are not related to one another).
  - Continuous.
  - Obtained via a simple random sample from the population

## 1. Define your null and alternative hypotheses and collect your data.

Let's say we want to determine if on average girls score more than 600 in the exam. We do not have the information related to variance (or standard deviation) for girls' scores. To perform t-test, we randomly collect the data of 10 girls with their marks

$$H_0: \mu \leq 600$$

$$H_1: \mu > 600$$

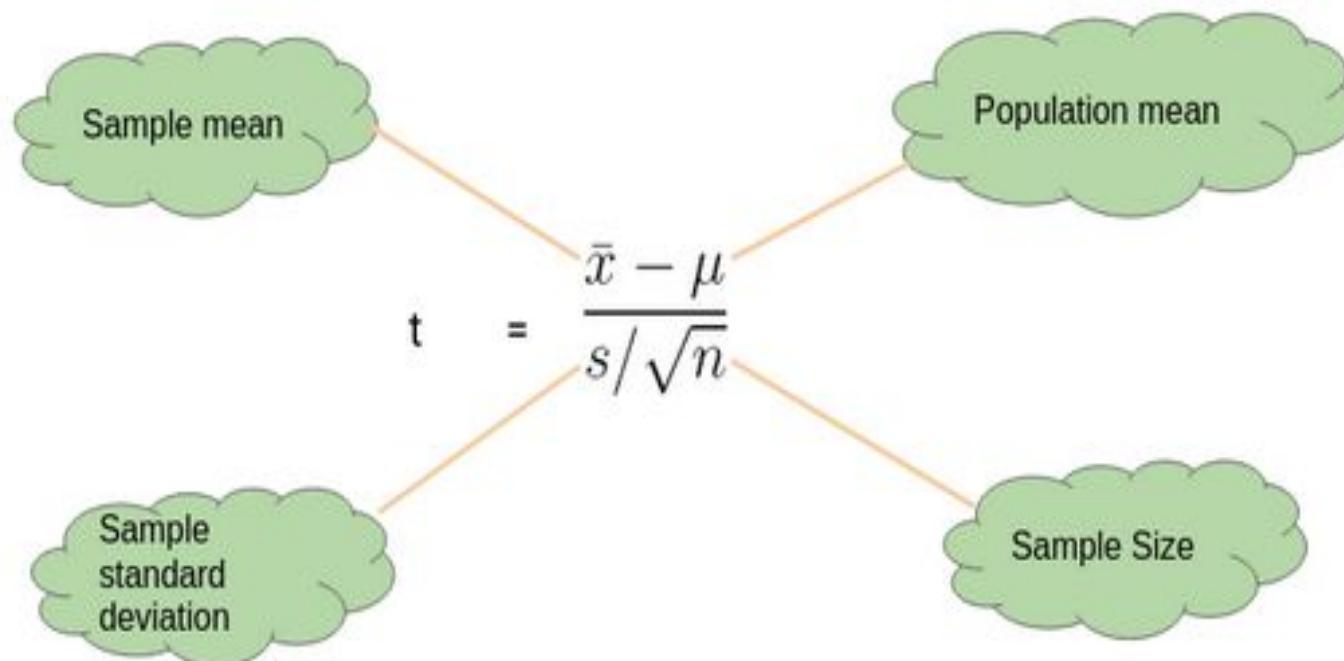
1. Define your null and alternative hypotheses and collect your data.



2. Decide on the **alpha value**. This involves deciding the risk you are willing to take of drawing the wrong conclusion. Most common value is  $\alpha=0.05$ .

$\alpha=0.05$ .

### 3. Check the data for errors.



### 3. Check the data for errors.

- The sample mean( $\bar{x}$ ) = 606.8
- The population mean( $\mu$ )= 600
- The sample standard deviation( $s$ ) = 13.14
- Number of observations( $n$ ) =10

### 3. Check the data for errors.

$$\begin{aligned} t &= \frac{\bar{x} - \mu}{s/\sqrt{n}} \\ &= \frac{606.8 - 600}{13.14/\sqrt{10}} \\ &= 1.64 \end{aligned}$$

#### 4. Check the assumptions for the test

|                  |  |  |
|------------------|--|--|
| Based on t score | $t \text{ score}_{\text{calculated}} < t \text{ score}_{\text{table}}$ | no statistically significant difference,<br><b>can not be rejected,</b><br>$H_0$         |
|                  | $t \text{ score}_{\text{calculated}} > t \text{ score}_{\text{table}}$ | statistically significant difference,<br><b><math>H_0</math> is rejected</b>             |
| Based on P value | $P \text{ value}_{\text{table}} > \alpha = 0.05$                       | no statistically significant difference,<br><b><math>H_0</math> can not be rejected.</b> |
|                  | $P \text{ value}_{\text{table}} < \alpha = 0.05$                       | statistically significant difference<br><b><math>H_0</math> is rejected</b>              |

## 4. Check the assumptions for the test

## Degree of Freedom and P value

For the goodness of fit test Degree of freedom is one fewer than the number of samples.

$$df = 10 - 1 = 9$$

- P value Calculator: [Click Here](#)
- $t \text{ score}_{\text{calculated}} = 1.64$
- $t \text{ score}_{\text{table}} = 1.833$  [Click here](#)
- $df = 9$
- P Value: 0.06

## 5. Perform the test and draw your conclusion.

- The value of **tscore** is **1.64**
- Since **1.64 < 1.83** ( $t \text{ score}_{\text{calculated}} < t \text{ score}_{\text{table}}$ )
- **we can not reject the null hypothesis.** and don't have enough evidence to support the hypothesis that on average, girls score more than 600 in the exam.

- The value of **P Value** is **0.06 < 0.05** ( $P \text{ value}_{\text{table}} > \alpha = 0.05$ )
- **we cannot reject the null hypothesis.**

## Two Sample T- Test

- to compare the mean of two samples.
- For a valid test, we need data values that are
  - randomly sampled from two normal populations
  - Obtained via a simple random sample from the population
  - do not have the information related to variance (or standard deviation)

## 1. Define your null and alternative hypotheses and collect your data.

let's say we want to determine if on average, boys score 15 marks more than girls in the exam. We do not have the information related to variance (or standard deviation) for girls' scores or boys' scores. To perform a t-test. we randomly collect the data of 10 girls and boys with their marks.

$$H_0: \mu_1 - \mu_2 \leq 15$$

$$H_1: \mu_1 - \mu_2 > 15$$

1. Define your null and alternative hypotheses and collect your data.



Girls\_Score

587

602

627

610

619

622

605

608

596

592



Boys\_Score

626

643

647

634

630

649

625

623

617

607

2. Decide on the **alpha value**. This involves deciding the risk you are willing to take of drawing the wrong conclusion. Most common value is  $\alpha=0.05$ .

$\alpha=0.05$ .

## 3. Check the data for errors.

Difference bw  
Sample mean  
 $\bar{x}_1 - \bar{x}_2$

Difference bw  
population mean  
 $\mu_1 - \mu_2$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Sample standard  
deviation  $s_{1, s_2}$

Sample Size  
 $n_1, n_2$

### 3. Check the data for errors.

- Mean Score for Boys is **630.1**
- Mean Score for Girls is **606.8**
- Difference between Population Mean **15**
- Standard Deviation for Boys' score is **13.42**
- Standard Deviation for Girls' score is **13.14**

3. Check the data for errors.

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$
$$\frac{(630.1 - 606.8) - (15)}{\sqrt{\frac{(13.42)^2}{10} + \frac{(13.14)^2}{10}}}$$
$$= 1.833$$

#### 4. Check the assumptions for the test

|                  |  |  |
|------------------|--|--|
| Based on t score | $t \text{ score}_{\text{calculated}} < t \text{ score}_{\text{table}}$ | no statistically significant difference,<br><b>can not be rejected,</b><br>$H_0$ |
|                  | $t \text{ score}_{\text{calculated}} > t \text{ score}_{\text{table}}$ | statistically significant difference,<br>$H_0$ is rejected                       |
| Based on P value | $P \text{ value}_{\text{table}} > \alpha = 0.05$                       | no statistically significant difference,<br>$H_0$ can not be rejected.           |
|                  | $P \text{ value}_{\text{table}} < \alpha = 0.05$                       | statistically significant difference<br>$H_0$ is rejected                        |

## 4. Check the assumptions for the test

## Degree of Freedom and P value

For the goodness of fit test Degree of freedom is degrees of freedom for the problem is the smaller of

$$n_1 - 1 \text{ and } n_2 - 1. \quad df = (10-1) + (10-1) = 18$$

- P value Calculator: [Click Here](#)
- $t \text{ score}_{\text{calculated}} = 1.833$
- $t \text{ score}_{\text{table}} = 1.73$  [Click here](#)
- $df = 18$
- P Value: 0.041

## 5. Perform the test and draw your conclusion.

- The value of **tscore** is **1.64**
- Since **1.833 > 1.73** ( $t \text{ score}_{\text{calculated}} > t \text{ score}_{\text{table}}$ )
- we reject the null hypothesis and conclude that on average boys score 15 marks more than girls in the exam.

- The value of **P Value** is **0.04 < 0.05** ( $P \text{ value}_{\text{table}} < \alpha=0.05$ )
- We reject the null hypothesis.

## Paired T- Test

### Paired Samples T-Test Example

**Observation 1:** A group of people were evaluated at baseline.

**Observation 2:** This same group of people were evaluated after a 12-week exercise program.

**Variable of interest:** Cholesterol levels.

## Paired T- Test

- To compare means from the same group at different times
- Subjects must be independent. Measurements for one subject do not affect measurements for any other subject.
- Each of the paired measurements must be obtained from the same subject. measured differences are normally distributed.

## 1. Define your null and alternative hypotheses and collect your data.

An instructor wants to use two exams in her classes next year. This year, she gives both exams to the students. She wants to know if the exams are equally difficult and wants to check this by looking at the differences between scores. If the mean difference between scores for students is “close enough” to zero, she will make a practical conclusion that the exams are equally difficult.

$$H_0: \mu_d = 0$$

$$H_1: \mu_d \neq 0$$

## 1. Define your null and alternative hypotheses and collect your data.

| Student   | Exam 1 Score | Exam 2 Score |
|-----------|--------------|--------------|
| Pooja     | 63           | 69           |
| Nisha     | 65           | 65           |
| Gayatri   | 56           | 62           |
| Khushbu   | 100          | 91           |
| Raj       | 88           | 78           |
| Akanksha  | 83           | 87           |
| Aishwarya | 77           | 79           |
| Govinda   | 92           | 88           |
| Nayan     | 90           | 85           |
| Tejal     | 84           | 92           |
| Sakshi    | 68           | 69           |
| Shruti    | 74           | 81           |
| Amol      | 87           | 84           |
| Neha      | 64           | 75           |
| Charmy    | 71           | 84           |
| Varun     | 88           | 82           |

2. Decide on the **alpha value**. This involves deciding the risk you are willing to take of drawing the wrong conclusion. Most common value is  $\alpha=0.05$ .

$\alpha=0.05$ .

### 3. Check the data for errors.

$$t = \frac{\bar{x}_d - \mu_d}{\left( \frac{s_d}{\sqrt{n}} \right)}, \quad df = n - 1$$

$\bar{x}_d$  : sample mean difference

$\mu$  : population mean difference

$s$  : sample difference standard deviation

$n$  : sample size

## 3. Check the data for errors.

| Student   | Exam 1 Score | Exam 2 Score | Difference | Difference - Mean | (Difference - Mean) <sup>2</sup> |
|-----------|--------------|--------------|------------|-------------------|----------------------------------|
| Pooja     | 63           | 69           | 6          | 4.6875            | 21.97                            |
| Nisha     | 65           | 65           | 0          | -1.3125           | 1.72                             |
| Gayatri   | 56           | 62           | 6          | 4.6875            | 21.97                            |
| Khushbu   | 100          | 91           | -9         | -10.3125          | 106.35                           |
| Raj       | 88           | 78           | -10        | -11.3125          | 127.97                           |
| Akanksha  | 83           | 87           | 4          | 2.6875            | 7.22                             |
| Aishwarya | 77           | 79           | 2          | 0.6875            | 0.47                             |
| Govinda   | 92           | 88           | -4         | -5.3125           | 28.22                            |
| Nayan     | 90           | 85           | -5         | -6.3125           | 39.85                            |
| Tejal     | 84           | 92           | 8          | 6.6875            | 44.72                            |
| Sakshi    | 68           | 69           | 1          | -0.3125           | 0.1                              |
| Shruti    | 74           | 81           | 7          | 5.6875            | 32.35                            |
| Amol      | 87           | 84           | -3         | -4.3125           | 18.6                             |
| Neha      | 64           | 75           | 11         | 9.6875            | 93.85                            |
| Charmy    | 71           | 84           | 13         | 11.6875           | 136.6                            |
| Varun     | 88           | 82           | -6         | -7.3125           | 53.47                            |

## 3. Check the data for errors.

| Student   | Exam 1 Score | Exam 2 Score | Difference | Difference - Mean | (Difference - Mean) <sup>2</sup> |
|-----------|--------------|--------------|------------|-------------------|----------------------------------|
| Pooja     | 63           | 69           | 6          | 4.6875            | 21.97                            |
| Nisha     | 65           | 65           | 0          | -1.3125           | 1.72                             |
| Gayatri   | 56           | 62           | 6          | 4.6875            | 21.97                            |
| Khushbu   | 100          | 91           | -9         | -10.3125          | 106.35                           |
| Raj       | 88           | 78           | -10        | -11.3125          | 127.97                           |
| Akanksha  | 83           | 87           | 4          | 2.6875            | 7.22                             |
| Aishwarya | 77           | 79           | 2          | 0.6875            | 0.47                             |
| Govinda   | 92           | 88           | -4         | -5.3125           | 28.22                            |
| Nayan     | 90           | 85           | -5         | -6.3125           | 39.85                            |
| Tejal     | 84           | 92           | 8          | 6.6875            | 44.72                            |
| Sakshi    | 68           | 69           | 1          | -0.3125           | 0.1                              |
| Shruti    | 74           | 81           | 7          | 5.6875            | 32.35                            |
| Amol      | 87           | 84           | -3         | -4.3125           | 18.6                             |
| Neha      | 64           | 75           | 11         | 9.6875            | 93.85                            |
| Charmy    | 71           | 84           | 13         | 11.6875           | 136.6                            |
| Varun     | 88           | 82           | -6         | -7.3125           | 53.47                            |

|  |        |
|--|--------|
| Average of Difference                  | 1.3125 |
| <b>(Difference - Mean)<sup>2</sup></b> |        |
| Total                                  | 735.43 |
| Average                                | 49.03  |
| Standard Deviation                     | 7      |

## 3. Check the data for errors.

- The average score difference is :  $\bar{x}_d = 1.31$
- The Population difference is  $\mu_d = 0$
- Standard deviation of sample difference= 7
- The sample size n= 16

$$t = \frac{\bar{x}_d - \mu_d}{\left( \frac{s_d}{\sqrt{n}} \right)}$$
$$t = \frac{1.31 - 0}{\frac{7}{\sqrt{16}}}$$

$$t= 0.75$$

#### 4. Check the assumptions for the test

|                  |  |  |
|------------------|--|--|
| Based on t score | $t \text{ score}_{\text{calculated}} < t \text{ score}_{\text{table}}$ | no statistically significant difference,<br><b>can not be rejected,</b><br>$H_0$ |
|                  | $t \text{ score}_{\text{calculated}} > t \text{ score}_{\text{table}}$ | statistically significant difference,<br>$H_0$ is rejected                       |
| Based on P value | $P \text{ value}_{\text{table}} > \alpha = 0.05$                       | no statistically significant difference,<br>$H_0$ can not be rejected.           |
|                  | $P \text{ value}_{\text{table}} < \alpha = 0.05$                       | statistically significant difference<br>$H_0$ is rejected                        |

## 4. Check the assumptions for the test

## Degree of Freedom and P value

For the goodness of fit test Degree of freedom is number of instances -1

$$df = n - 1 = 16 - 1 = 15$$

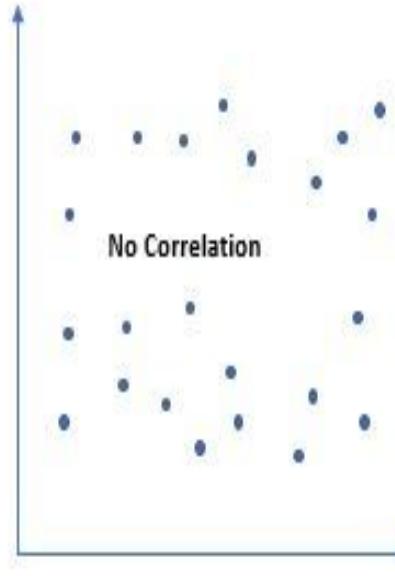
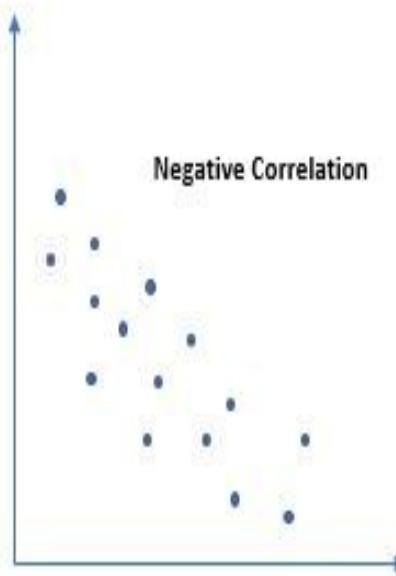
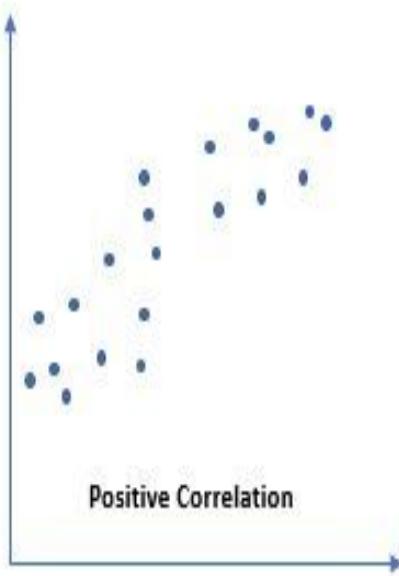
- P value Calculator: [Click Here](#)
- $t \text{ score}_{\text{calculated}} = 0.75$
- $t \text{ score}_{\text{table}} = 2.131$  (Two Tail Test) [Click here](#)
- $df = 15$
- P Value: 0.46

## 5. Perform the test and draw your conclusion.

- The value of **tscore** is **0.75**
- Since  $0.75 < 2.131$  ( $t \text{ score}_{\text{calculated}} < t \text{ score}_{\text{table}}$ )
- we fail to reject the null hypothesis and conclude. The instructor can go ahead with her plan to use both exams next year, and give half the students one exam and half the other exam.

- The value of **P Value** is **0.46 > 0.05** ( $P \text{ value}_{\text{table}} > \alpha = 0.05$ )
- We fail to reject the null hypothesis.

- Correlation is a bi-variate analysis that measures **the strength of association between two variables and the direction of the relationship.**
- The correlation coefficient varies between **+1 and -1.**
- A value of  **$\pm 1$  indicates a perfect degree of association** between the two variables.
- As the correlation coefficient value goes towards 0, the relationship between the two variables will be **weaker.**
- **Four Types**
  - Pearson correlation, Kendall rank correlation, Spearman correlation, Point-Biserial correlation.



- Pearson correlation coefficient is a measure of the strength of a linear association between two variables
- denoted by r

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

Where,

r = Pearson Correlation Coefficient

$x_i$  = x variable samples

$\bar{x}$  = mean of values in x variable

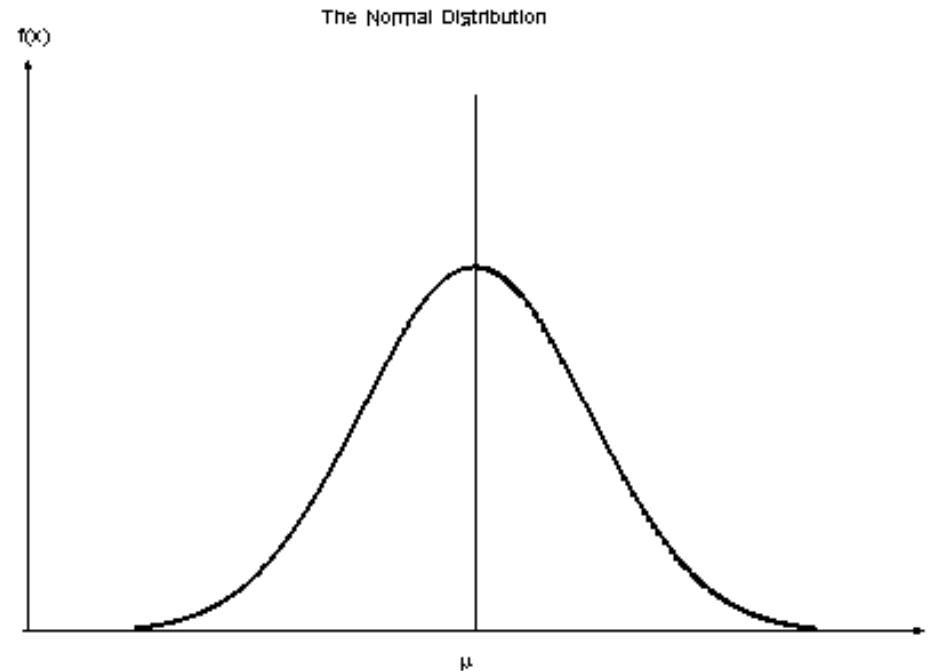
$y_i$  = y variable sample

$\bar{y}$  = mean of values in y variable

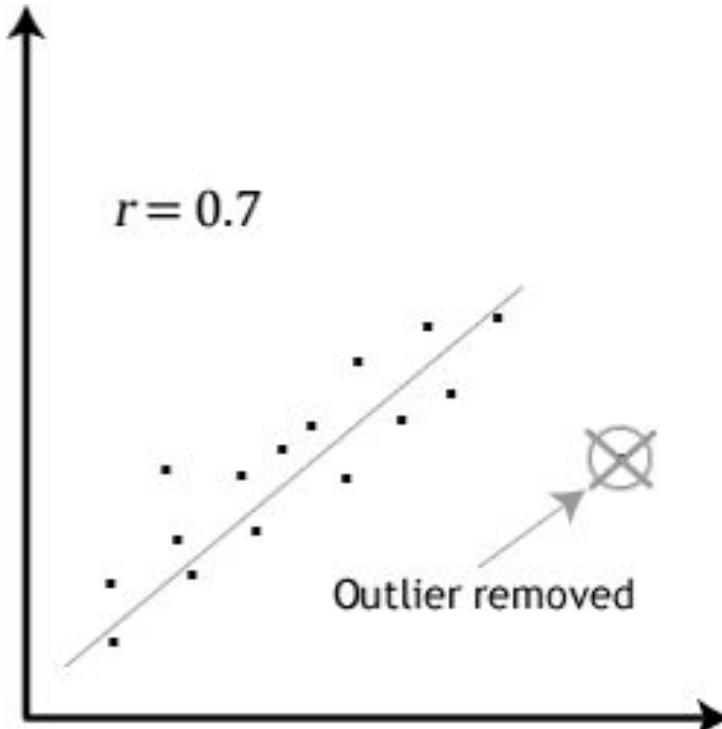
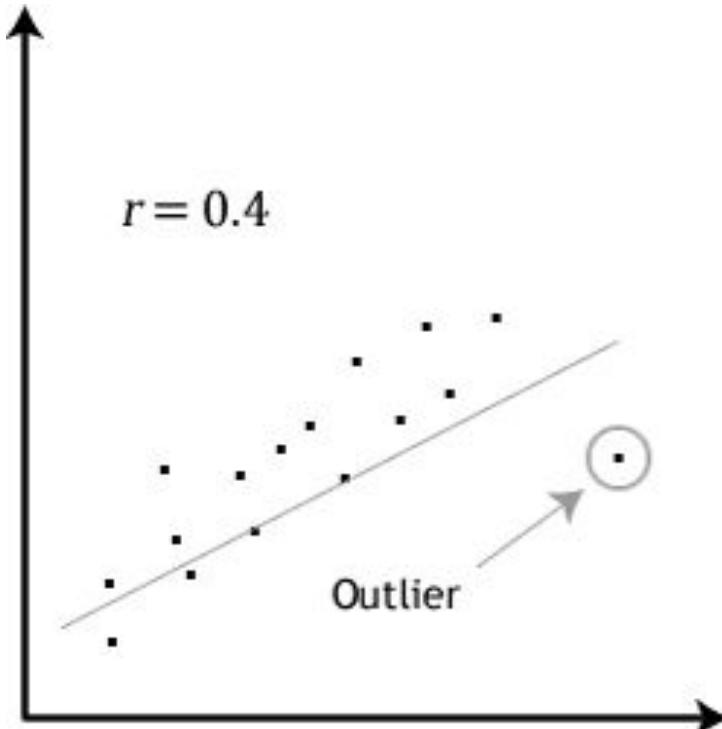
- Is there a statistically **significant relationship between age and height?**
- Is there a relationship **between temperature and ice cream sales?**
- Is there a relationship **among job satisfaction, productivity, and income?**
- Which two variable have the **strongest correlation between age, height, weight, size of family and family income?**

1. Both variables should be normally distributed.

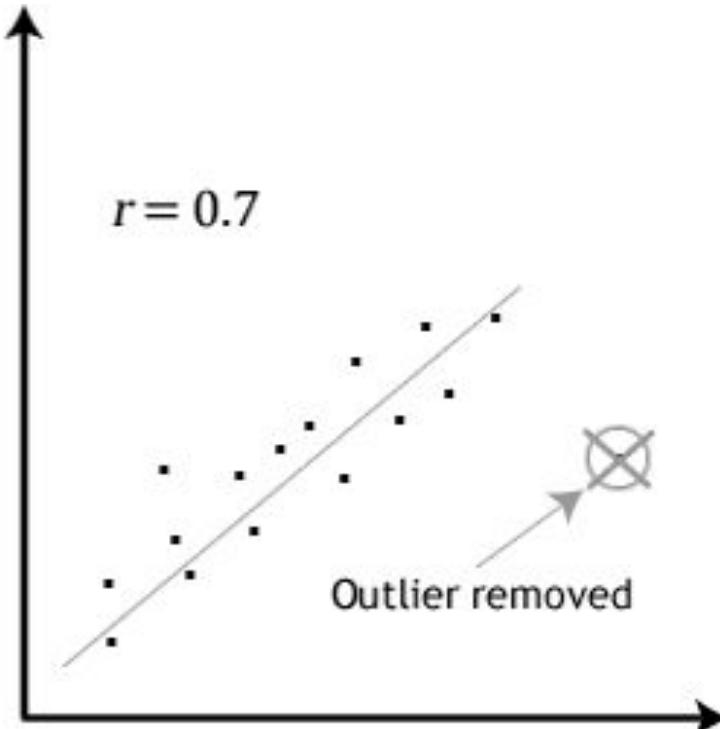
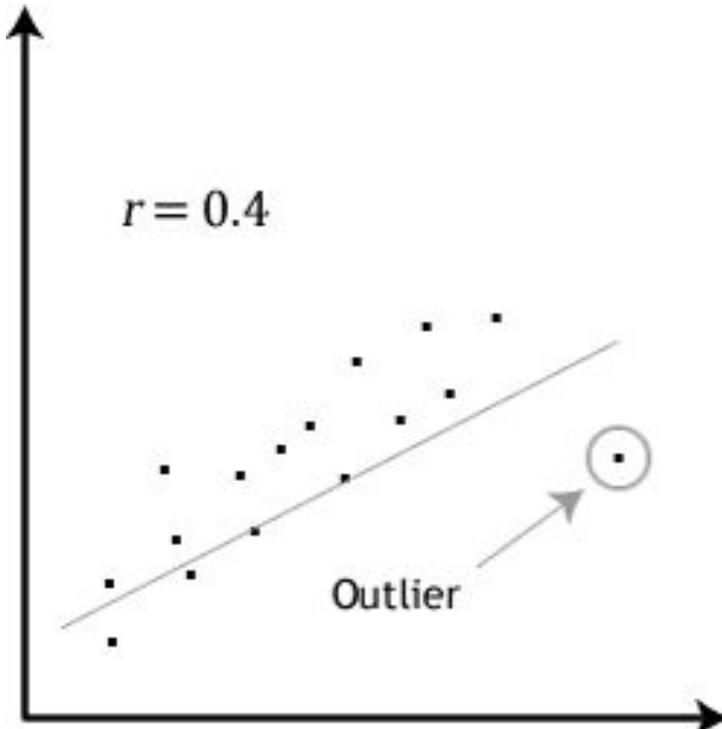
This is sometimes called the ‘Bell Curve’ or the ‘Gaussian Curve’



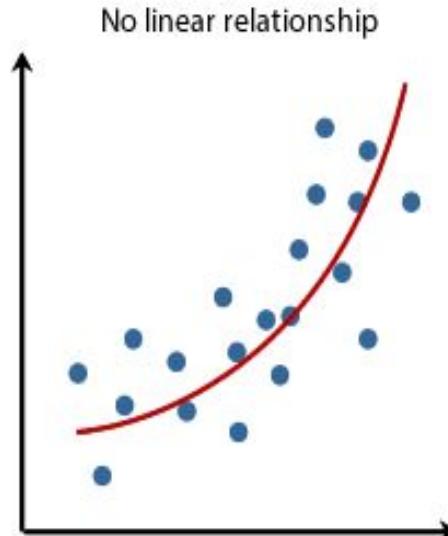
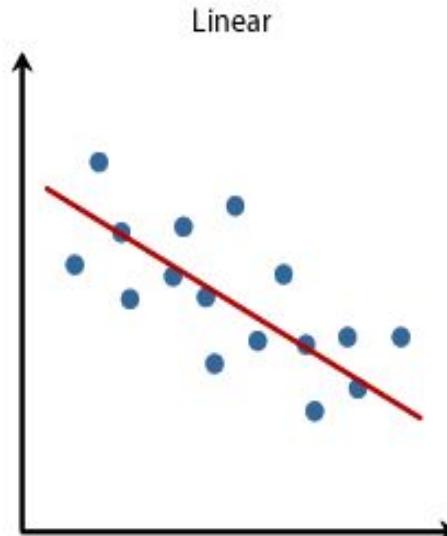
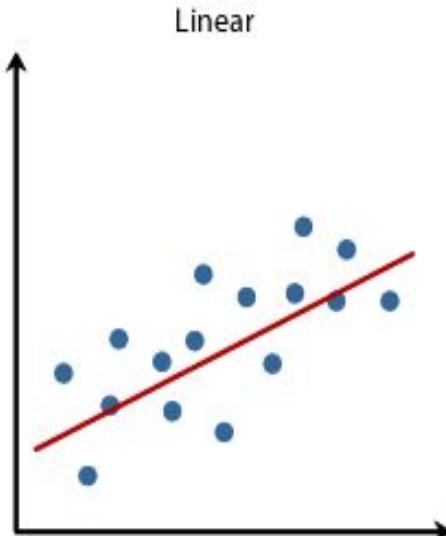
2. There should be no significant outliers



4. There should be no significant outliers



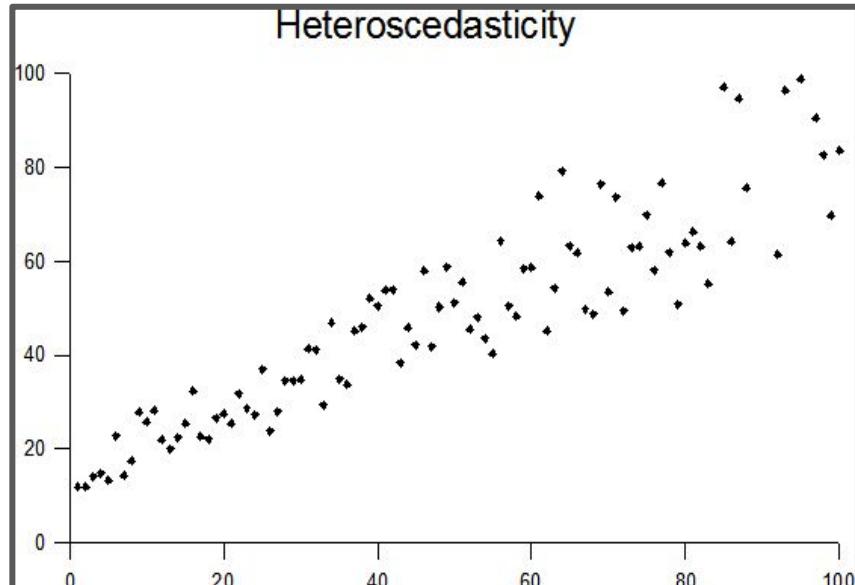
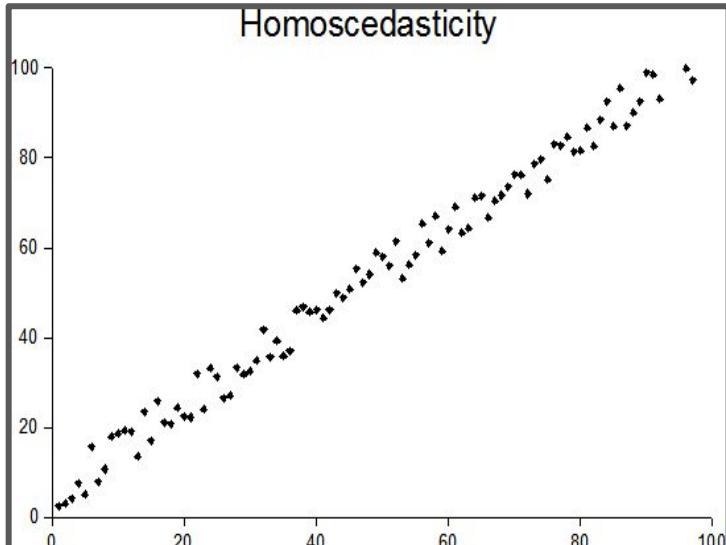
3. Each variable **should be continuous i.e. interval or ratios** for example weight, time, height, age etc
4. The two variables have a **linear relationship**



5. The observations are **paired observations**. That is, for every observation of the **independent variable**, there **must be a corresponding observation of the dependent variable**. For example if you're calculating the correlation between age and weight. If there are 12 observations of weight, you should have 12 observations of age. i.e. no blanks.

6. There should be **Homoscedasticity**, which means the variance around the line of best fit should be similar.

Homoscedasticity describes a situation in which the error term is the same across all values of the independent variables. A scatter-plot makes it easy to check for this. If the points lie equally on both sides of the line of best fit, then the data is homoscedastic.



# Pearson Correlation Example

|    | A                  | B                     | C                     | D             | E                   | F                   | G                                   | H |
|----|--------------------|-----------------------|-----------------------|---------------|---------------------|---------------------|-------------------------------------|---|
| 5  |                    |                       |                       |               |                     |                     |                                     |   |
| 6  | Hours Played Sport | Test Score            | $x - \bar{x}$         | $y - \bar{y}$ | $(x_i - \bar{x})^2$ | $(y_i - \bar{y})^2$ | $(x_i - \bar{x}) * (y_i - \bar{y})$ |   |
| 7  | x                  | y                     |                       |               |                     |                     |                                     |   |
| 8  | 3                  | 74                    | 0.43                  | 1.71          | 0.18                | 2.94                | 0.73                                |   |
| 9  | 1                  | 68                    | -1.57                 | -4.29         | 2.47                | 18.37               | 6.73                                |   |
| 10 | 1                  | 66                    | -1.57                 | -6.29         | 2.47                | 39.51               | 9.88                                |   |
| 11 | 3                  | 72                    | 0.43                  | -0.29         | 0.18                | 0.08                | -0.12                               |   |
| 12 | 4                  | 80                    | 1.43                  | 7.71          | 2.04                | 59.51               | 11.02                               |   |
| 13 | 2                  | 68                    | -0.57                 | -4.29         | 0.33                | 18.37               | 2.45                                |   |
| 14 | 4                  | 78                    | 1.43                  | 5.71          | 2.04                | 32.65               | 8.16                                |   |
| 15 |                    |                       |                       |               |                     |                     |                                     |   |
| 16 |                    | $\bar{x}$ (Mean of x) | $\bar{y}$ (Mean of y) |               |                     |                     |                                     |   |
| 17 | Mean               | 2.57                  | 72.29                 |               |                     |                     |                                     |   |
| 18 |                    |                       |                       |               |                     |                     |                                     |   |

# Pearson Correlation Example

|    | A                    | B            | C                   | D                   | E                                   | F                   | G                                   | H |
|----|----------------------|--------------|---------------------|---------------------|-------------------------------------|---------------------|-------------------------------------|---|
| 5  |                      |              |                     |                     |                                     |                     |                                     |   |
| 6  | Hours Played Sport   | Test Score   | $x - \bar{x}$       | $y - \bar{y}$       | $(x_i - \bar{x})^2$                 | $(y_i - \bar{y})^2$ | $(x_i - \bar{x}) * (y_i - \bar{y})$ |   |
| 7  | x                    | y            |                     |                     |                                     |                     |                                     |   |
| 8  | 3                    | 74           | 0.43                | 1.71                | 0.18                                | 2.94                | 0.73                                |   |
| 9  | 1                    | 68           | -1.57               | -4.29               | 2.47                                | 18.37               | 6.73                                |   |
| 10 | 1                    | 66           | -1.57               | -6.29               | 2.47                                | 39.51               | 9.88                                |   |
| 11 | 3                    | 72           | 0.43                | -0.29               | 0.18                                | 0.08                | -0.12                               |   |
| 12 | 4                    | 80           | 1.43                | 7.71                | 2.04                                | 59.51               | 11.02                               |   |
| 13 | 2                    | 68           | -0.57               | -4.29               | 0.33                                | 18.37               | 2.45                                |   |
| 14 | 4                    | 78           | 1.43                | 5.71                | 2.04                                | 32.65               | 8.16                                |   |
| 15 |                      |              |                     |                     |                                     |                     |                                     |   |
| 19 | Sum is calculated as |              |                     |                     |                                     |                     |                                     |   |
| 20 |                      |              |                     |                     |                                     |                     |                                     |   |
| 21 |                      |              | $(x_i - \bar{x})^2$ | $(y_i - \bar{y})^2$ | $(x_i - \bar{x}) * (y_i - \bar{y})$ |                     |                                     |   |
| 22 | Formula              | =SUM(E8:E14) | =SUM(F8:F14)        | =SUM(G8:G14)        |                                     |                     |                                     |   |
| 23 | Sum                  | 9.71         | 171.43              | 38.86               |                                     |                     |                                     |   |
| 24 |                      |              |                     |                     |                                     |                     |                                     |   |

|    | A                                   | B                   | C                   | D                                   | E |
|----|-------------------------------------|---------------------|---------------------|-------------------------------------|---|
| 20 |                                     |                     |                     |                                     |   |
| 21 |                                     | $(x_i - \bar{x})^2$ | $(y_i - \bar{y})^2$ | $(x_i - \bar{x}) * (y_i - \bar{y})$ |   |
| 22 | Sum                                 | 9.71                | 171.43              | 38.86                               |   |
| 23 |                                     |                     |                     |                                     |   |
| 24 | Standard Deviation is calculated as |                     |                     |                                     |   |
| 25 |                                     |                     |                     |                                     |   |
| 26 |                                     | $\sigma_x$          | $\sigma_y$          |                                     |   |
| 27 | Formula                             | =SQRT(B22)          | =SQRT(C22)          |                                     |   |
| 28 | Standard Deviation                  | 3.12                | 13.09               |                                     |   |
| 29 |                                     |                     |                     |                                     |   |

|    | A   | B                   | C                   | D                                   | E |
|----|---|---------------------|---------------------|-------------------------------------|---|
| 20 |   |                     |                     |                                     |   |
| 21 |   | $(x_i - \bar{x})^2$ | $(y_i - \bar{y})^2$ | $(x_i - \bar{x}) * (y_i - \bar{y})$ |   |
| 22 | Sum   | 9.71                | 171.43              | 38.86                               |   |
| 25 |   |                     |                     |                                     |   |
| 26 |   | $\sigma_x$          | $\sigma_y$          |                                     |   |
| 27 | Standard Deviation  | 3.12                | 13.09               |                                     |   |
| 28 |   |                     |                     |                                     |   |
| 29 | Pearson Correlation Coefficient is calculated using the formula given below   |                     |                     |                                     |   |
| 30 | <b>Pearson Correlation Coefficient = <math>\rho(x,y) = \Sigma[(x_i - \bar{x}) * (y_i - \bar{y})] / (\sigma_x * \sigma_y)</math></b> |                     |                     |                                     |   |
| 31 |   |                     |                     |                                     |   |
| 32 | Pearson Correlation<br>Coefficient Formula  | =D22/(B27*<br>C27)  |                     |                                     |   |
| 33 | Pearson Correlation<br>Coefficient  | 0.95                |                     |                                     |   |
| 34 |   |                     |                     |                                     |   |

- Pearson Correlation Coefficient =  $38.86 / (3.12 * 13.09)$
- Pearson Correlation Coefficient = 0.95

We have an output of 0.95; this indicates that when the number of hours played to increase, the test scores also increase. These two variables are positively correlated.

<https://sonalake.com/latest/an-introduction-to-hypothesis-testing/>

<https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/>

<https://towardsdatascience.com/hypothesis-testing-p-value-13b55f4b32d9>

<https://www.analyticsvidhya.com/blog/2020/12/quick-guide-to-perform-hypothesis-testing/>

<https://www.simplilearn.com/tutorials/statistics-tutorial/hypothesis-testing-in-statistics>

<https://www.youtube.com/watch?v=kx-pcQAPvoc>

<https://www.analyticsvidhya.com/blog/2021/01/an-introduction-to-hypothesis-testing/>

<https://www.analyticsvidhya.com/blog/2020/06/statistics-analytics-hypothesis-testing-z-test-t-test/>

<https://analyticsindiamag.com/importance-of-hypothesis-testing-in-data-science/>

[https://wwwjmp.com/en\\_be/statistics-knowledge-portal/chi-square-test/chi-square-goodness-of-fit-test.html](https://wwwjmp.com/en_be/statistics-knowledge-portal/chi-square-test/chi-square-goodness-of-fit-test.html)