Data Science & Big Data Analytics

Subject Code: 310251

T. E. Computer (2019 Pattern)

UNIT II

Dr. S.R.Khonde

UNIT II

Unit II	Statistical Inference	07 Hours
Need of statistics in Data Science and 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.		
#Exemplar/Case Studies	For an employee dataset, create measure of central tendency and its measure of dispersion for statistical analysis of given data.	
***Mapping of Course Outcomes for Unit II**	CO2	

COURSE OUTCOME

Apply statistics for Big Data Analytics

OUTLINE

- Need of statistics in Data Science & Big Data Analytics
- Measures of Central Tendency
- Measures of Dispersion

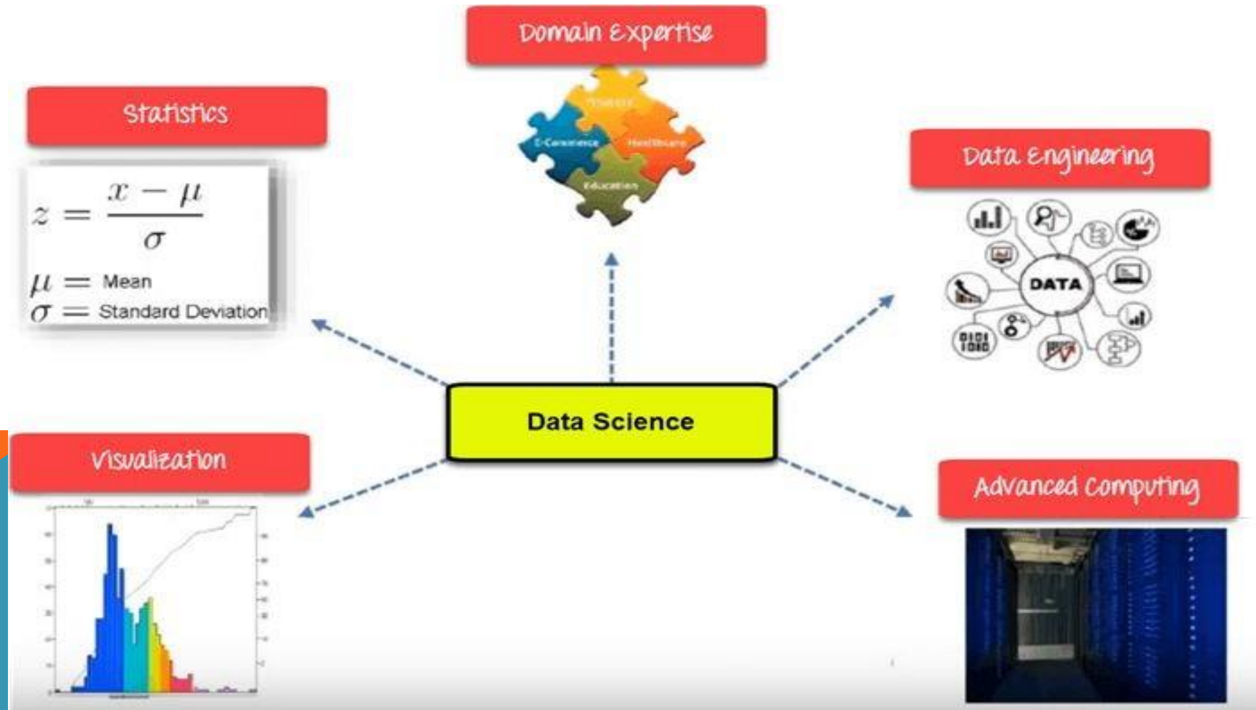
OUTLINE

- **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



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

In

large quantities

To get

useful insights

NEED OF STATISTICS

Need of statistics in Data Science & Big Data Analytics

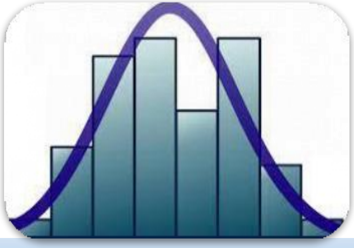
1

Identify

the

Importance

of



various statistical tests.

By using

Relevant Features

Dr. S.R.Khonde

NEED OF STATISTICS

Need of statistics in Data Science & Big Data Analytics

#	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	...

2

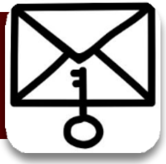
Finding the Relationships between Features

To eliminate Duplicate Features.

NEED OF STATISTICS

Need of statistics in Data Scienc & Big Data Analytics

3



Convert the Features into Required Format

NEED OF STATISTICS

Need of statistics in Data Science & Big Data Analytics

4

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

Dr. S.R.Khonde

NEED OF STATISTICS

Need of statistics in Data Science & Big Data Analytics

6

After

Processing the data

Identify the

Right mathematical approach/model

NEED OF STATISTICS

Need of statistics in Data Science & Big Data Analytics

7

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**

Attributes

Objects

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

NEED OF STATISTICS

Know your Data

Dimension	• Data Warehousing
Feature	• Machine learning
Variable	• Statisticians
Attribute	• Data mining and database professional

NEED OF STATISTICS

Attribute Values

- Attribute values are **numbers or symbols assigned to an attribute**

Heights (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

NEED OF STATISTICS

Attribute Values

- Attribute values **are numbers or symbols assigned to an attribute**

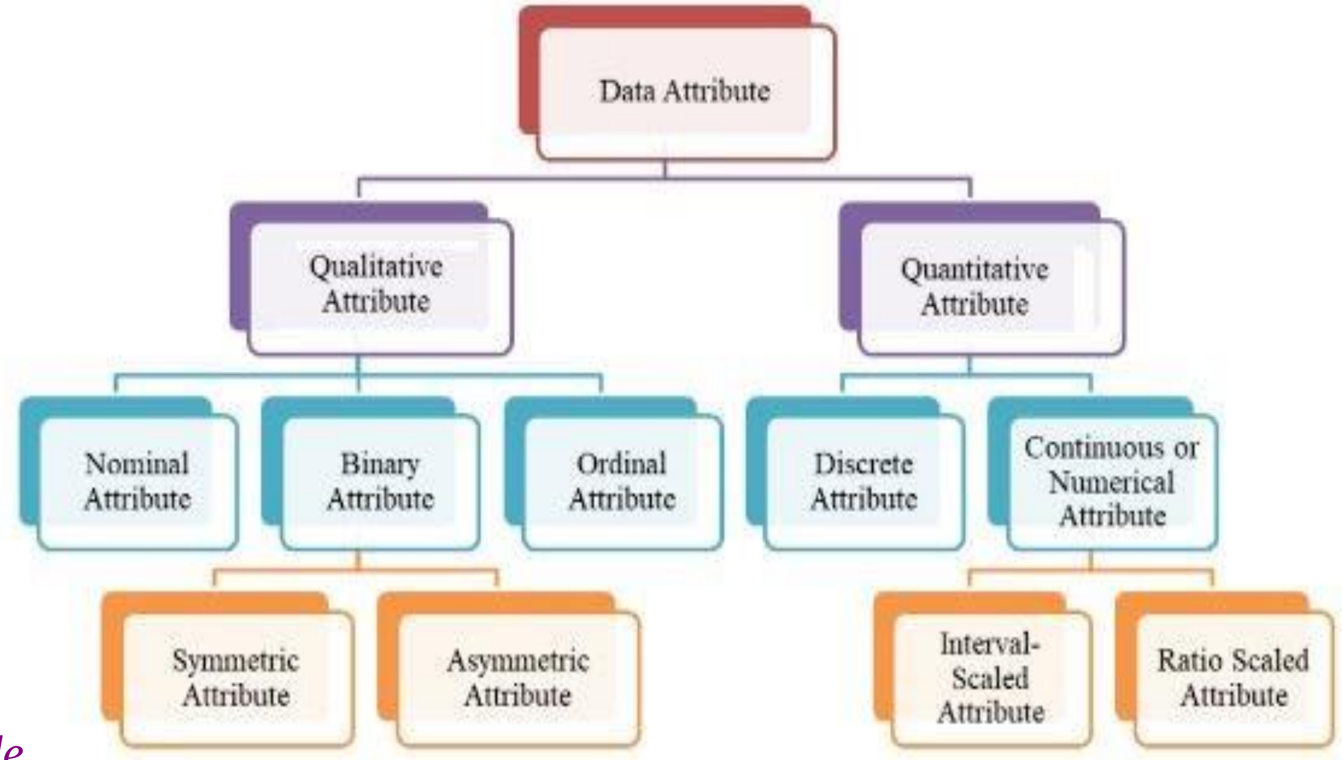
Height	Hair	Eyes	CLASS
short	blonde	blue	$\oplus$
short	dark	blue	$\ominus$
tall	dark	brown	$\ominus$
tall	blonde	brown	$\ominus$
tall	dark	blue	$\ominus$
short	blonde	brown	$\ominus$
tall	red	blue	$\oplus$
tall	blonde	blue	$\oplus$

Multivariate Data

NEED OF STATISTICS

Measures of Central Tendency

Types of Data Attribute



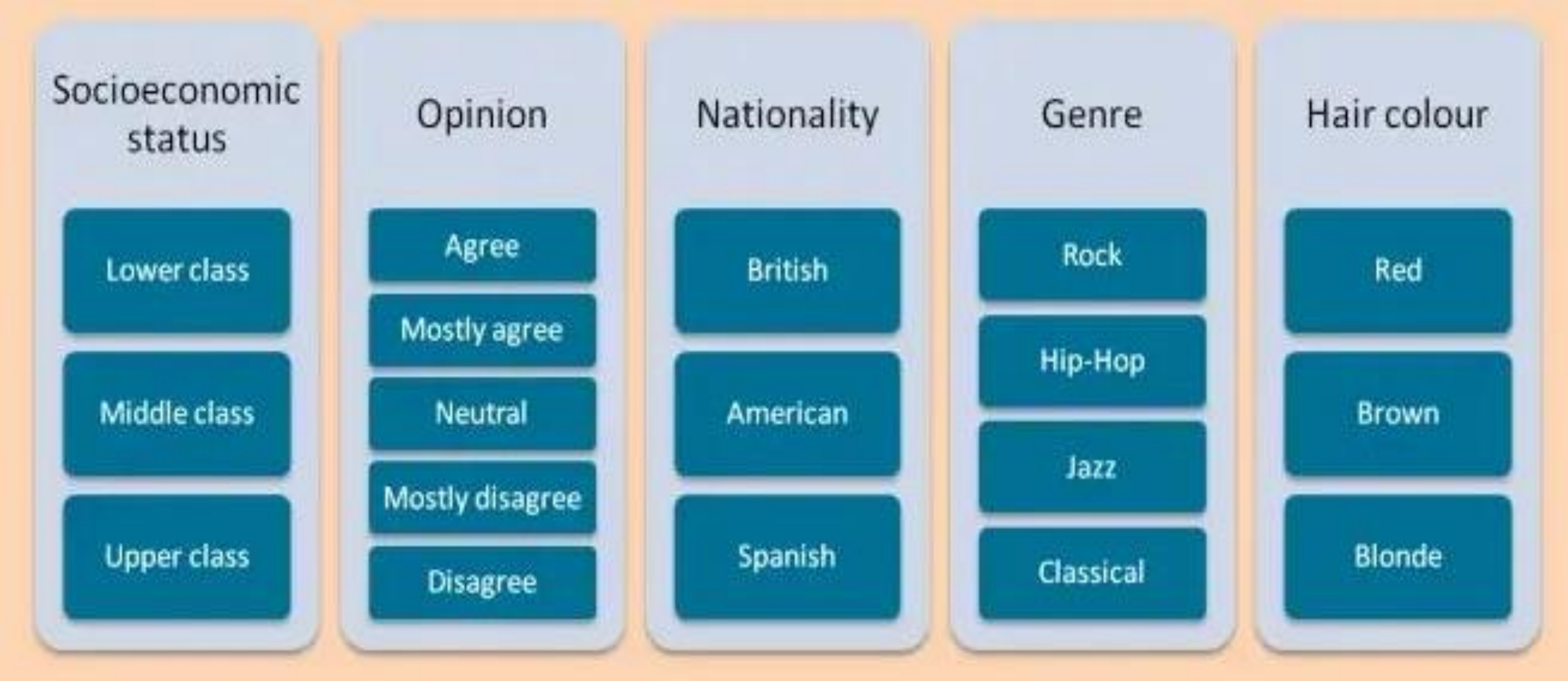
DATA ATTRIBUTE

- 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



DATA ATTRIBUTE

Quantitative Attributes

- attributes **can be measured and assigned in 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

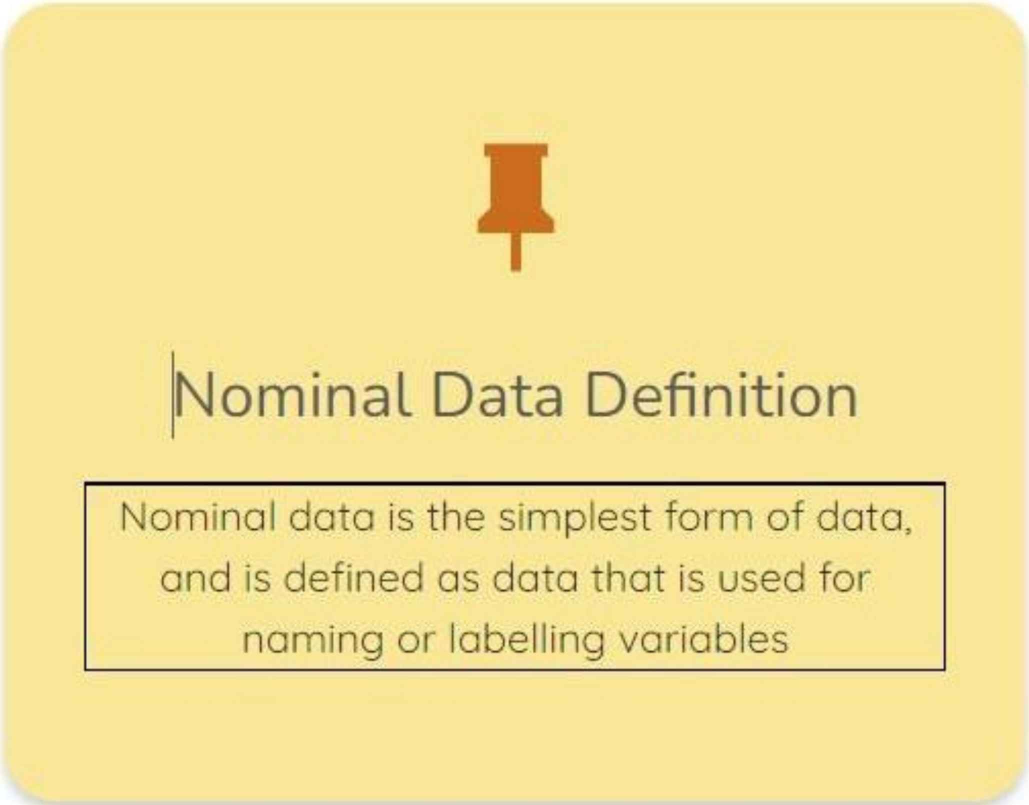
DATA ATTRIBUTE

Qualitative Attributes

1 **Nominal**

Qualitative Attributes

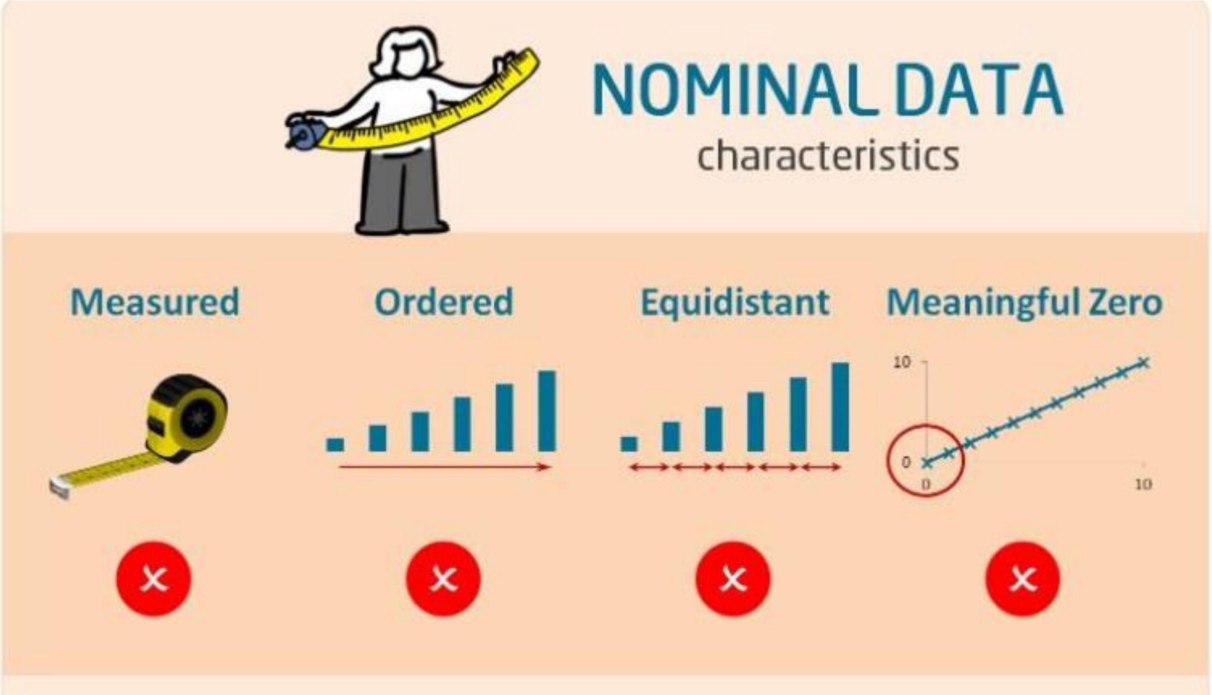
1. **Nominal**



DATA ATTRIBUTE

Qualitative Attributes

1 Nominal



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
Independent	Democrat	Democrat

Simple Frequency Distribution

Percentage Frequency Distribution

DATA ATTRIBUTE

Qualitative Attributes

References:
https://www.scribbr.com/statistics/nominal-data/

1 Nominal Data Analysis

Simple Frequency Distribution

Political preference	Frequency
Democrat	13
Republican	9
Independent	5

Percentage Frequency Distribution

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

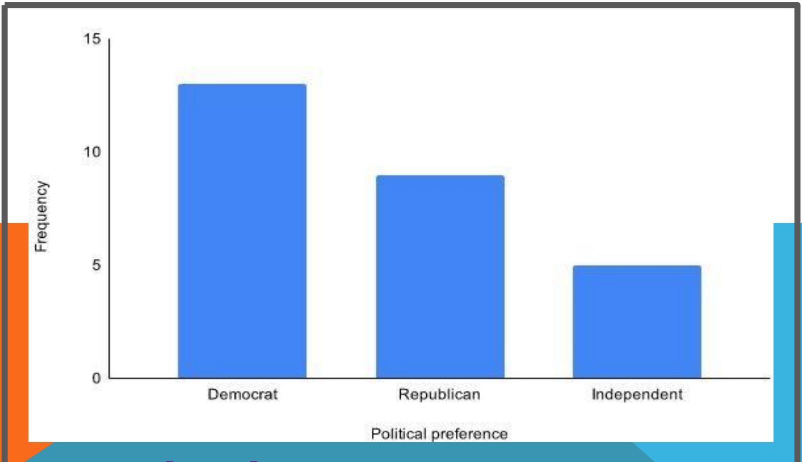
DATA ATTRIBUTE

Qualitative Attributes

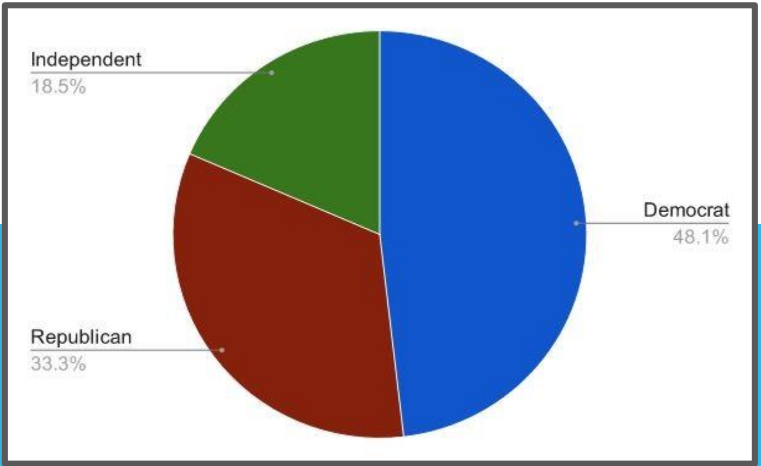
References:
https://www.scribbr.com/statistics/nominal-data/

1 Nominal Data Analysis

Bar Chart



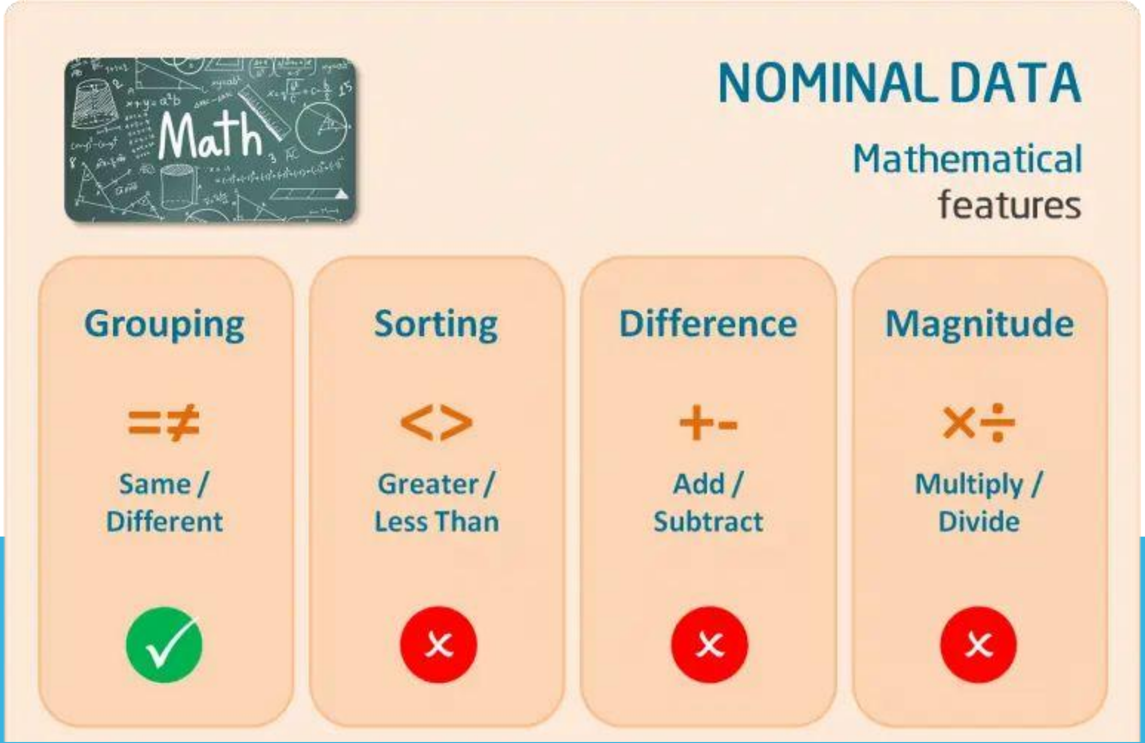
Pie Chart



DATA ATTRIBUTE

Qualitative Attributes

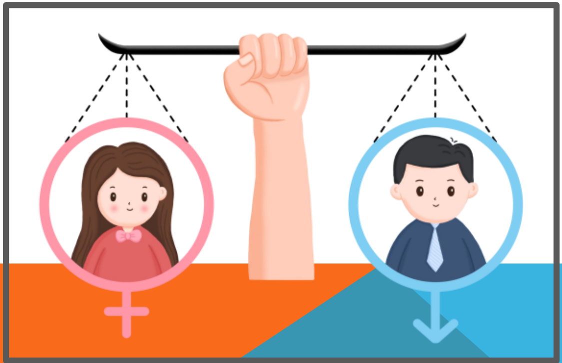
1 Nominal

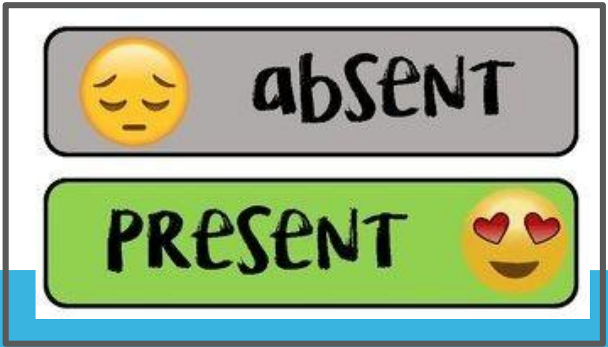


Qualitative Attributes

2 Binary

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





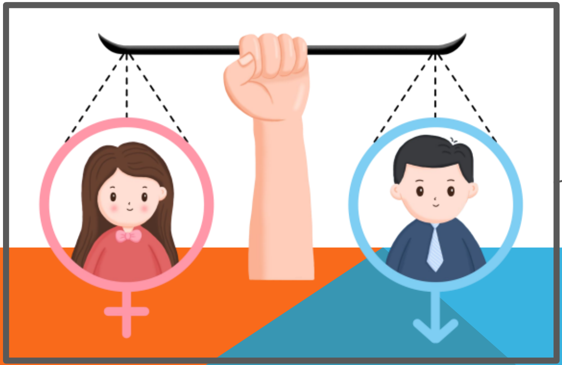
DATA ATTRIBUTE

Qualitative Attributes

2 Binary

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

Symmetric Binary



Equal Important

1 Female
1 Male

1 Female
0 Male

DATA ATTRIBUTE

Qualitative Attributes

2 Binary

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

Asymmetric Binary

Not Equal Important

Most Important With 1

Other with 0

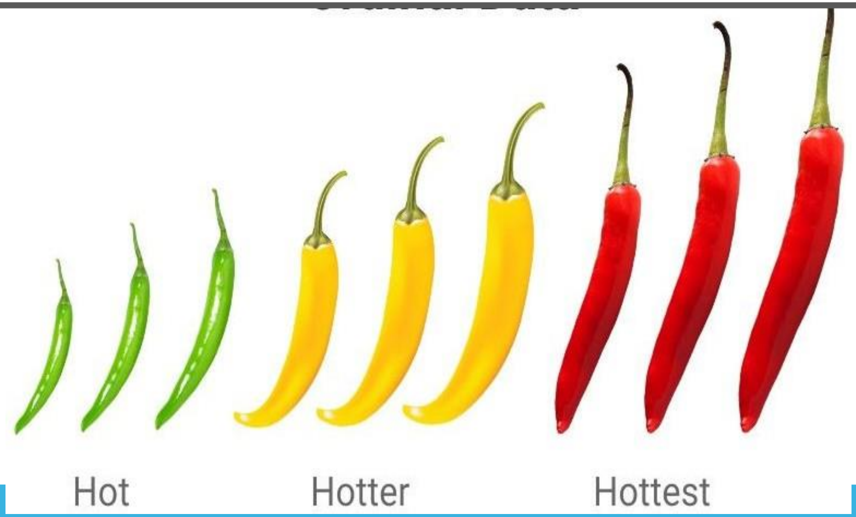
Absent 1
Present 0

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

DATA ATTRIBUTE

Qualitative Attributes

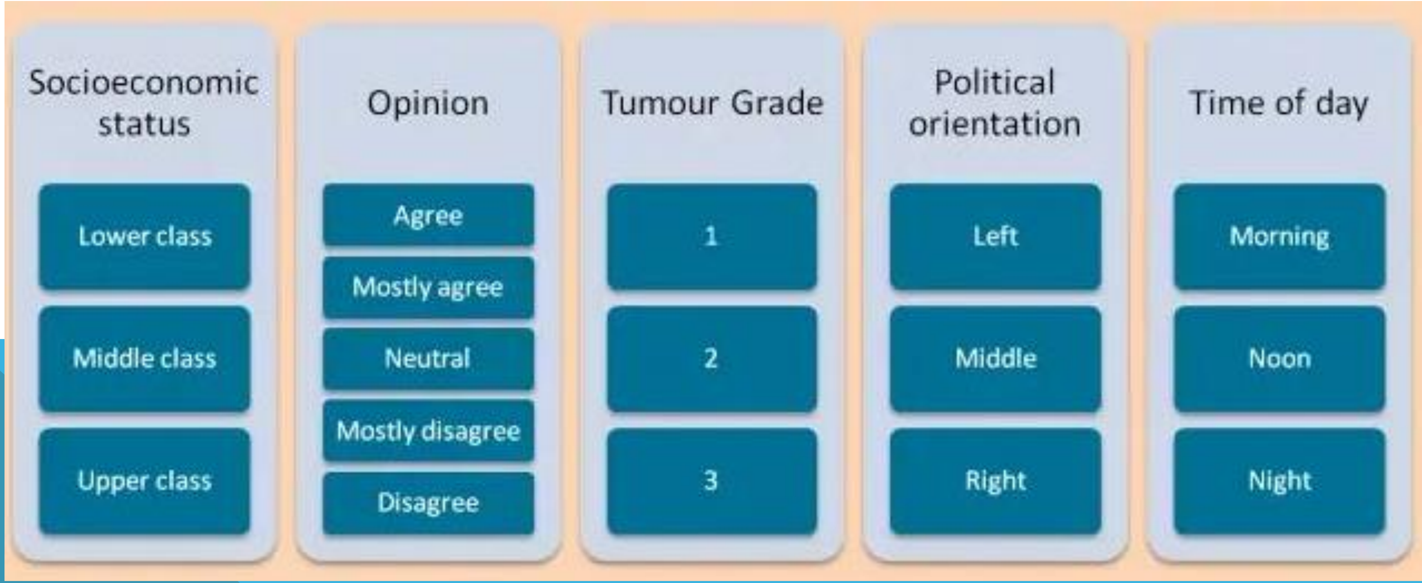
3 Ordinal



DATA ATTRIBUTE

Qualitative Attributes

3 **Ordinal**



DATA ATTRIBUTE

Qualitative Attributes

3 **Ordinal**

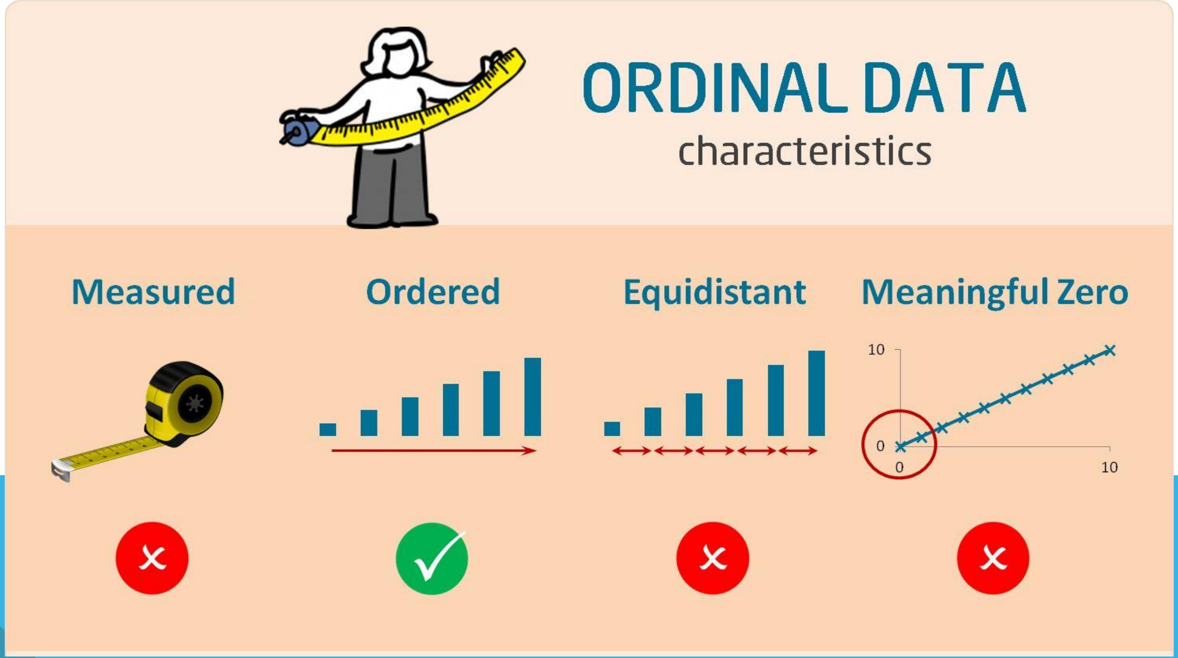
Ordinal Data Definition

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

DATA ATTRIBUTE

Qualitative Attributes

3 **Ordinal**



DATA ATTRIBUTE

Qualitative Attributes

3 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

3 Ordinal Data Analysis

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

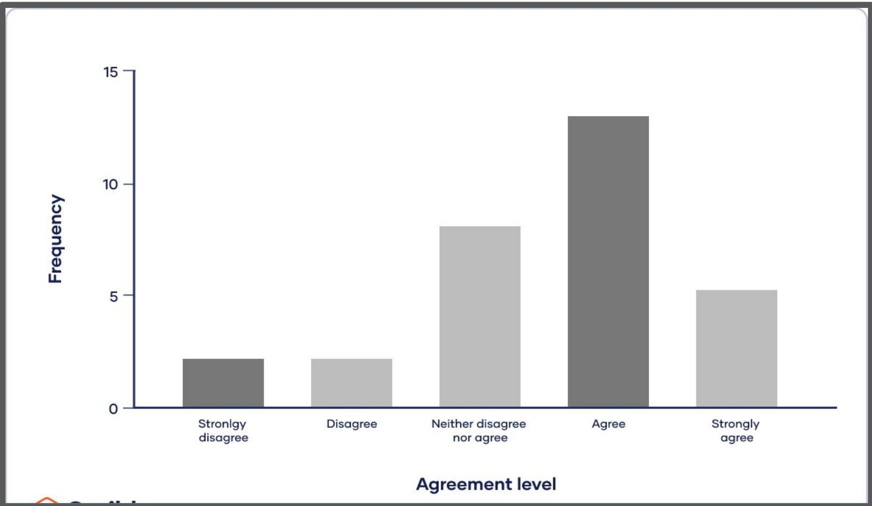
3 Ordinal Data Visualization

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

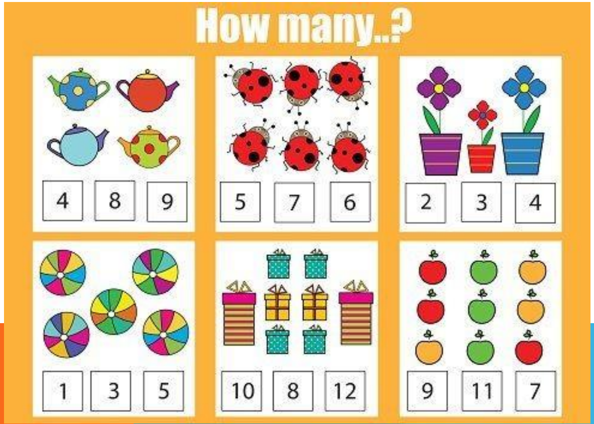
Qualitative Attributes

3 Ordinal

DATA ATTRIBUTE

Quantitative Attributes

1 Discrete



Discrete data is counted

2 Continues/Numeric

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

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

Continuous data is measured

DATA ATTRIBUTE

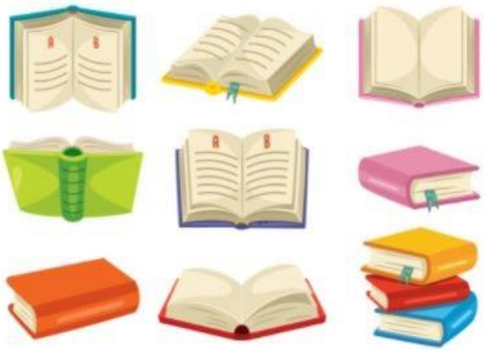
Quantitative Attributes

1 Discrete

Number of books in a bookshelf

2 Continues/Numeric

Length of pages of books present in a bookshelf



DATA ATTRIBUTE

Quantitative Attributes

1 Discrete

Number of students present in a class

2 Continues/Numeric

Weight of each student in a class

Dr. S.R.Khonde

DATA ATTRIBUTE

Quantitative Attributes

Check your Knowledge

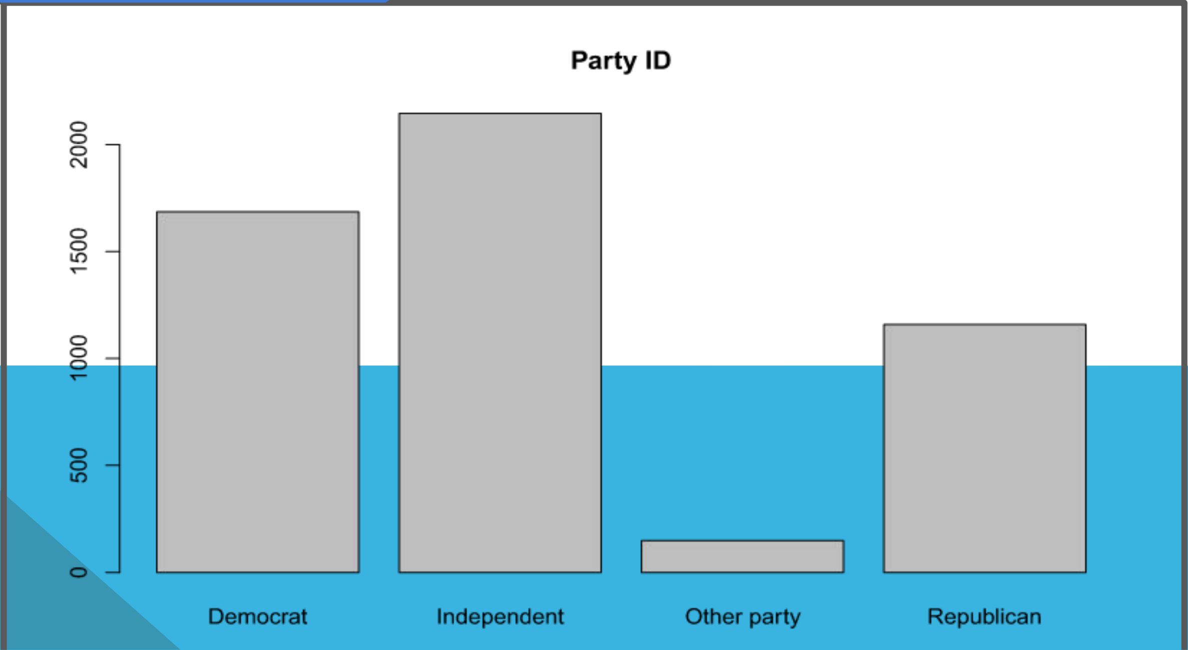
Temperature in a city on different days	Continues
Number of people travel in trains on different days of the week	Discrete
Sum of numbers on rolling three dice together.	Discrete
Volume of water in a water tank	Continues

DATA ATTRIBUTE

Quantitative Attributes

1 Discrete Data Analysis

Bar Graph

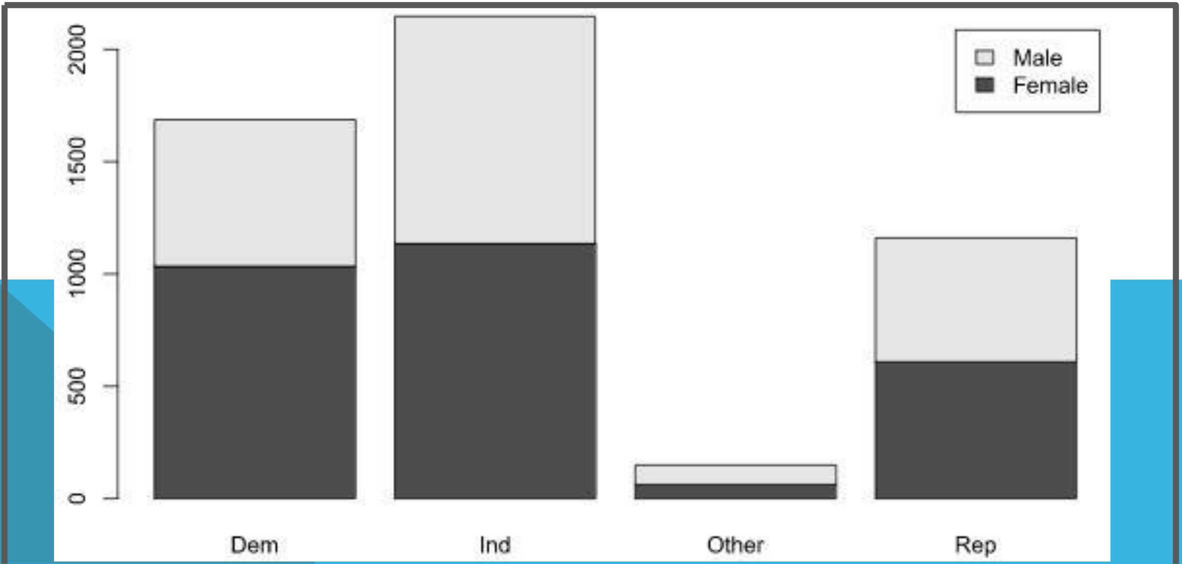


DATA ATTRIBUTE

Quantitative Attributes

1 Discrete Data Analysis

Bar Graph



Data Attribute

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

DATA ATTRIBUTE

Quantitative Attributes

2.1

Continues/Numeric : Interval

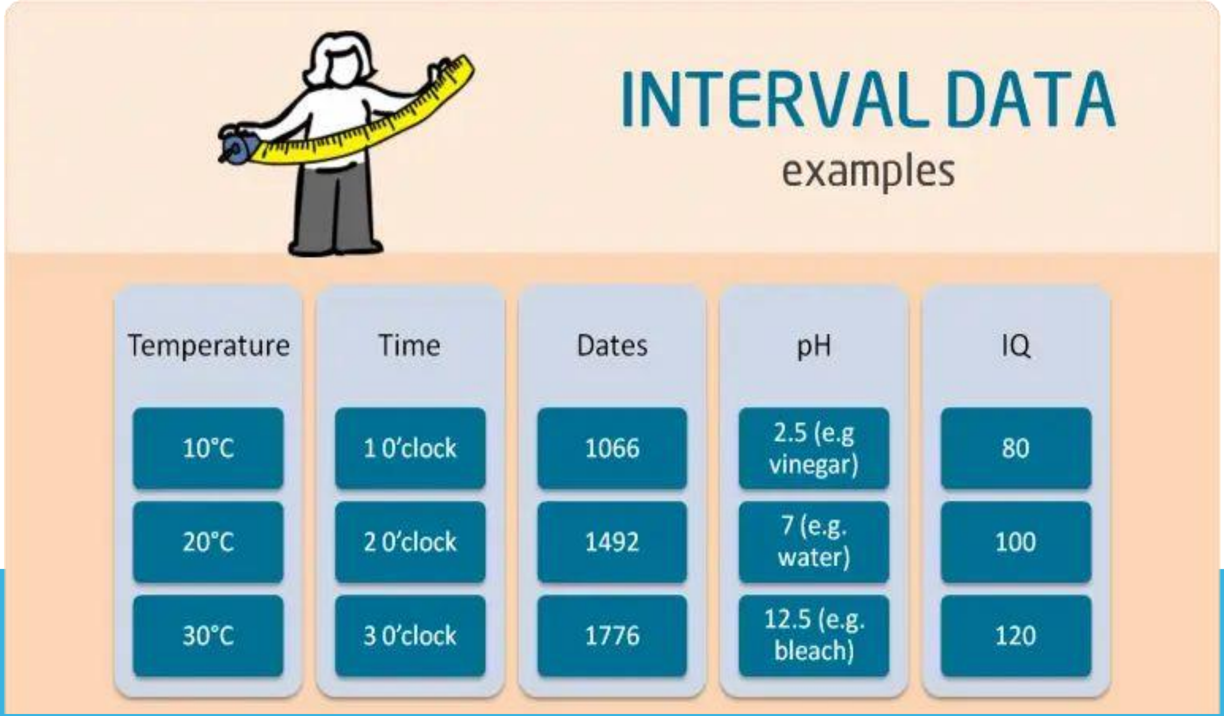
Interval Data Definition

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

DATA ATTRIBUTE

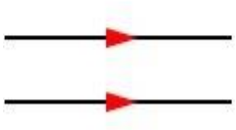
Quantitative Attributes

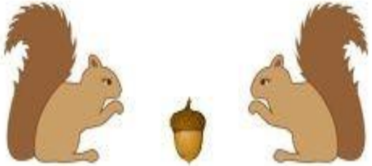
2.1 Continues/Numeric : Interval

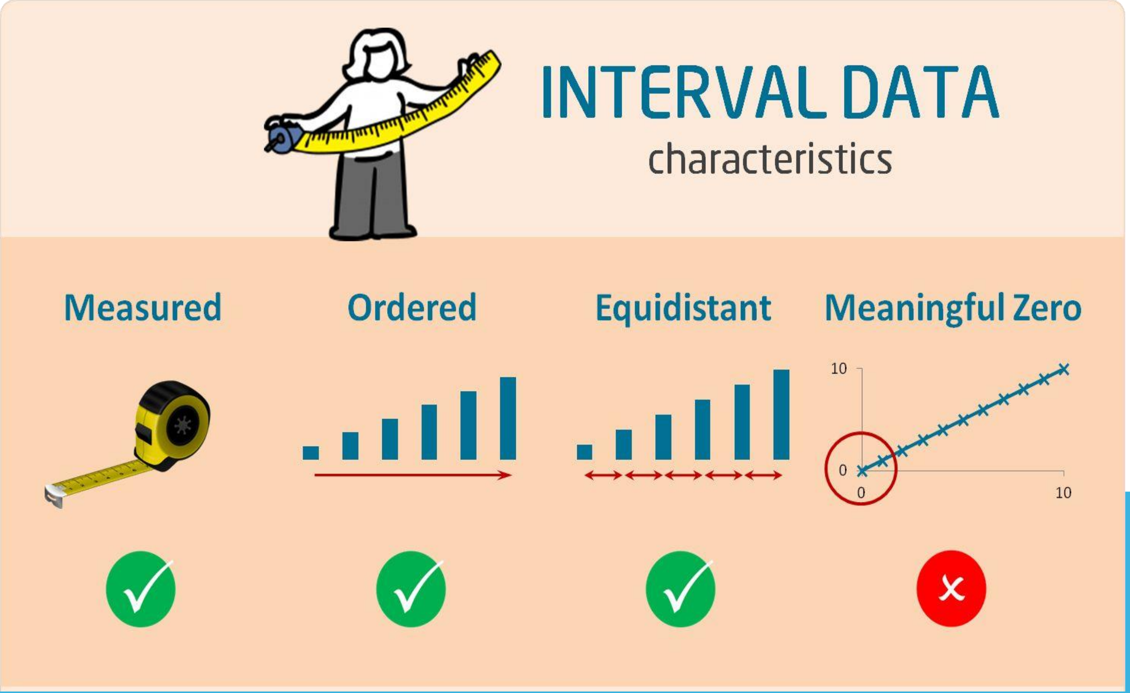


Quantitative Attributes

2.1 Continues/Numeric : Interval







Equidistance

Dr. S.R.Khonde

DATA ATTRIBUTE

Quantitative Attributes

2.1

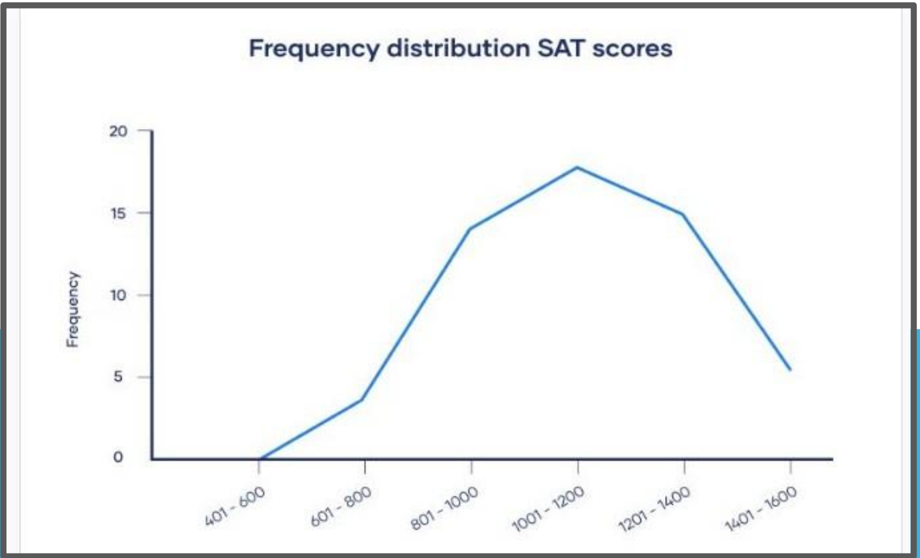
Continues/Numeric

: 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

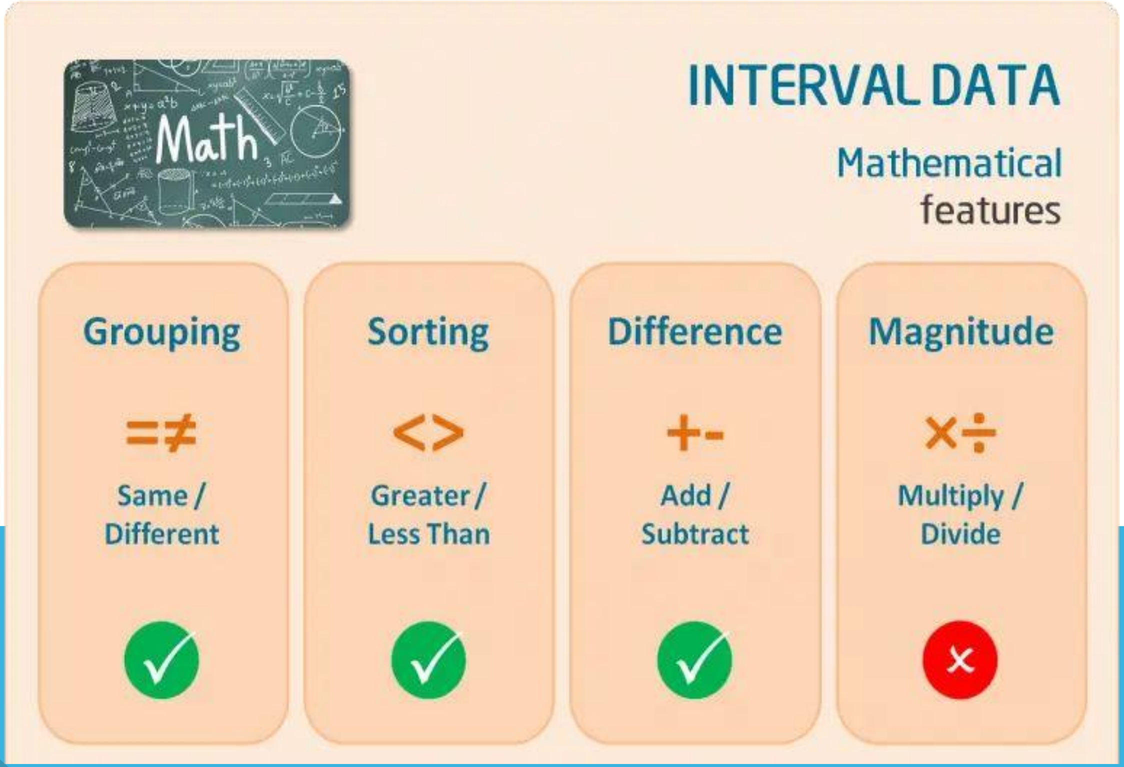


Quantitative Attributes

2.1

Continues/Numeric

: Interval



Quantitative Attributes

2.2 Continues/Numeric:

Ratio Scaled

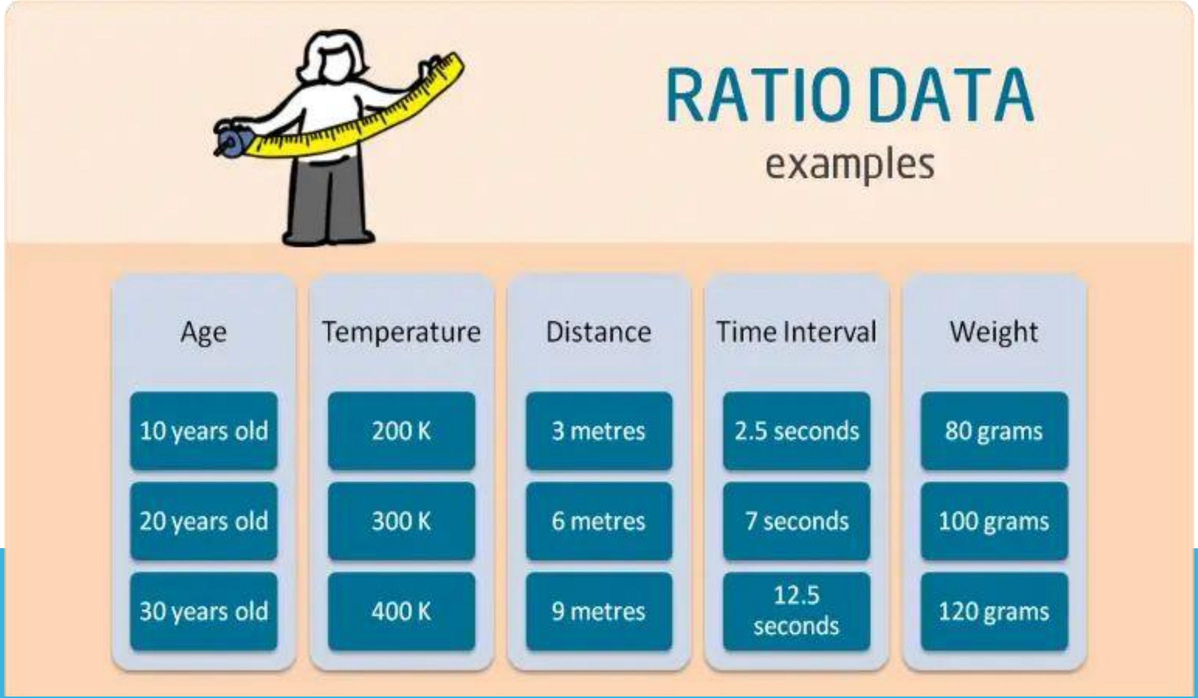
Ratio Data Definition

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

DATA ATTRIBUTE

Quantitative Attributes

2.2 Continues/Numeric : Interval



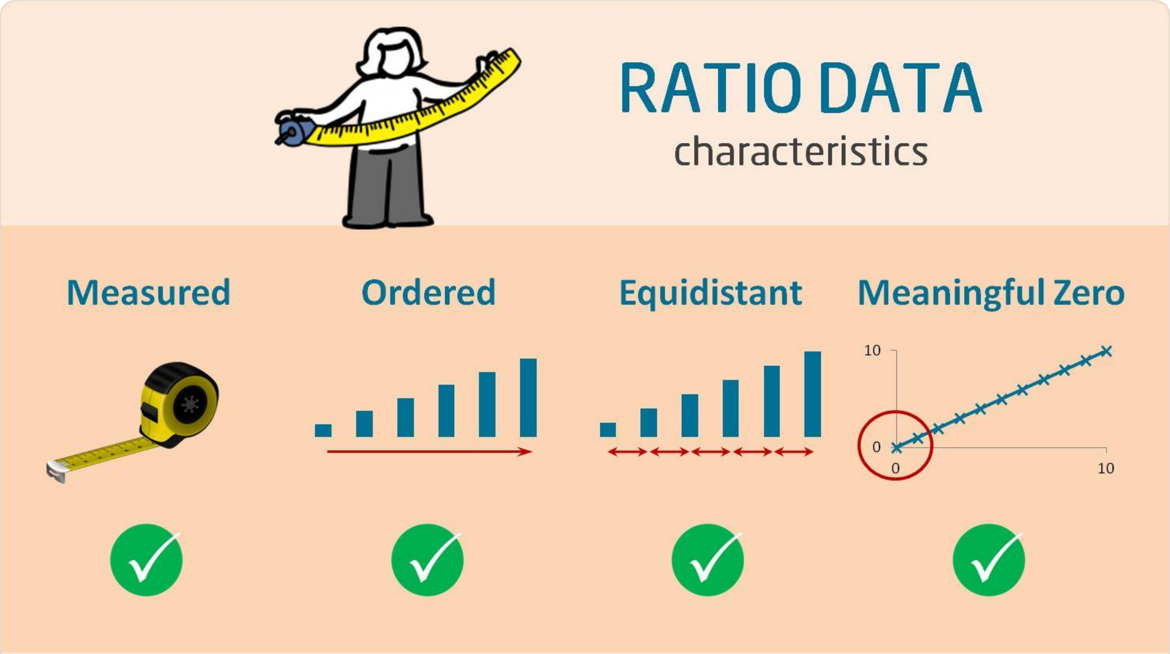
DATA ATTRIBUTE

Quantitative Attributes

2.2

Continues/Numeric:

Ratio Scaled

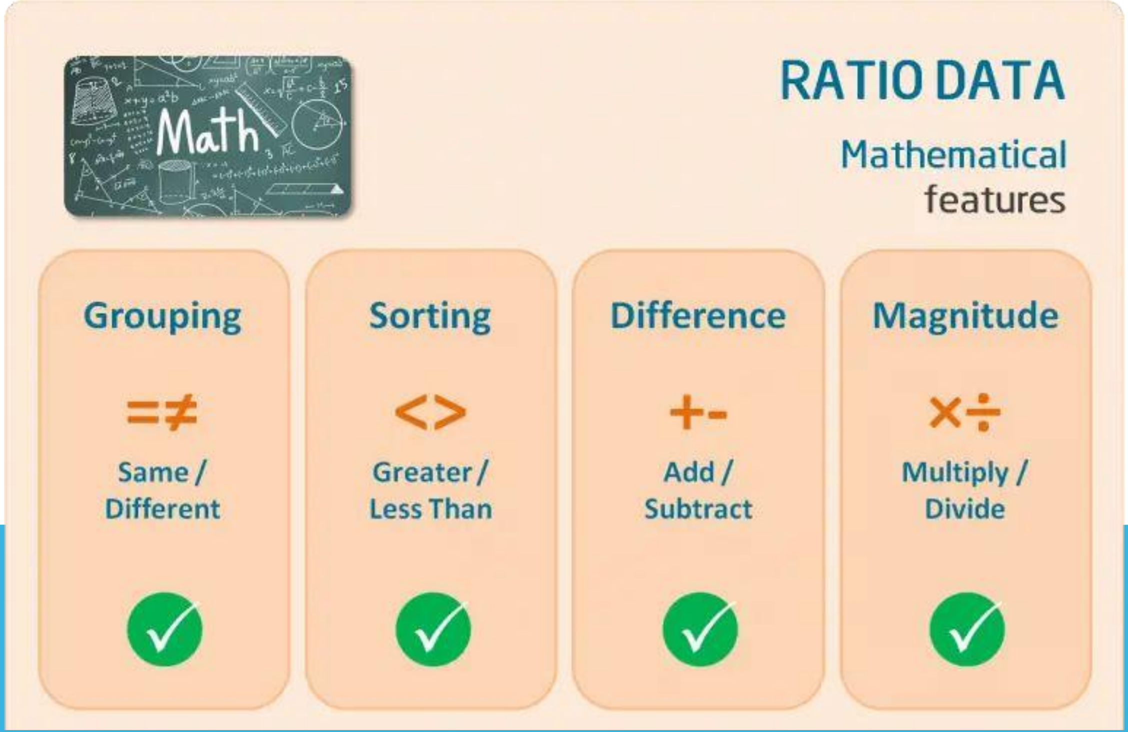


Quantitative Attributes

2.2

Continues/Numeric:

Ratio Scaled



Quantitative Attributes

Interval Vs Ratio

I am type of Interval

I am type of Ratio

10 $ less Than Rajan

66% greater than Poojan

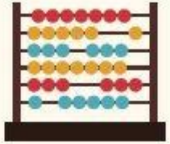
20 $ /hour

30 $ /hour

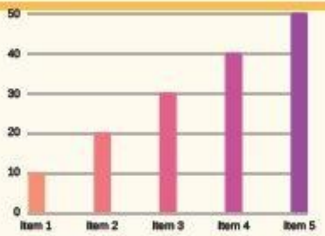
Poojan

Rajan

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

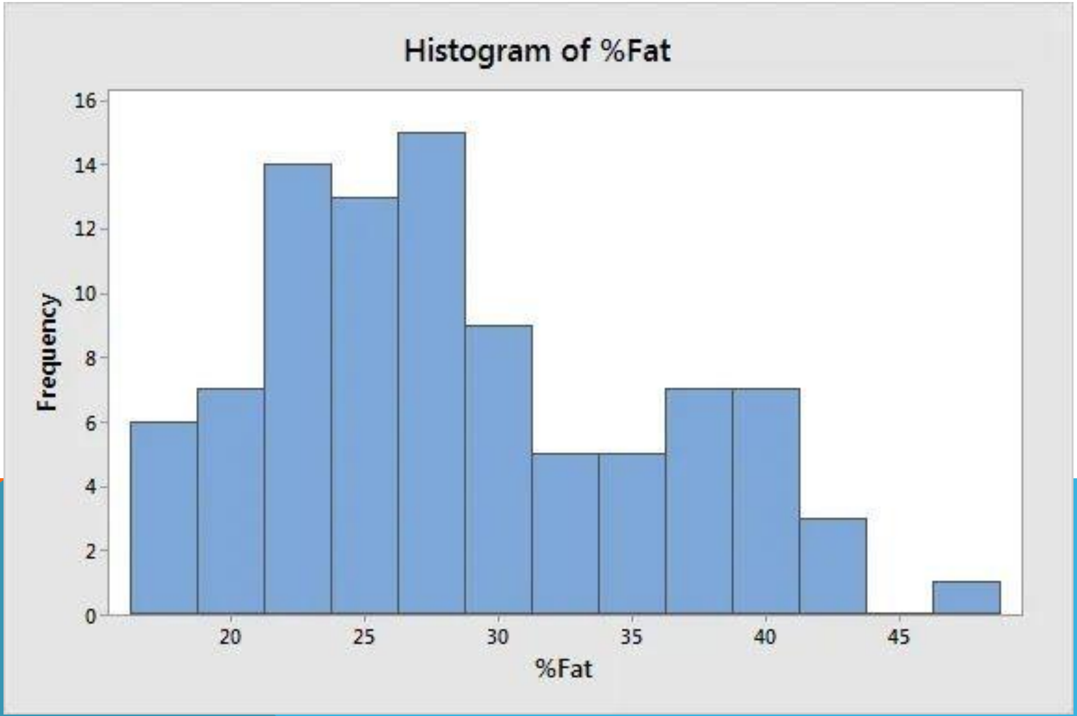
Points	Discrete Data	Continuous Data	
Values	It has a finite number of possible values. The values cannot be divided into smaller pieces and add additional meaning.	It has an infinite number of possible values within an interval. The values can be subdivided into smaller and smaller pieces.	
Graphical Representation	Bar Chart	Histogram	

KEY POINTS TO FOCUS

Points	Discrete Data	Continuous Data	
Examples	• The number of students in a class. • The number of workers in a company. • The number of parts damaged during transportation. • Shoe sizes. • Number of languages an individual speaks. • The number of home runs in a baseball game. • The number of test questions you answered correctly.	• The amount of time required to complete a project. • The height of children. • The amount of time it takes to sell shoes. • The amount of rain, in inches, that falls in a storm. • The square footage of a two-bedroom house. • The weight of a truck. • The speed of cars. • Time to wake up.	 

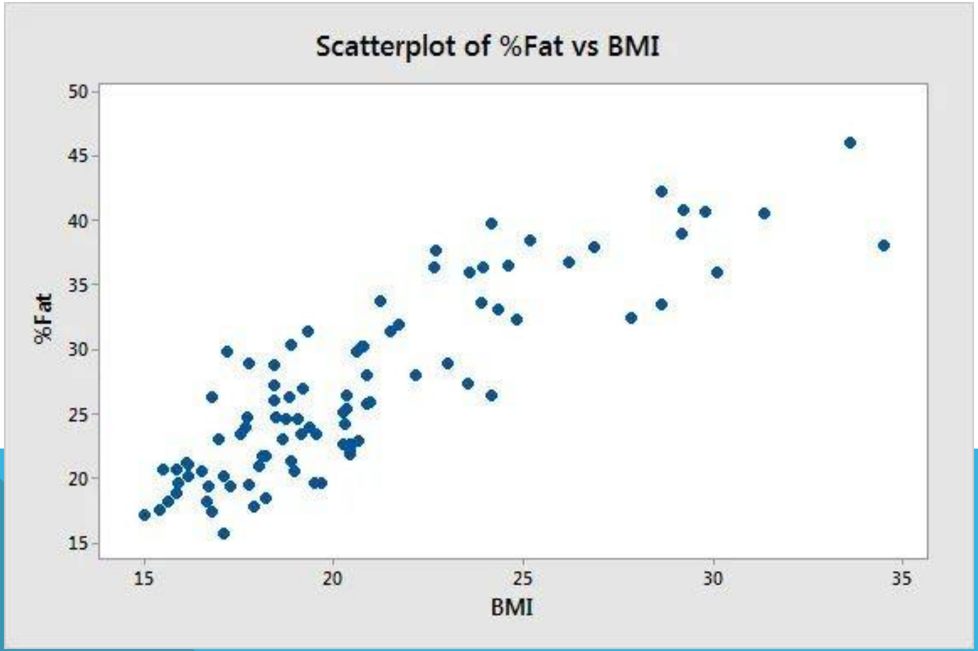
STATISTICS: DATA TYPE AND APPROPRIATE GRAPH

continuous variable



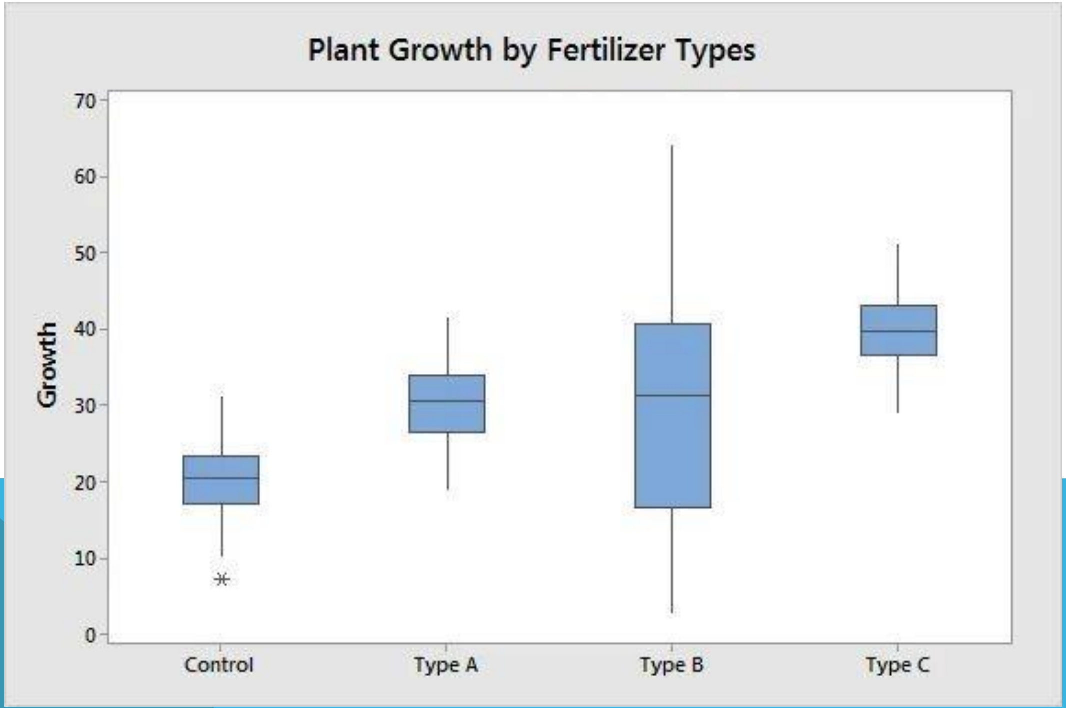
STATISTICS: DATA TYPE AND APPROPRIATE GRAPH

- two continuous variable



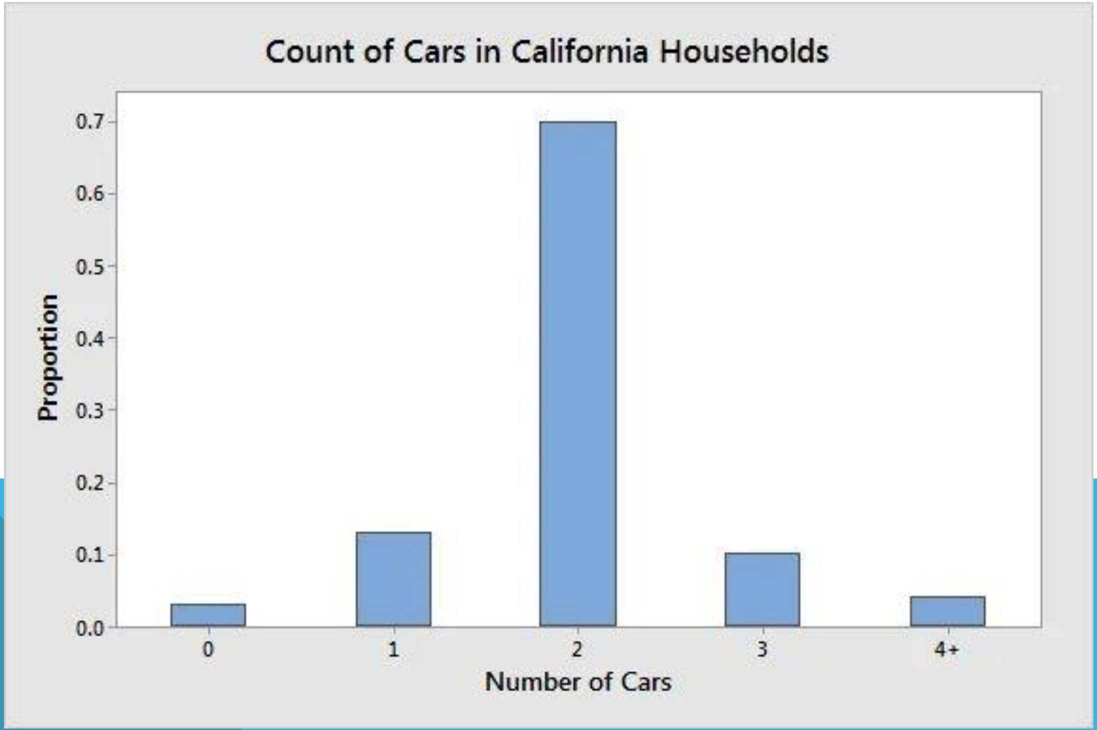
STATISTICS: DATA TYPE AND APPROPRIATE GRAPH

Groupwise continuous variable



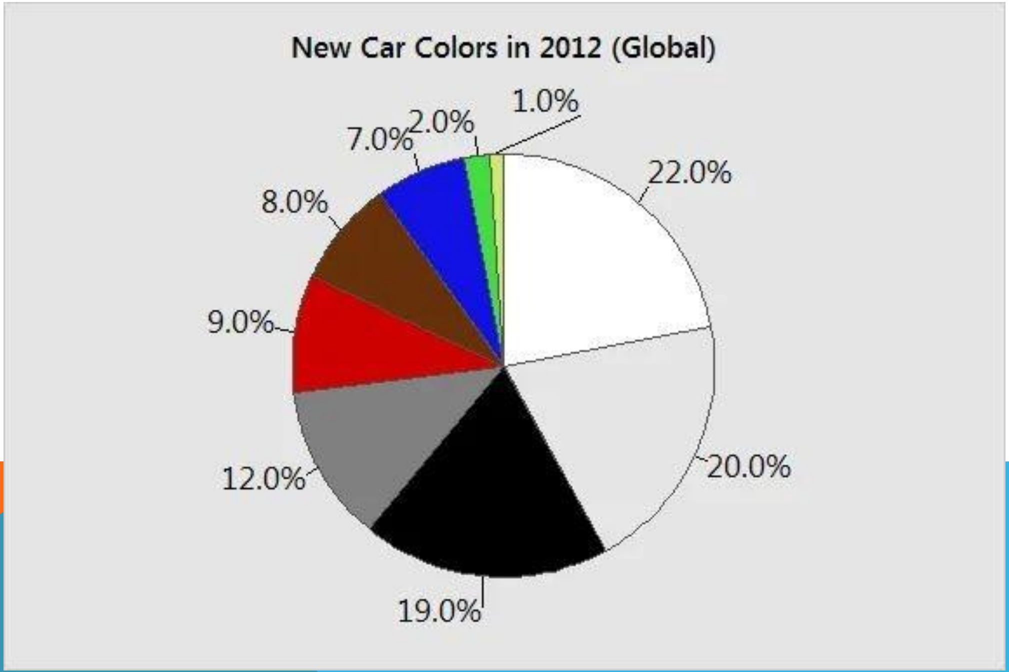
STATISTICS: DATA TYPE AND APPROPRIATE GRAPH

Discrete data



STATISTICS: DATA TYPE AND APPROPRIATE GRAPH

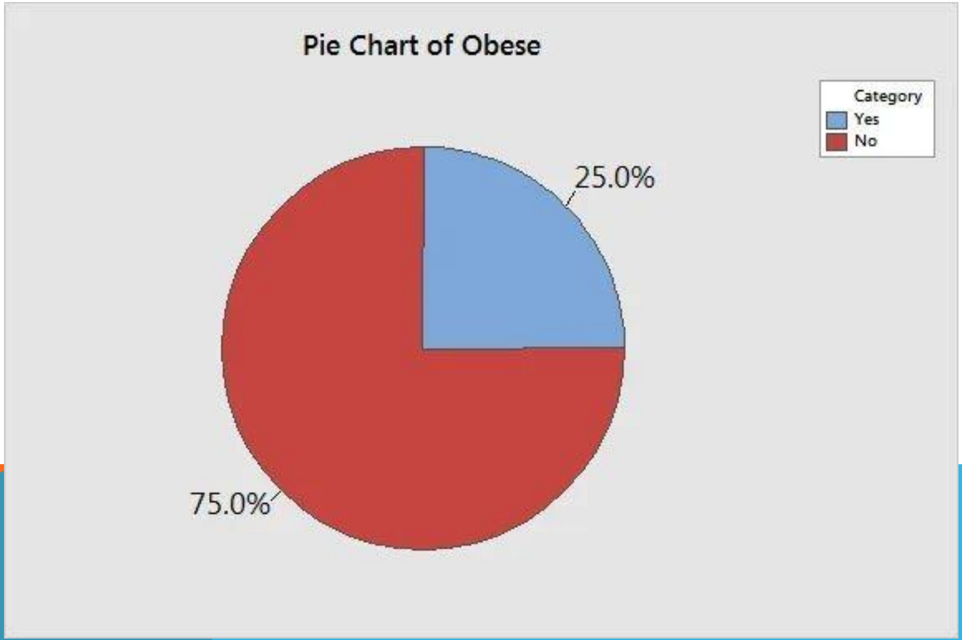
Categorical data



Color
White
Silver
Black
Gray
Red

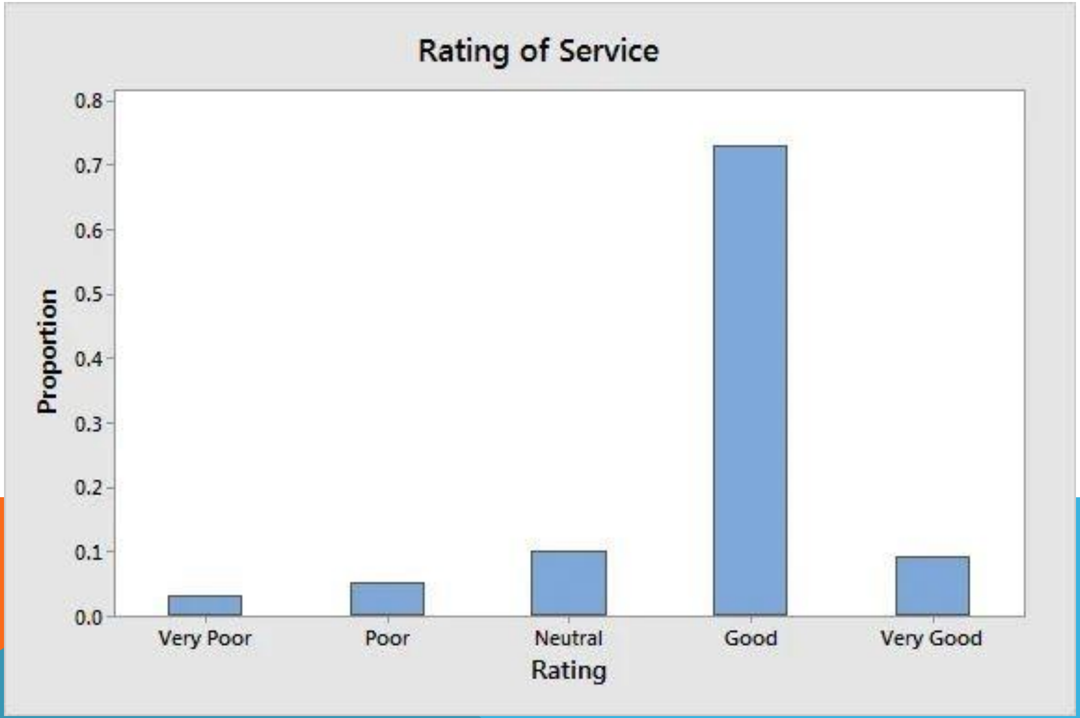
STATISTICS: DATA TYPE AND APPROPRIATE GRAPH

Binary data



STATISTICS: DATA TYPE AND APPROPRIATE GRAPH

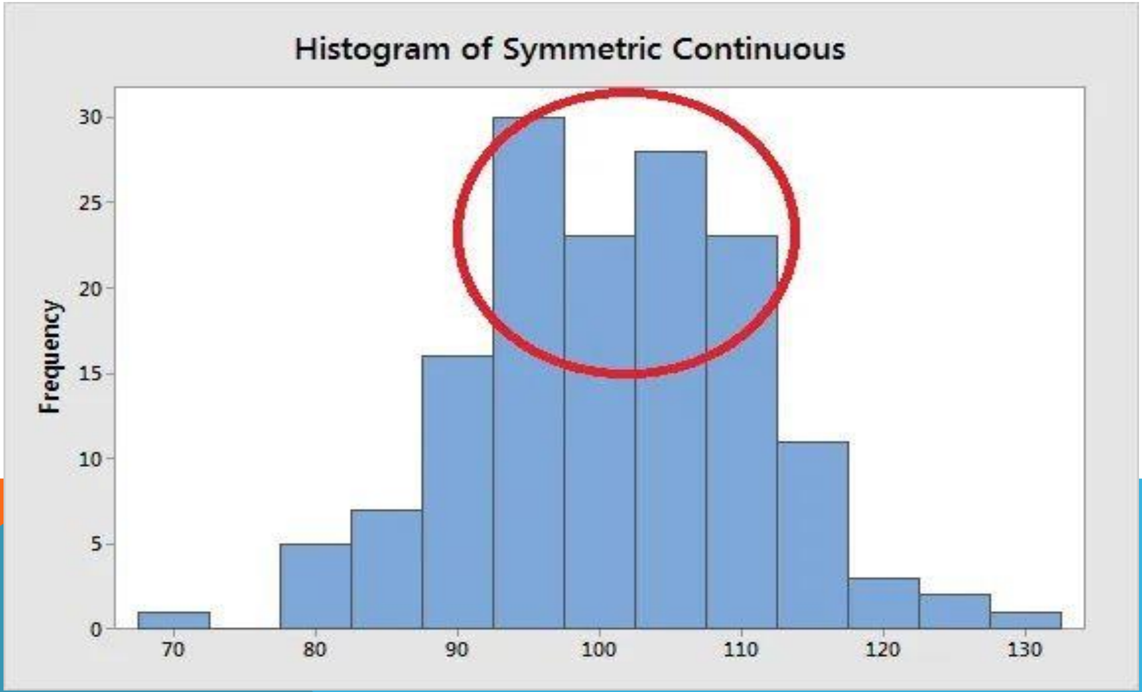
Ordinal data



Rating
Very Poor
Poor
Neutral
Good
Very Good

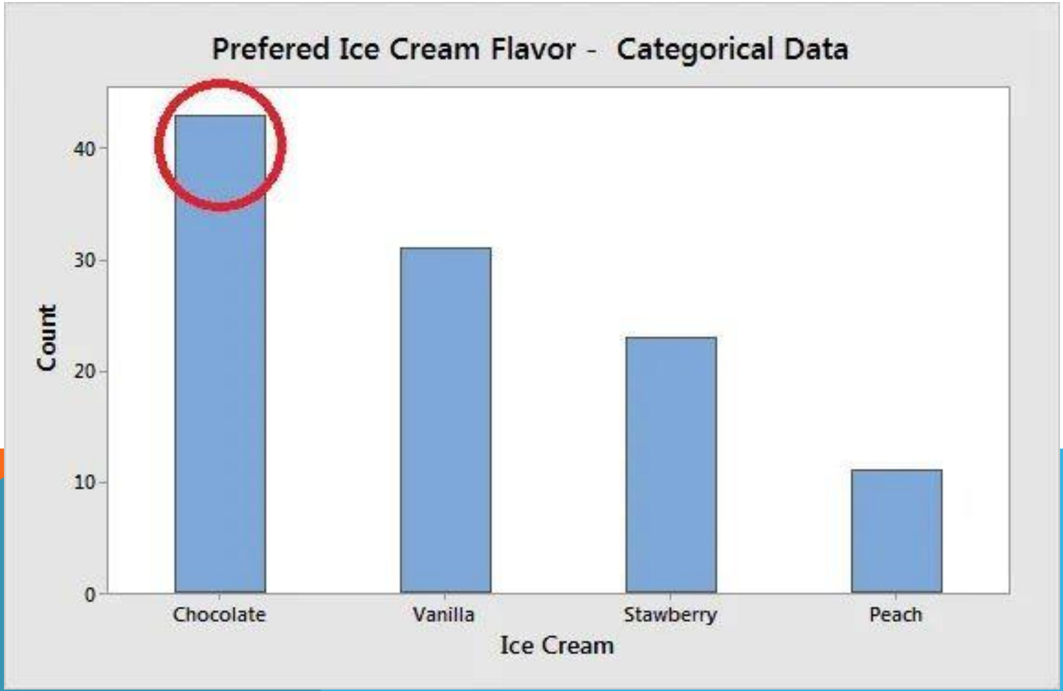
STATISTICS: DATA TYPE AND APPROPRIATE GRAPH

Locating the Center of Your Data



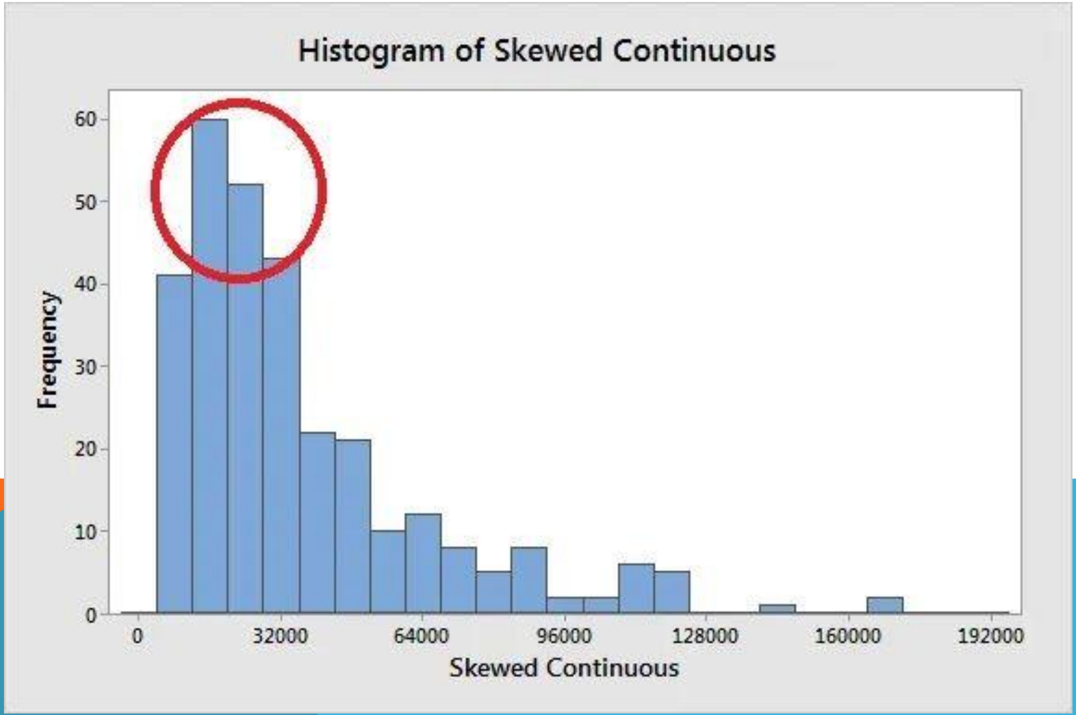
STATISTICS: DATA TYPE AND APPROPRIATE GRAPH

Locating the Center of Your Data



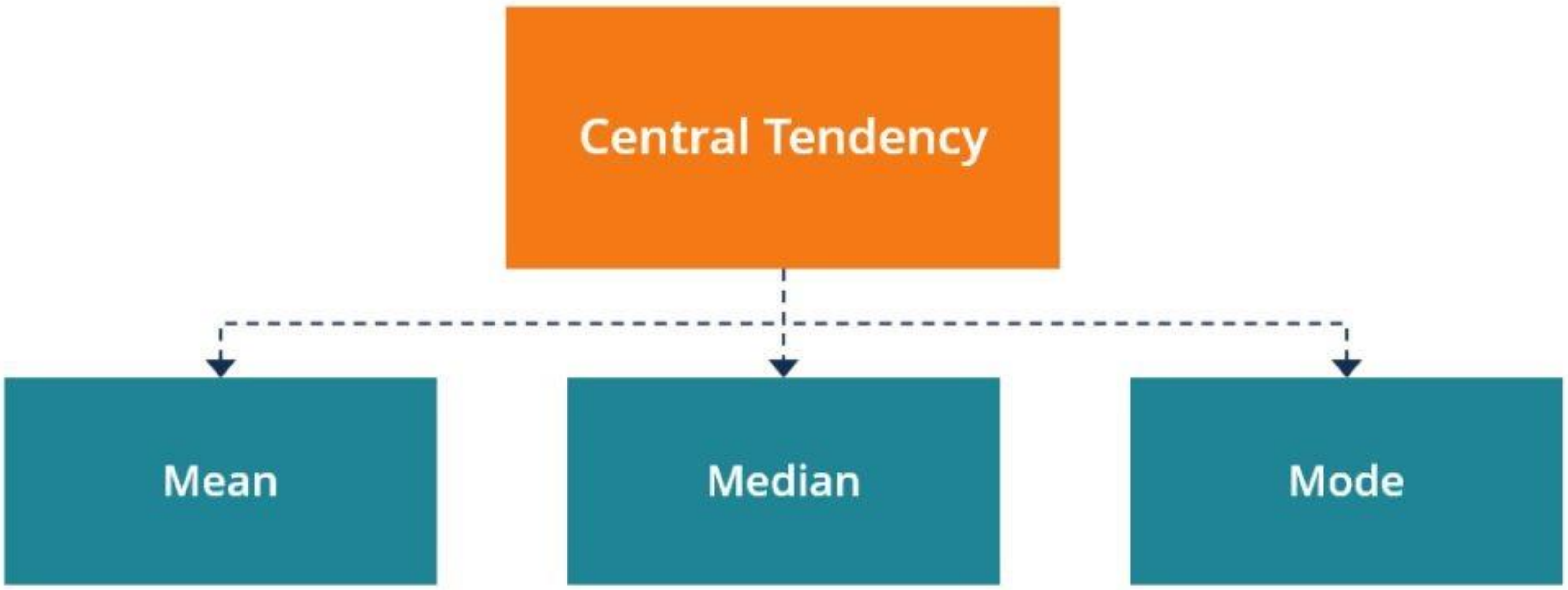
STATISTICS: DATA TYPE AND APPROPRIATE GRAPH

Locating the Center of Your Data



NEED OF STATISTICS

Measures of Central Tendency



NEED OF STATISTICS

Measures of Central Tendency

1 Mean

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

$$30 \div 6 = 5$$

The mean number is

$\boxed{5}$

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

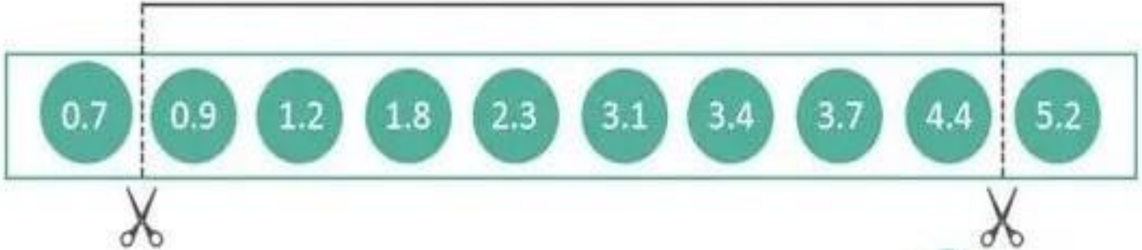
Measures of Central Tendency

1 Mean

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



Loss of Information

- 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+15+12=17+90+95) /10

= 307/10

= 30.7

- Average salary by trimmed mean is 19.7 k
- It is not the best way to accurately reflect the typical salary of a worker

NEED OF STATISTICS

REFERENCES

Measures of Central Tendency

2 Median

Middle of value → 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

28

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

Dr. S.R.Khonde

NEED OF STATISTICS

Measures of Central Tendency

2 Median

Odd number of Samples:

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

Even number of Samples:

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

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 Median = 7+1 / 2 = 4th positional value

Step 4 Median = 40 rooms

NEED OF STATISTICS

Measures of Central Tendency

2 **Median**

Median for Discrete grouped data:

i. Calculate the cumulative frequencies

ii. Find (N+1)/2, where $N=\Sigma f$=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) x	Frequency f	Cumulative Frequency $c.f$
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) x	Frequency f	Cumulative Frequency $c.f$
10	4	4
20	7	11
30	12	23
40	15	38
50	13	51
60	5	56
70	4	60
Total	N = 60	

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

Dr. S.R.Khonde

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
Total	N = 200	

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 = 10

Step 6

f= frequency of 440-450=54

Dr. S.R.Khonde

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
Total	N = 200	

Step 7

Median = 440+((100-76)/54)*10

= 440+4.44

= 444.44

Dr. S.R.Khonde

NEED OF STATISTICS

Measures of Central Tendency

3 Mode

Most frequent of value → 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

NEED OF STATISTICS

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.

f1 = frequency of the modal class

f0 = frequency of the class preceding the modal class

f2 = frequency of the class succeeding the modal class

c = width of the class limits

Dr. S.R.Khonde

NEED OF STATISTICS

Measures of Central Tendency

3 Mode

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

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

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

Step 2

Highest Frequency is 18. Modal class is 300-400

l= lower boundary of 300-400 = 300

Step 3

f1= frequency of 300-400 =18

Step 4

f0= frequency of 200-300 =12

Step 5

f2= frequency of 440-500 =16

Step 6

Mode= 300+ (18-12)/(2*18-12-16)*100

=300+6/(36-28) *100

=300+ 600/8 = 300+75= **375**

NEED OF STATISTICS

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

Average → of → Largest and Smallest instance → of → dataset

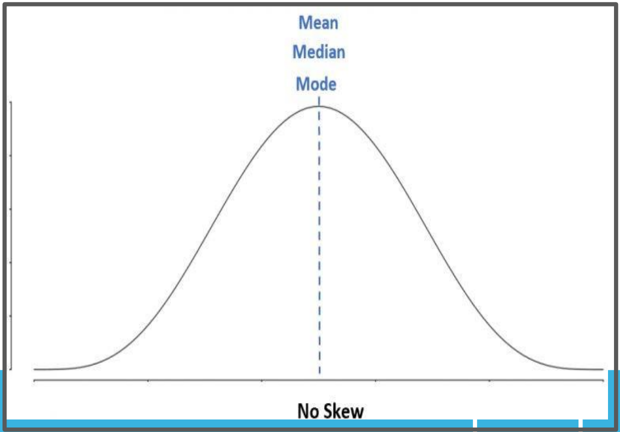
Example		
Problem	**Find the range and midrange for the following set of numbers: 2, 4, 7, 10, 14, 35.**	
	range: $35 - 2 = 33$	Subtract the least value from the greatest value to find the range.
	midrange: $\frac{35+2}{2} = \frac{37}{2} = 18.5$	Add together the greatest value and the least value and divide by 2.
Answer	The range is 33. The midrange is 18.5.	

RELATIONSHIP AMONG MEAN, MEDIAN AND MODE

No of days spend in training	
Team1	4
Team2	5
Team3	6
Team4	6
Team5	6
Team6	7
Team7	7
Team8	7
Team9	7
Team10	7
Team11	7
Team12	8
Team13	8
Team14	8
Team15	9
Team16	10

Mean	7
Median	7
Mode	7

Mean= Median=Mode



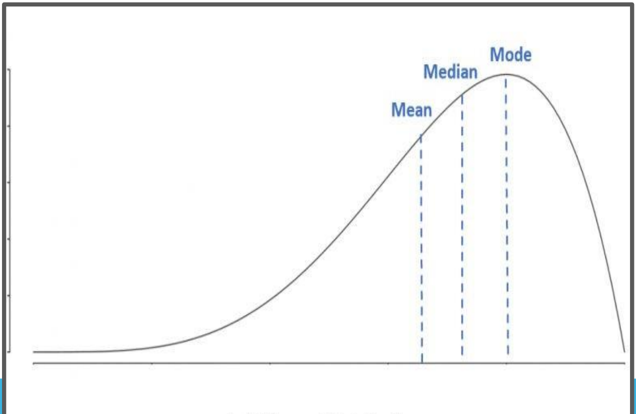
Symmetric Data Distribution

RELATIONSHIP AMONG MEAN, MEDIAN AND MODE

No of days spend in training	
Team1	4
Team2	5
Team3	6
Team4	6
Team5	6
Team6	7
Team7	7
Team8	7
Team9	7
Team10	8

Mean	6.3
Median	6.5
Mode	7

Mean < Median < Mode



Left Skewed Distribution

Left Skew data distribution

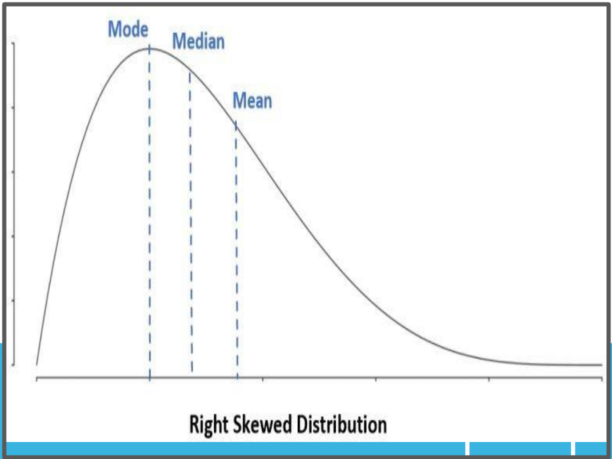
Dr. S.R.Khonde

RELATIONSHIP AMONG MEAN, MEDIAN AND MODE

No of days spend in training	
Team1	6
Team2	7
Team3	7
Team4	7
Team5	7
Team6	8
Team7	8
Team8	8
Team9	9
Team10	10

Mean	7.7
Median	7.5
Mode	7

Mode < Median < Mean



Right Skew data distribution

Dr. S.R.Khonde

NEED OF STATISTICS

Measures of Central Tendency

Empirical Relationship among mean, median and mode

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

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

$$0.2 \underset{\approx}{\rightarrow} 0.26$$

For Right Skew Data

Mean	7.7
Median	7.5
Mode	7

NEED OF STATISTICS

Measures of Central Tendency

Empirical Relationship among mean, median and mode

Mean- Median= ⅓(Mean-Mode)

6.3 -6.5= ⅓ (6.3-7)

(-0.2) ⥵(-0.26)

For Left Skew Data

Mean	6.3
Median	6.5
Mode	7

Dr. S.R.Khonde

NEED OF STATISTICS

Measures of Central Tendency

Empirical Relationship among mean, median and mode

Mean- Median= ⅓(Mean-Mode)

mean - mode = 3´(mean - median)

Dr. S.R.Khonde

NEED OF STATISTICS

Measures of Dispersion

01 Range

02 Variance

03 Standard Deviation

04 Mean Deviation

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

MEASURES OF DISPERSION

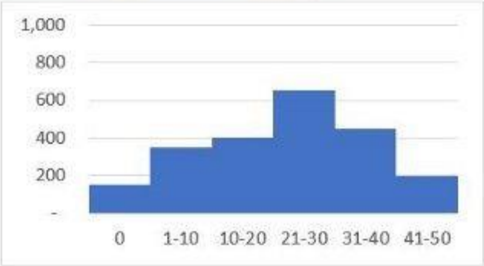
Unit - II

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
Total	**2,200**

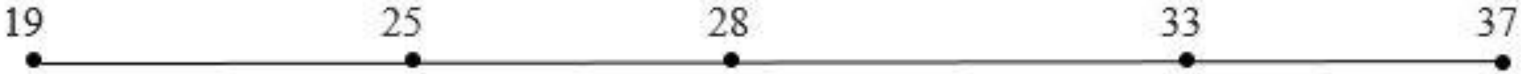
Distribution 2	
Value of X	Frequency
0	150
1-10	350
10-20	400
21-30	650
31-40	450
41-50	200
Total	**2,200**



MEASURES OF DISPERSION

01 Range

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



Range = H – L

= 37-19 $\Longrightarrow$ 18

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

19 25 28 33 60

Range = H – L

= 60-19 ⟹ 41 Now range of marks is 41.

Range of Sequence is 18

Dr. S.R.Khonde

MEASURES OF DISPERSION

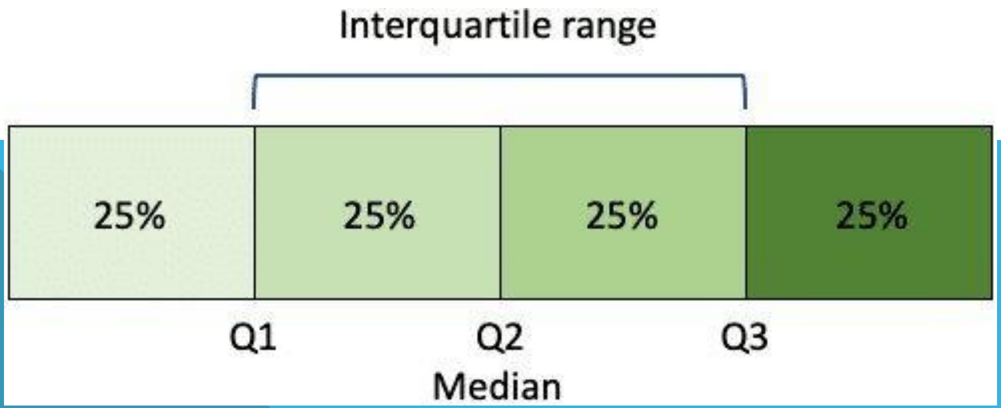
01

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



Dr. S.R.Khonde

MEASURES OF DISPERSION

01

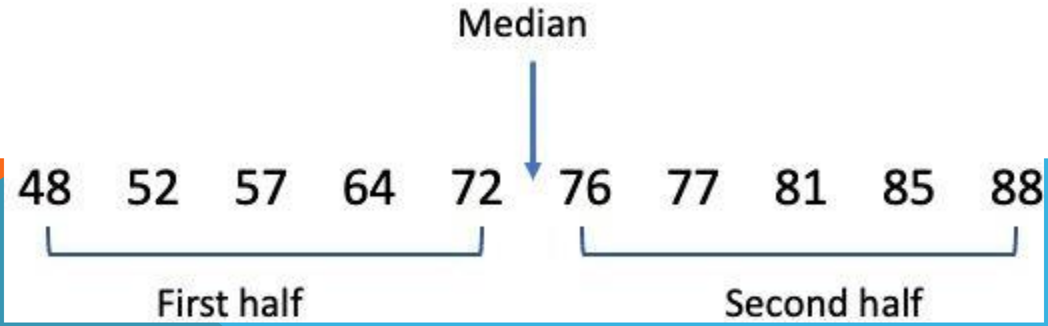
Inter-quartile Range

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.



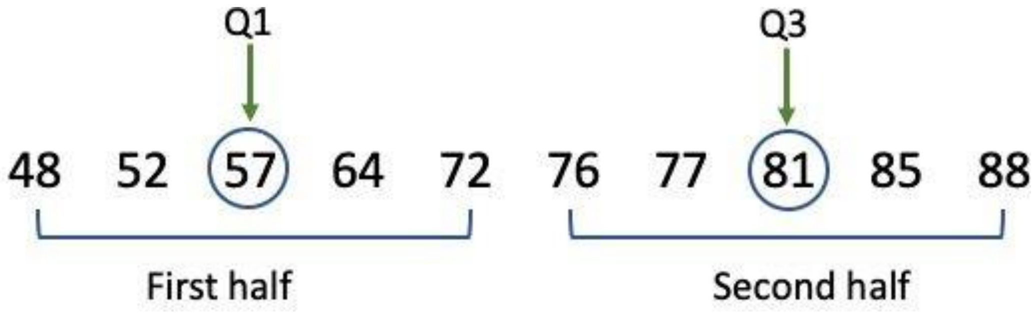
MEASURES OF DISPERSION

01

Inter-quartile Range

Even number of Elements

Find Q1 and Q3.



IQR

$$IQR = Q3 - Q1$$
$$IQR = 81 - 57 = 24$$

Dr. S.R.Khonde

MEASURES OF DISPERSION

01

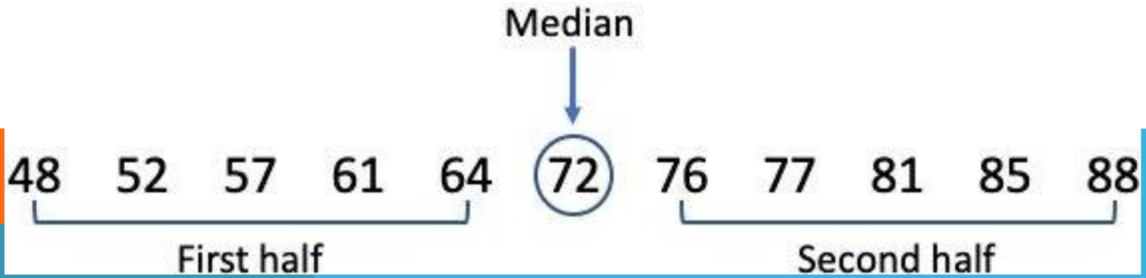
Inter-quartile Range

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.



Dr. S.R.Khonde

MEASURES OF DISPERSION

01

Inter-quartile Range

Even number of Elements

With Inclusion

Find Q1 and Q3.

$$\underbrace{48 \quad 52 \quad 57 \quad 61 \quad 64 \quad 72}_{\text{First half}} \quad \underbrace{72 \quad 76 \quad 77 \quad 81 \quad 85 \quad 88}_{\text{Second half}}$$

Q1 Q3

$$\underbrace{48 \quad 52 \quad 57 \quad 61 \quad 64 \quad 72}_{\text{First half}} \quad \underbrace{72 \quad 76 \quad 77 \quad 81 \quad 85 \quad 88}_{\text{Second half}}$$

$$Q1 = \frac{57 + 61}{2} = 59 \qquad Q3 = \frac{77 + 81}{2} = 79$$

Dr. S.R.Khonde

MEASURES OF DISPERSION

01

Inter-quartile Range

Even number of Elements

With Inclusion

IQR

$$IQR = Q3 - Q1$$
$$IQR = 79 - 59 = 20$$

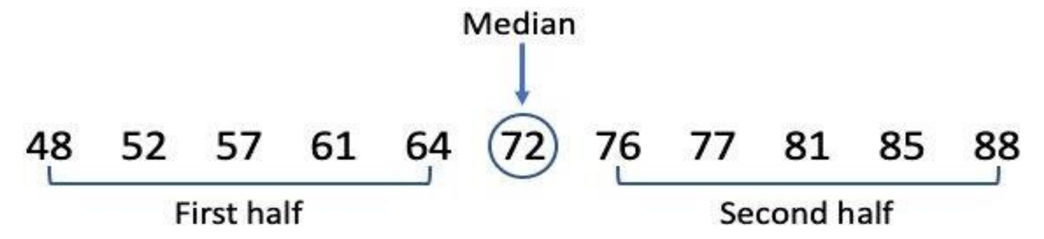
01

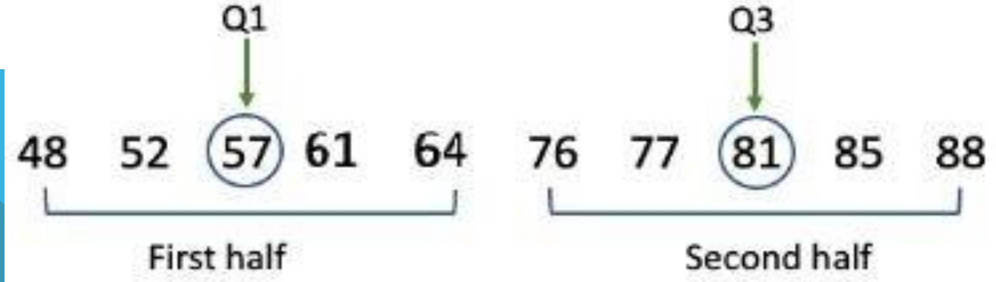
Inter-quartile Range

Even number of Elements

Exclusive Method

Find Q1 and Q3.





Dr. S.R.Khonde

MEASURES OF DISPERSION

01

Inter-quartile Range

Even number of Elements

Exclusive Method

IQR

$$IQR = Q3 - Q1$$
$$IQR = 81 - 57 = 24$$

MEASURES OF DISPERSION

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

MEASURES OF DISPERSION

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

Dr. S.R.Khonde

MEASURES OF DISPERSION

$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**

MEASURES OF DISPERSION

02

Variance

Variance is the average of the squared differences from the mean

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

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

Mean : 70

Mean : 70

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

MEASURES OF DISPERSION

02 Variance

Formula to Compute Variance

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

X – Input variable

$\bar{X}$ - mean of input dataset

02

Variance

	Score A	$x - \bar{x}$	$(x - \bar{x})^2$
1	30		
2	50		
3	70		
4	100		
5	100		
Total	**350**		

Mean : 70

MEASURES OF DISPERSION

02

Variance

	Score 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	
Total	**350**		

MEASURES OF DISPERSION

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
Total	**350**		3800

Dr. S.R.Khonde

MEASURES OF DISPERSION

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
Total	**350**		3800

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

Variance

= 3800/5

=760

Dr. S.R.Khonde

MEASURES OF DISPERSION

02

Variance

	Score 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
Totals	350		0

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

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?

MEASURES OF DISPERSION

02

Variance

Dive	Mark's Score X	$X-\overline{X}$	$(X-\overline{X})^2$
1	28	5	25
2	22	-1	1
3	21	-2	4
4	26	3	9
5	18	-5	25
Totals	**115**	**0**	**64**

Dive	Myrna's Score X	$X-\overline{X}$	$(X-\overline{X})^2$
1	27	4	16
2	27	4	16
3	28	5	25
4	06	-17	289
5	27	4	16
Totals	**115**	**0**	**362**

MEASURES OF DISPERSION

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

- Mean=(5+7+9+3)/4=6

$$\textit{Mean Deviation} = \frac{(5-6)+(7-6)+(9-6)+(3-6)}{4}$$

$$= \frac{(-1)+(1)+(3)+(-3)}{4} \Rightarrow 0$$

$$\textit{Mean Absolute Deviation} = \frac{|5-6|+|7-6|+|9-6|+|3-6|}{4}$$

$$= \frac{(1)+(1)+(3)+(3)}{4} \Rightarrow 2$$

MEASURES OF DISPERSION

03

Mean Deviation

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

Where μ = mean, X = score, ∑= the sum of, N = number of scores, ∑X = "add up all the scores",

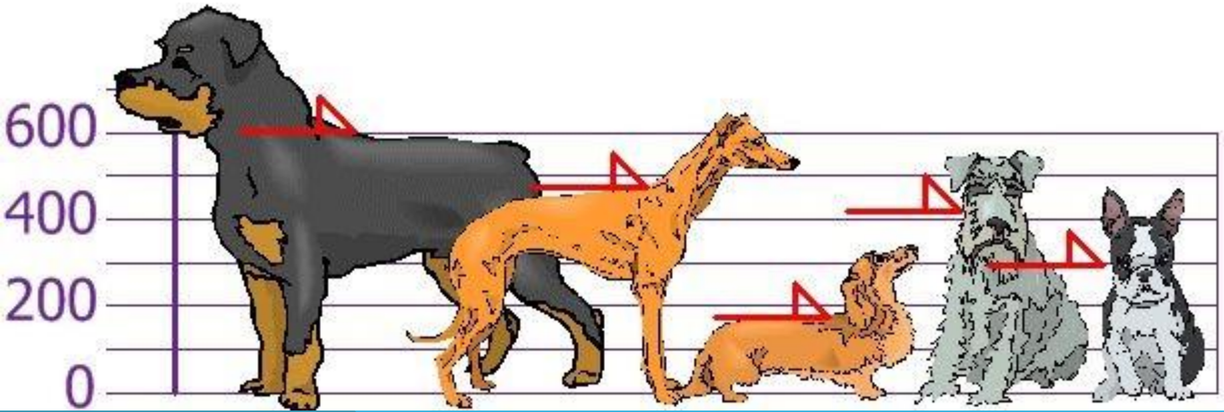
| | = take the absolute value (i.e. ignore the minus sign).

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**:

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

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

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

MEASURES OF DISPERSION

03

Mean Deviation

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

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

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 σ (the greek letter sigma)

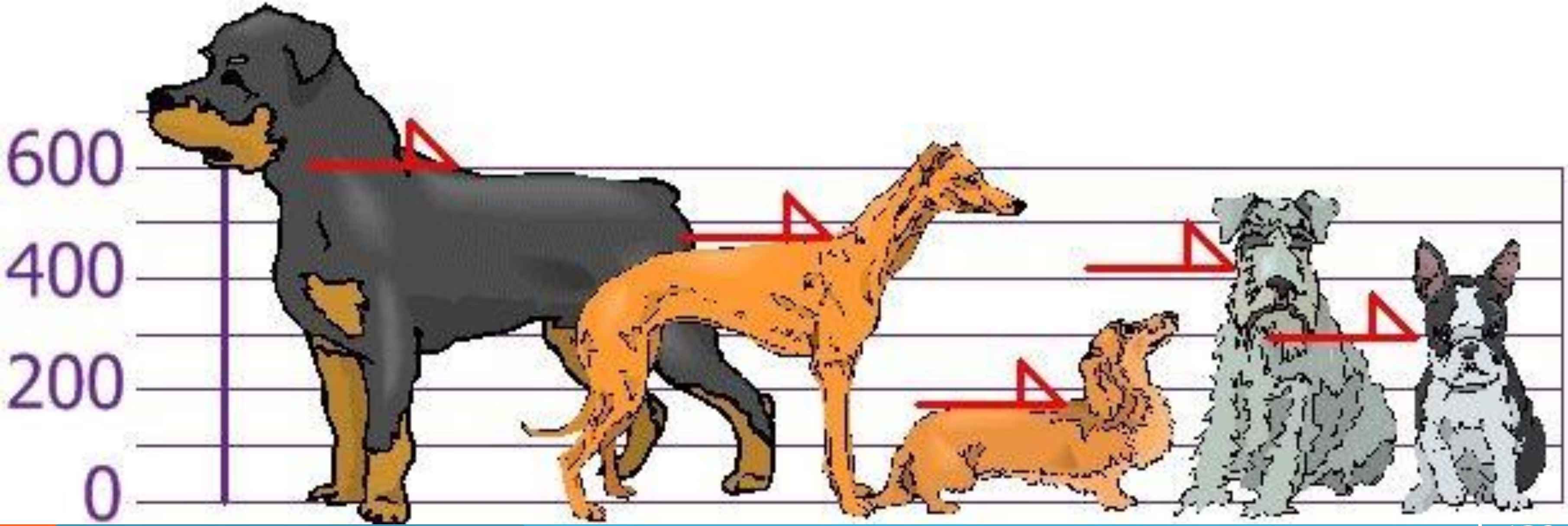
- It is the square root of the Variance

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

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.

Mean= 600 + 470 + 170 + 430 + 300/ 5

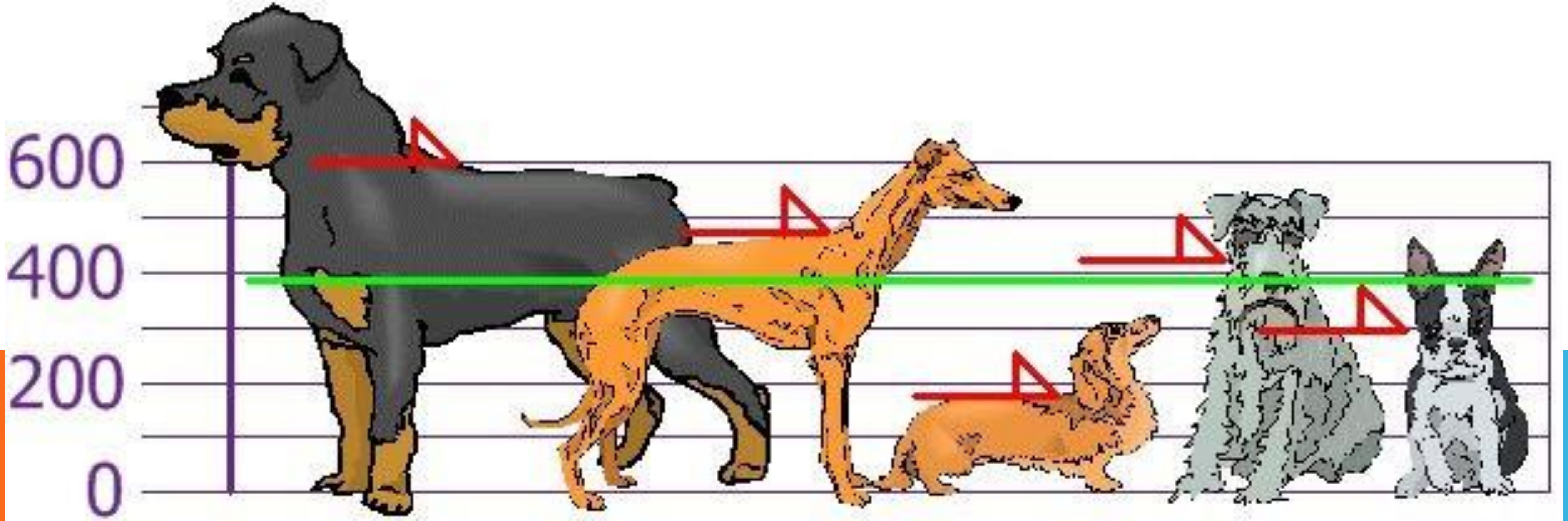
= 1970 / 5

= 394

MEASURES OF DISPERSION

Mean, Variance and Standard Deviation

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

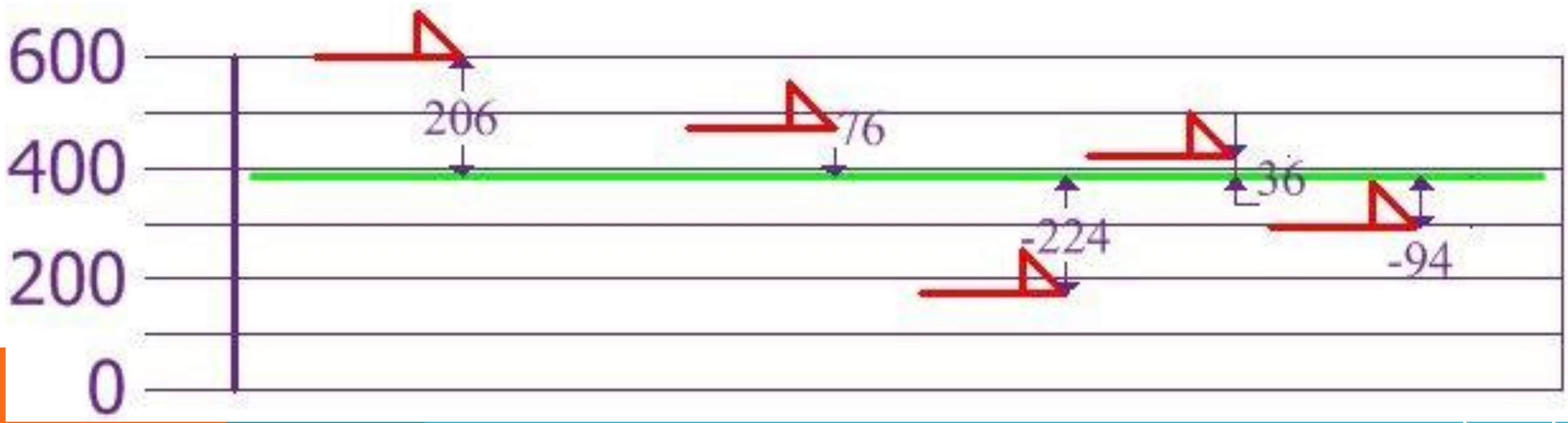


Dr. S.R.Khonde

MEASURES OF DISPERSION

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

$$\sigma^2 = \frac{206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2}{5}$$

$$= \frac{42436 + 5776 + 50176 + 1296 + 8836}{5}$$

$$= \frac{108520}{5}$$

$$= 21704$$

So the Variance is **21,704**

Mean, Variance and Standard Deviation

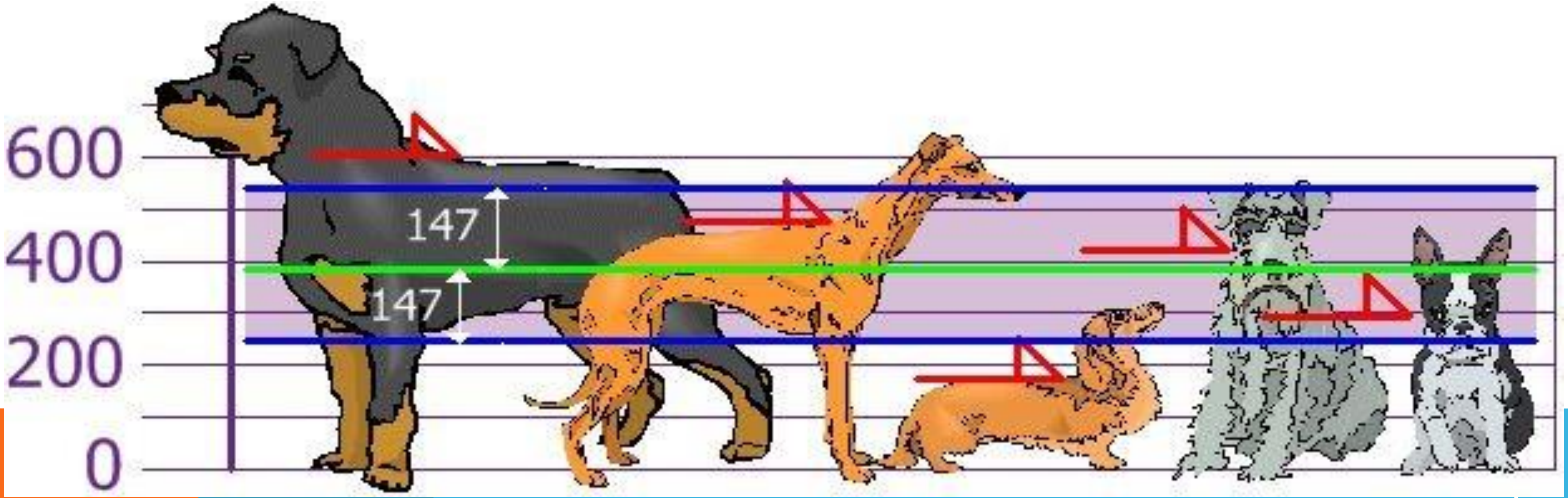
Standard Deviation

$$\sigma = \sqrt{21704}$$
$$= 147.32...$$
$$= \mathbf{147}\ \textit{(to the nearest mm)}$$

MEASURES OF DISPERSION

Mean, Variance and Standard Deviation

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

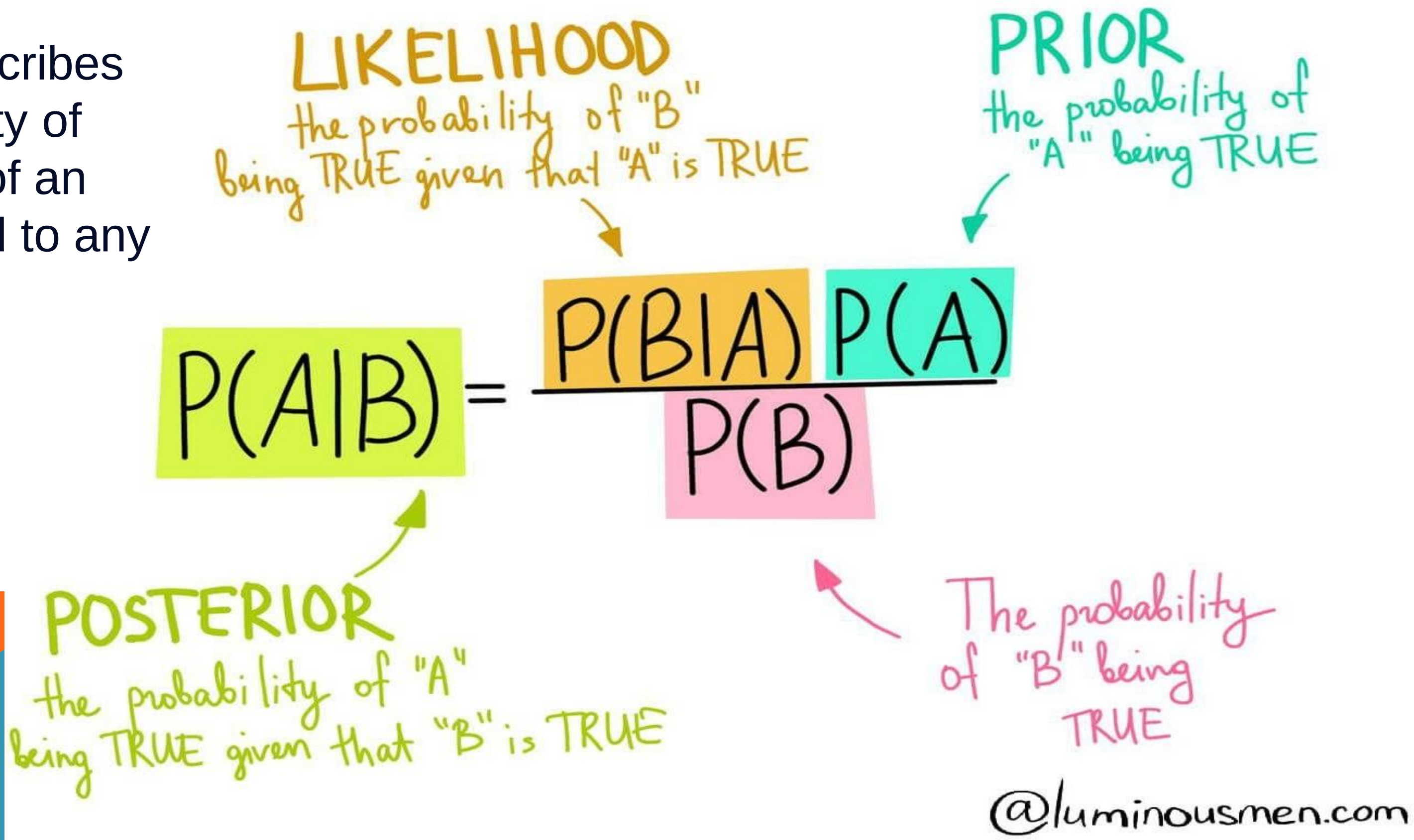


So, using the Standard Deviation we have a "standard" way of knowing what is normal

Data type	Mathematical operations	Measures of central tendency	Measures of variability
Nominal	• Equality (=,)	• Mode	• None
Ordinal	• Equality (=,) • Comparison (>, <)	• Mode • Median	• Range • Interquartile range
Interval	• Equality (=,) • Comparison (>, <) • Addition, subtraction (+,)	• Mode • Median • Arithmetic mean	• Range • Interquartile range • Standard deviation • Variance
Ratio	• Equality (=,) • Comparison (>, <) • Addition, subtraction (+,) • Multiplication, division (×, ÷)	• Mode • Median • Arithmetic mean • *Geometric mean	• Range • Interquartile range • Standard deviation • Variance • **Relative standard deviation

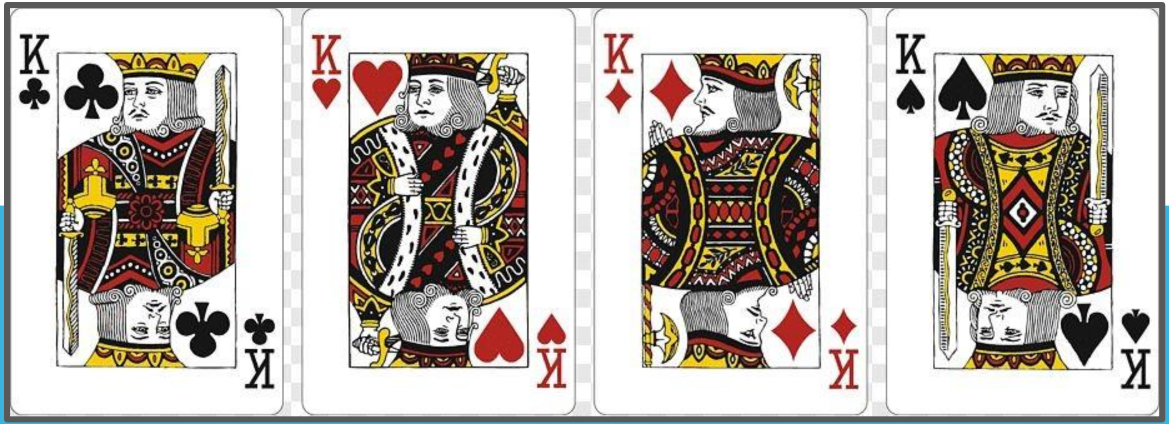
BAYES THEOREM

Bayes' theorem describes the probability of occurrence of an event related to any condition.



BAYES THEOREM

Probability of King



BAYES THEOREM

Marginal Probability:

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

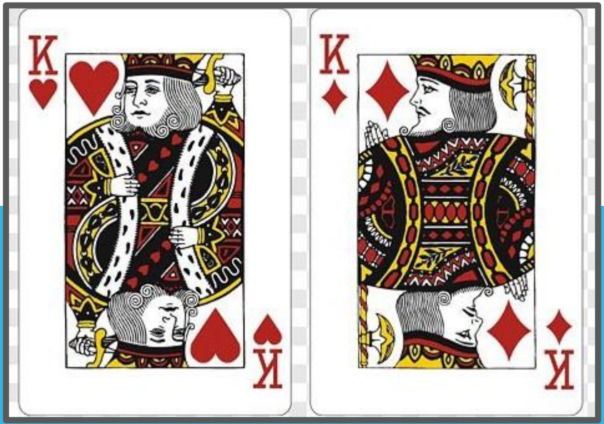
P(King)

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

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 and B) or P(A, B)

P(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

BAYES THEOREM

Conditional Probability:

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

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

P(A | B) = P(A ∩B)/ P(B)

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

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

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

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

BAYES THEOREM

- 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**?

BAYES THEOREM

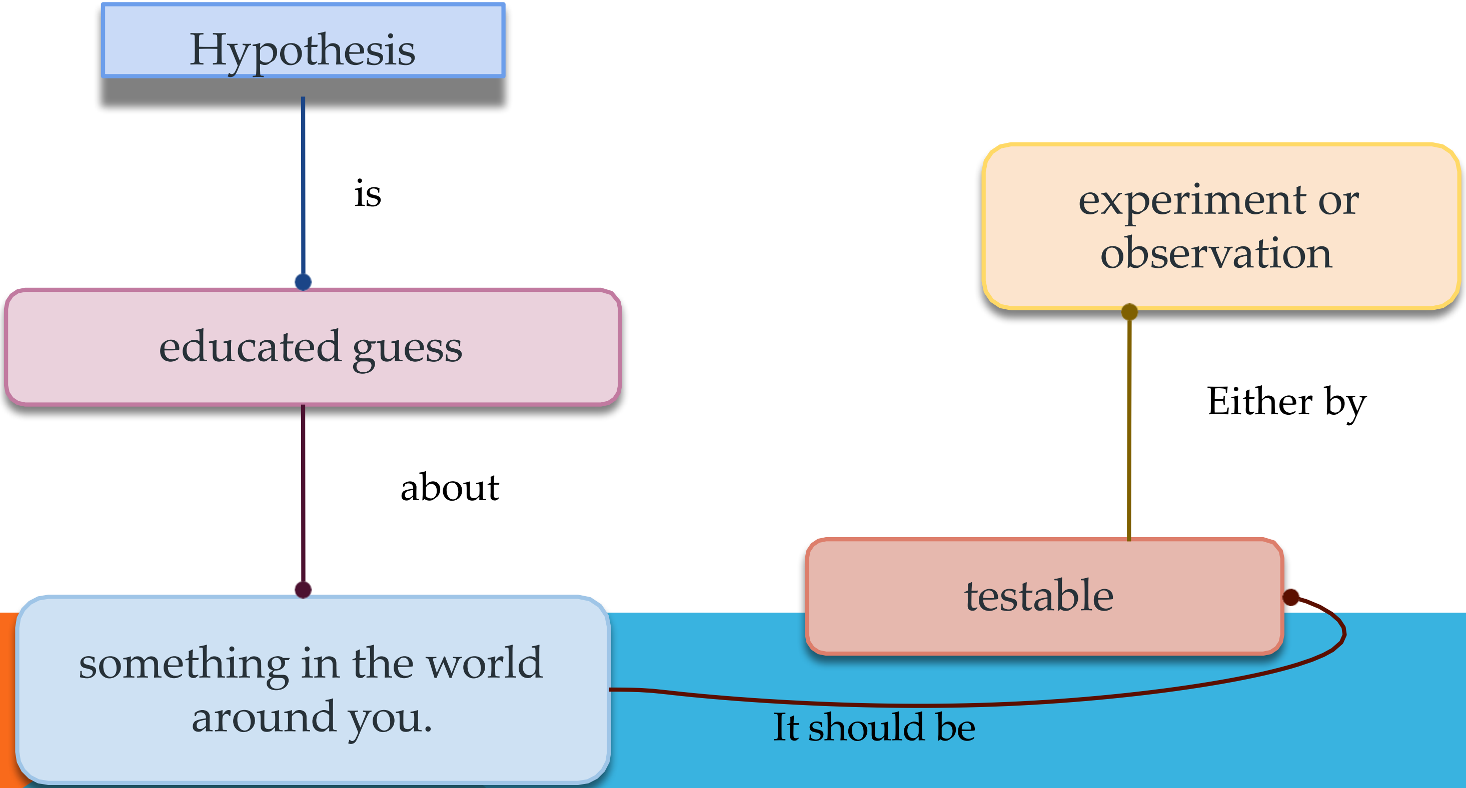
- Let A, B and C be the events that notebooks are provided by A, B and C respectively.
- Let D be the event that notebooks are defective
- Then,
- P(A) = 0.25, P(B) = 0.35, P(C) = 0.4
- P(D|A) = 0.05, P(D|B) = 0.04, P(D|C) = 0.02
- P(A | D) = (P(D | A)*P(A))/(P(D | A) * P(A) + P(D | B) * P(B) + P(D | C) * P(C))

=(0.05*0.25)/((0.05*0.25)+(0.04*0.35)+(0.02*0.4))

= 2000/(80*69)

=25/69.

BASICS OF HYPOTHESIS



BASIC OF HYPOTHESIS

Which is better?

Research Question

BASIC OF HYPOTHESIS

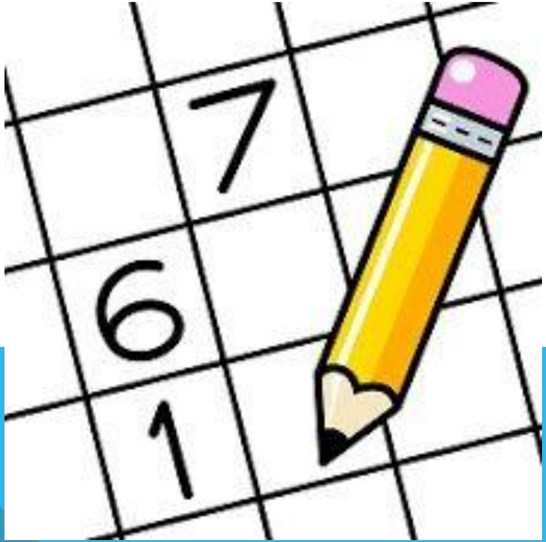
Both vaccination has same efficiency

Hypothesis Statement

BASIC OF HYPOTHESIS

Who can solve Sudoku Faster? Girls/Boys

Research Question

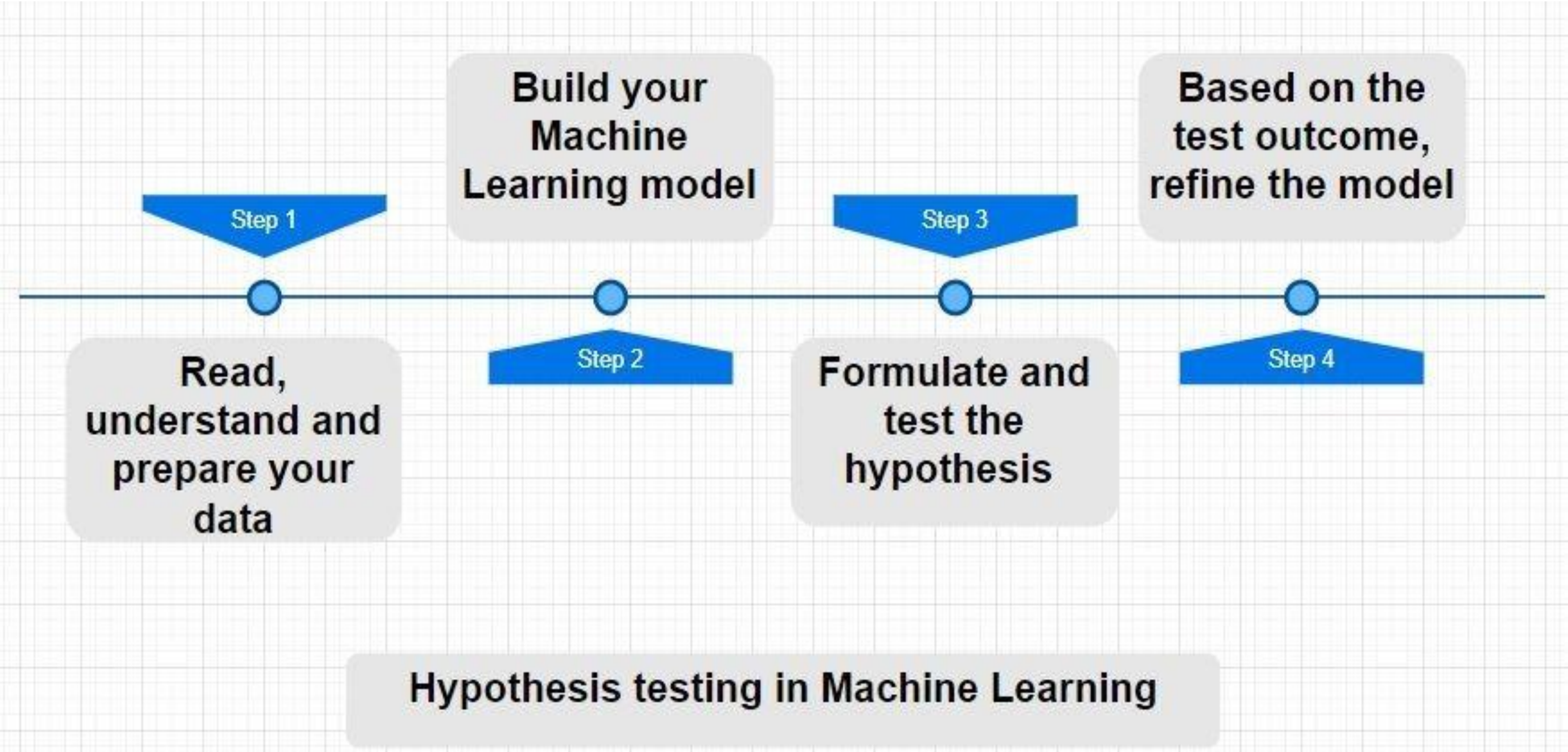


BASIC OF HYPOTHESIS

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

Hypothesis

NEED OF HYPOTHESIS



HYPOTHESIS TESTING

- 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

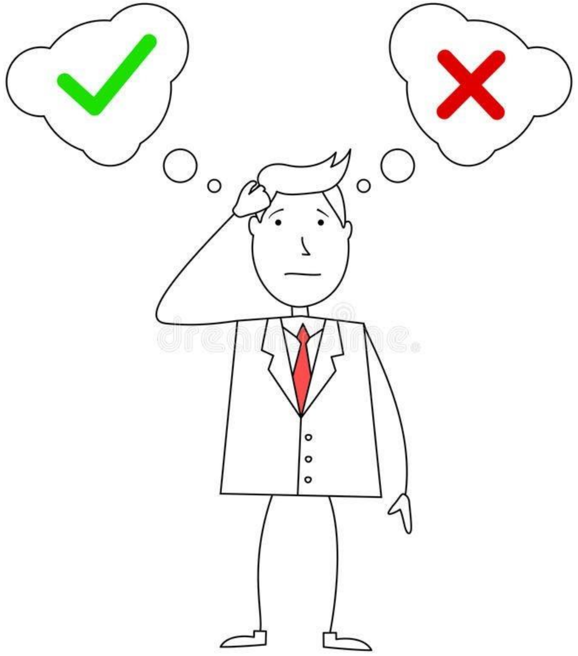
HYPOTHESIS TESTING

- The best way to determine if a hypothesis was true would be to examine the entire population

- Usually impractical (time, money, resources)

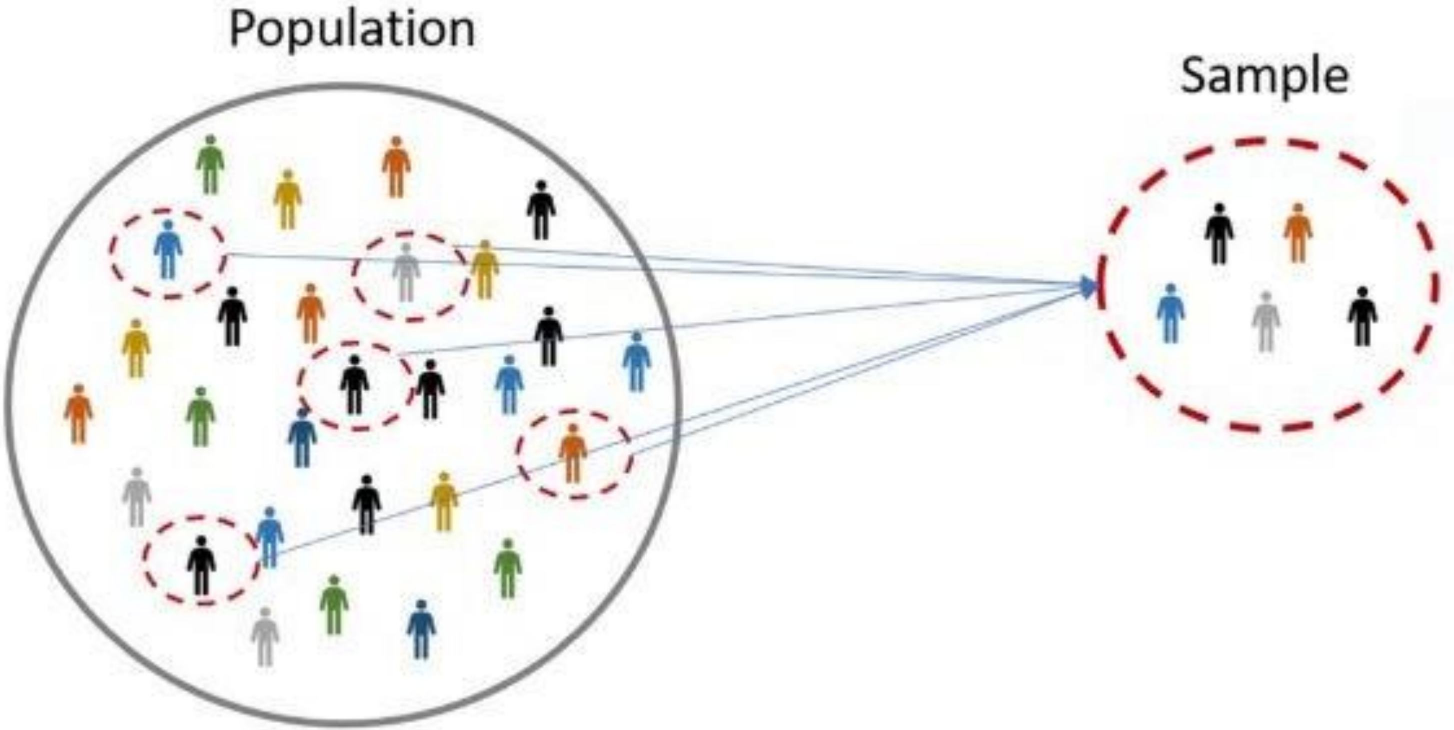
- Examine random samples from population

- If sample data are not consistent with hypothesis – reject

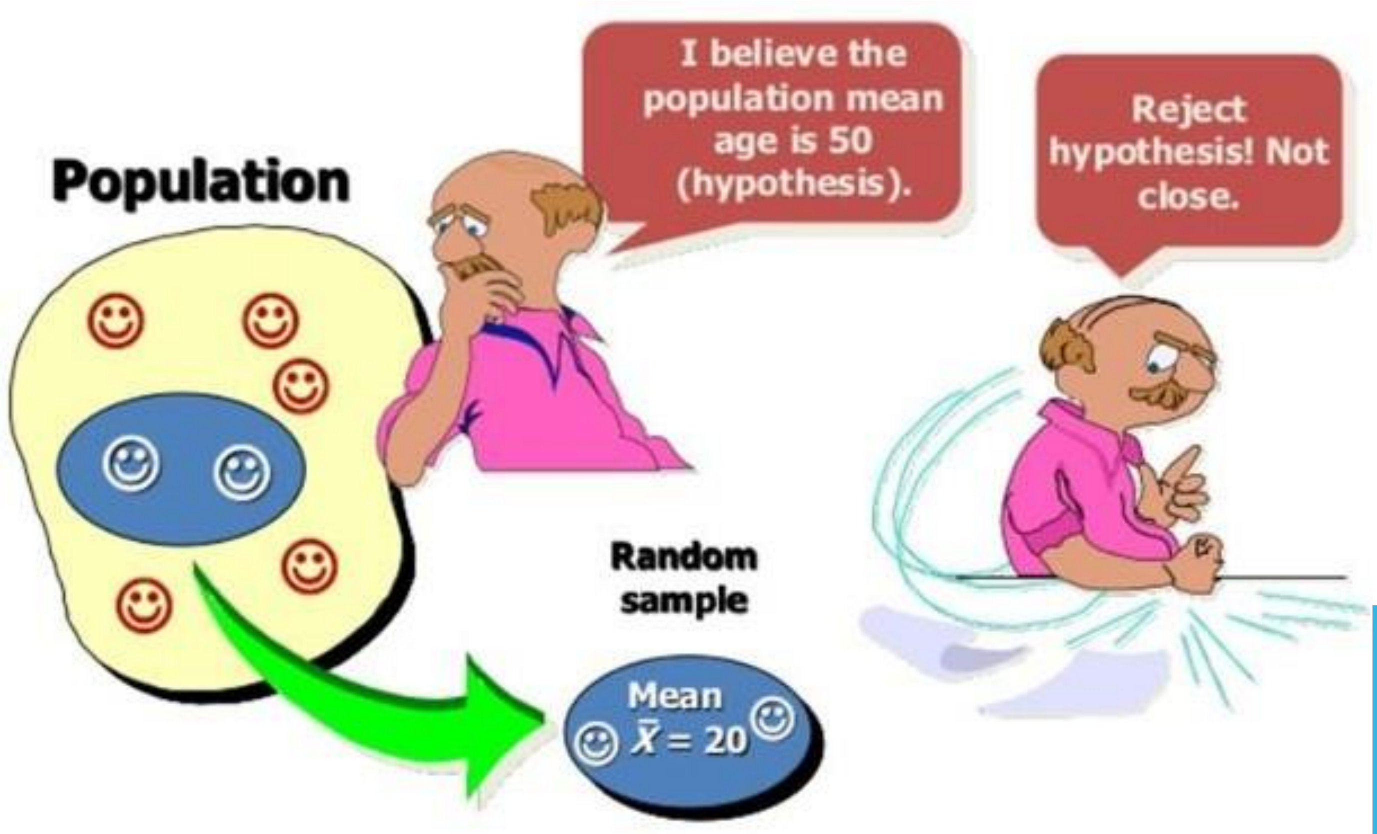


Dr. S.R.Khonde

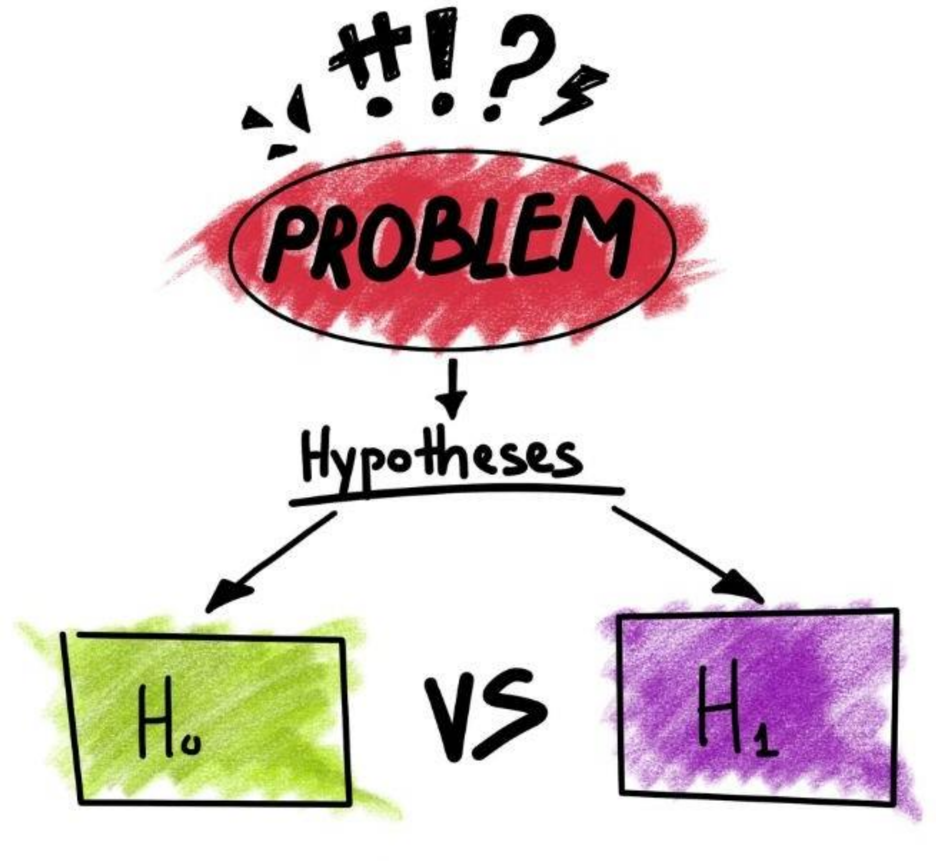
HYPOTHESIS TESTING



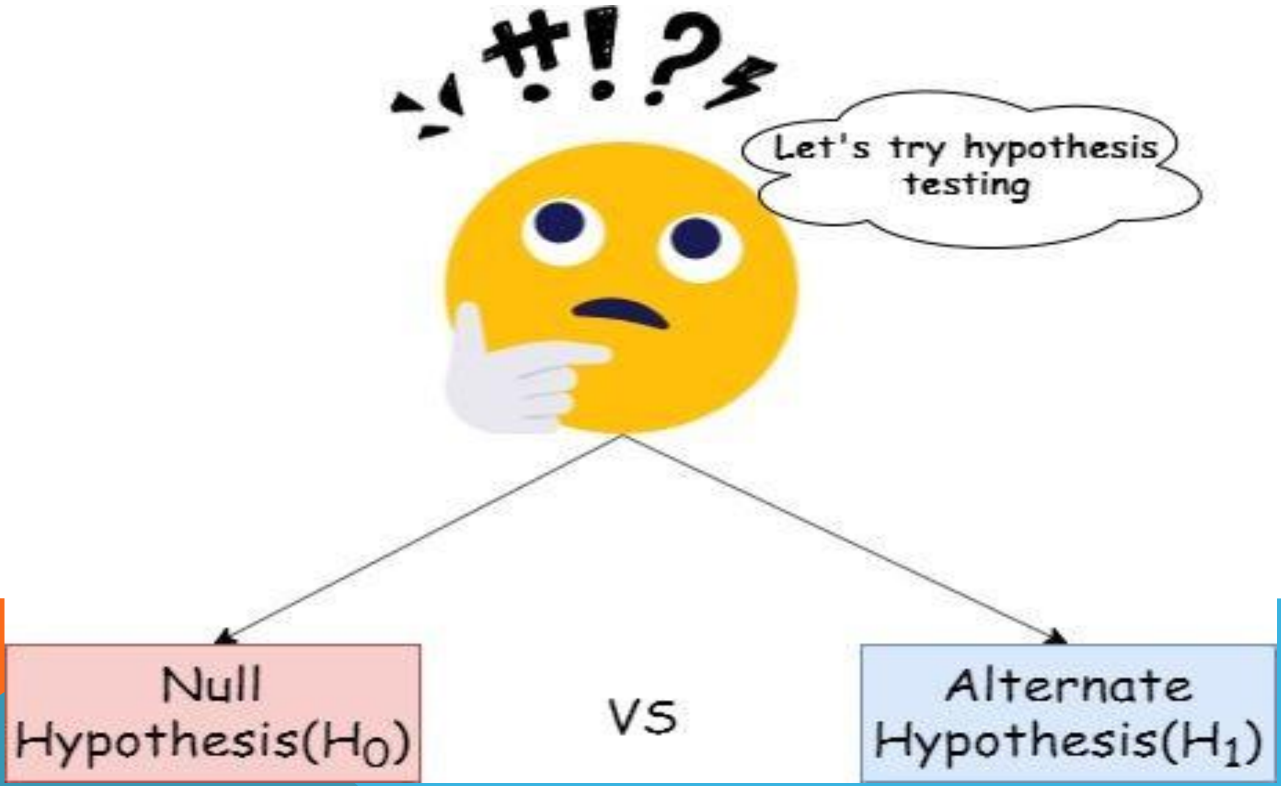
HYPOTHESIS TESTING



HYPOTHESIS TESTING



HYPOTHESIS TESTING



HYPOTHESIS TESTING

Null Hypothesis(H_0)

All dogs have Four Lags

HYPOTHESIS TESTING

Alternate Hypothesis(H_1)

5% dogs have Three Legs

HYPOTHESIS TESTING

Statistical Hypothesis

Null hypothesis

H_0

- 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

HYPOTHESIS TESTING

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.

HYPOTHESIS TESTING

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

HYPOTHESIS 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

Dr. S.R.Khonde

HYPOTHESIS TESTING

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: $\neq$ or < or >
Signs used in Minitab: = or $\geq$ or $\leq$	

HYPOTHESIS TESTING

Assignment is due for my subject

Hypothesis:

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

HYPOTHESIS TESTING

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

μ

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

Suppose the sample mean is 6.1 days

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

Dr. S.R.Khonde

HYPOTHESIS TESTING

Stating the Null and Alternative Hypothesis

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

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

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

HYPOTHESIS TESTING

Two–Tail Hypothesis Test

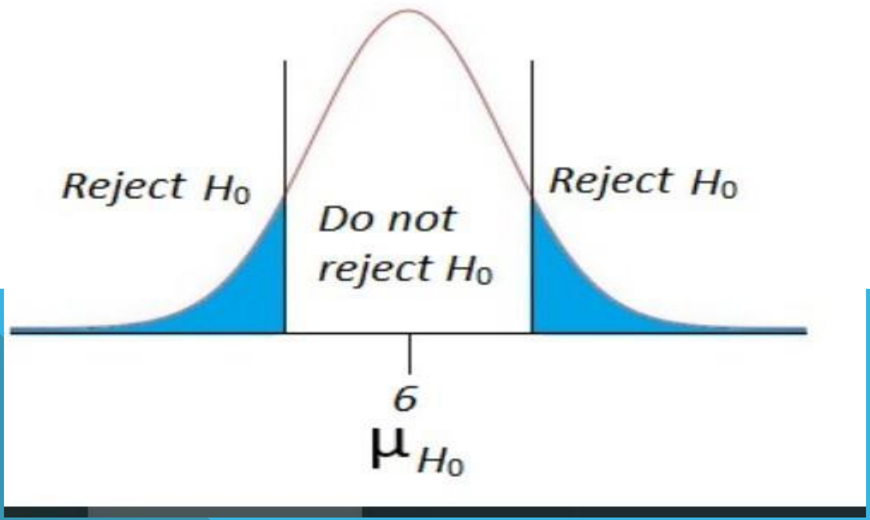
- Two-tail hypothesis test is used whenever the alternative hypothesis is stated as ≠
- The assignment example would require a two-tail test because the alternative hypothesis is stated as:

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

HYPOTHESIS TESTING

Two–Tail Hypothesis Test

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



HYPOTHESIS TESTING

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

HYPOTHESIS TESTING

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.

HYPOTHESIS TESTING

One Tail Hypothesis Test

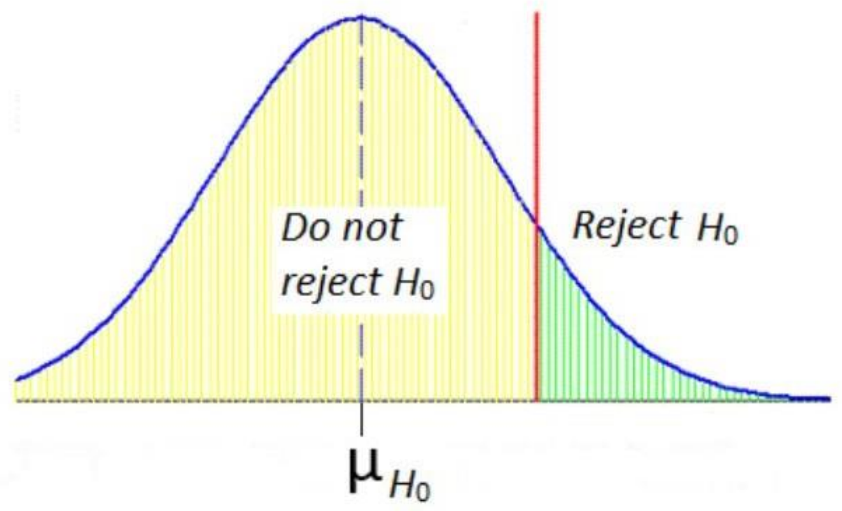
- One -tail hypothesis test is used whenever the alternative hypothesis is stated as < or >
- The golf example would require a one-tail test because the alternative hypothesis is expressed as:

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

HYPOTHESIS TESTING

One Tail Hypothesis Test

Test and plot the sample mean, which represents the average increase in variable value



Dr. S.R.Khonde

HYPOTHESIS TESTING

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 yellow region – we do not reject null hypothesis

If sample mean falls in shaded region – reject null hypothesis

HYPOTHESIS TESTING METHOD - CHI SQUARE (χ^2)

- useful for analysing 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.

HYPOTHESIS TESTING METHOD - CHI SQUARE (χ^2)

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.

HYPOTHESIS TESTING METHOD – CHI SQUARE GOODNESS OF fiT TEST (χ^2)

- Test statistic:

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

O_i= Frequency of Outcome (Original Frequency)

E= Expected Frequency

HYPOTHESIS TESTING METHOD – CHI SQUARE GOODNESS OF fiT TEST (χ^2)

a random sample of 10 bags

100 pieces of candy and five flavors

Hypothesis Testing method - Chi Square goodness of fit test (χ^2)

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.

HYPOTHESIS TESTING METHOD – CHI SQUARE GOODNESS OF fiT TEST (χ^2)

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

Expected Frequency

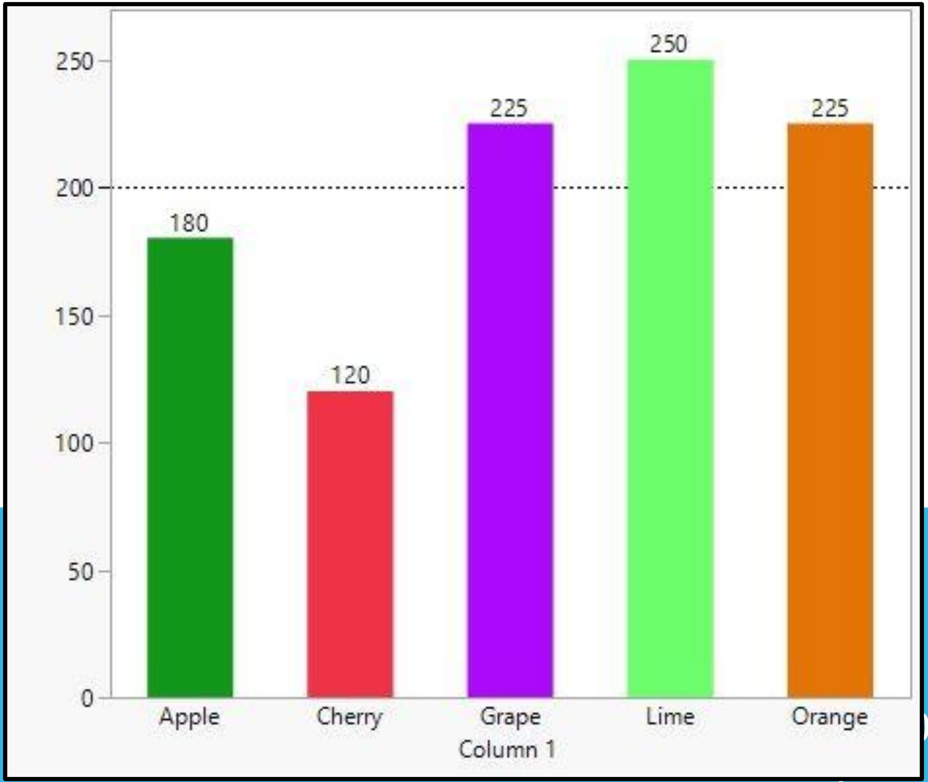
- Each bag has **100** pieces of candy.
- Each bag has **five** flavors of candy.
- We expect to have equal numbers for each flavor.
- This means we expect 100 / 5 = **20** pieces of candy in each flavor from each bag.
- For 10 bags in our sample, we expect **10 x 20 = 200** pieces of candy in each flavour

HYPOTHESIS TESTING METHOD – CHI SQUARE GOODNESS OF fiT TEST (χ^2)

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

Actual Frequency

Bar chart of counts of candy flavors from all 10 bags



Dr. S.R.Khonde

HYPOTHESIS TESTING METHOD – CHI SQUARE GOODNESS OF fiT TEST (χ^2)

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 α=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, α , to 0.05.

HYPOTHESIS TESTING METHOD – CHI SQUARE GOODNESS OF fiT TEST (χ^2)

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 (10 bags) Actual Frequency	Expected Number of Pieces of Candy
Apple	180	200
Lime	250	200
Cherry	120	200
Cherry	225	200
Grape	225	200

Dr. S.R.Khonde

3. Check the data for errors.

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

Flavor	Actual Number of Pieces of Candy (10 bags)	Expected Number of 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

Dr. S.R.Khonde

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:

$$2 + 12.5 + 32 + 3.125 + 3.125 = 52.75 \quad (\chi^2)$$

HYPOTHESIS TESTING METHOD – CHI SQUARE GOODNESS OF fiT TEST (χ^2)

4. Check the assumptions for the test

Based on χ^2	χ^2 (CALCULATED)< χ^2(TABLE)	no statistically significant difference, Ho can not be rejected,
	χ^2 (CALCULATED) > χ^2 (TABLE)	statistically significant difference, Ho is rejected
Based on P value	Pvalue_{table} > alpha= 0.05	no statistically significant difference, Ho can not be rejected.
	Pvalue_{table} < alpha= 0.05	statistically significant difference Ho is rejected

HYPOTHESIS TESTING METHOD – CHI SQUARE GOODNESS OF fiT TEST (χ^2)

4. Check the assumptions for the test

χ^2 Table

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
0.040	4.21	6.43	8.31	10.02	11.64
0.035	4.44	6.70	8.60	10.34	11.98
0.030	4.70	7.01	8.94	10.71	12.37
0.025	5.02	7.37	9.34	11.14	12.83
0.020	5.41	7.82	9.83	11.66	13.38
0.015	5.91	8.39	10.46	12.33	14.09
0.010	6.63	9.21	11.34	13.27	15.08
0.005	7.87	10.59	12.83	14.86	16.74
0.001	10.82	13.81	16.26	18.46	20.51
<0.001	>10.82	>13.81	>16.26	>18.46	>20.51

Right-tail area	df = 6	df = 7	df = 8	df = 9	df = 10
>0.100	<10.64	<12.01	<13.36	<14.68	<15.98
0.100	10.64	12.01	13.36	14.68	15.98
0.095	10.79	12.17	13.52	14.85	16.16
0.090	10.94	12.33	13.69	15.03	16.35
0.085	11.11	12.50	13.87	15.22	16.54
0.080	11.28	12.69	14.06	15.42	16.75
0.075	11.46	12.88	14.26	15.63	16.97
0.070	11.65	13.08	14.48	15.85	17.20
0.065	11.86	13.30	14.71	16.09	17.44
0.060	12.08	13.53	14.95	16.34	17.71
0.055	12.33	13.79	15.22	16.62	17.99
0.050	12.59	14.06	15.50	16.91	18.30
0.045	12.87	14.36	15.82	17.24	18.64
0.040	13.19	14.70	16.17	17.60	19.02
0.035	13.55	15.07	16.56	18.01	19.44
0.030	13.96	15.50	17.01	18.47	19.92
0.025	14.44	16.01	17.53	19.02	20.48
0.020	15.03	16.62	18.16	19.67	21.16
0.015	15.77	17.39	18.97	20.51	22.02
0.010	16.81	18.47	20.09	21.66	23.20
0.005	18.54	20.27	21.95	23.58	25.18
0.001	22.45	24.32	26.12	27.87	29.58
<0.001	>22.45	>24.32	>26.12	>27.87	>29.58

Dr. S.R.Khonde

HYPOTHESIS TESTING METHOD – CHI SQUARE GOODNESS OF fiT TEST (χ^2)

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 : area under the density curve of chi square
- χ^2 **= 52.75**
- **df = 4**
- **P Value: 0.00**

HYPOTHESIS TESTING METHOD – CHI SQUARE GOODNESS OF fiT TEST (χ^2)

4. Check the assumptions for the test **Degree of Freedom and P value**

α = 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

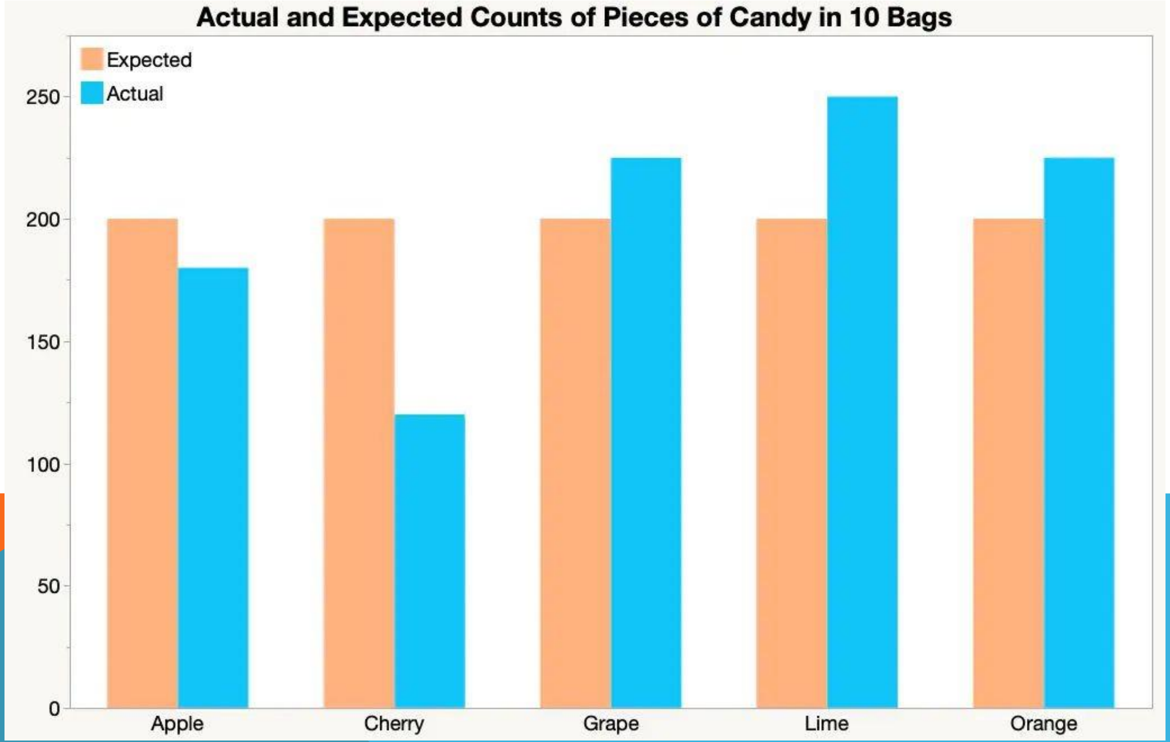
HYPOTHESIS TESTING METHOD – CHI SQUARE GOODNESS OF fiT TEST (χ^2)

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 (** $X^2_{calculated}$ > X^2_{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_{table} value > alpha)
- we reject the null hypothesis that the proportions of flavors of candy are equal

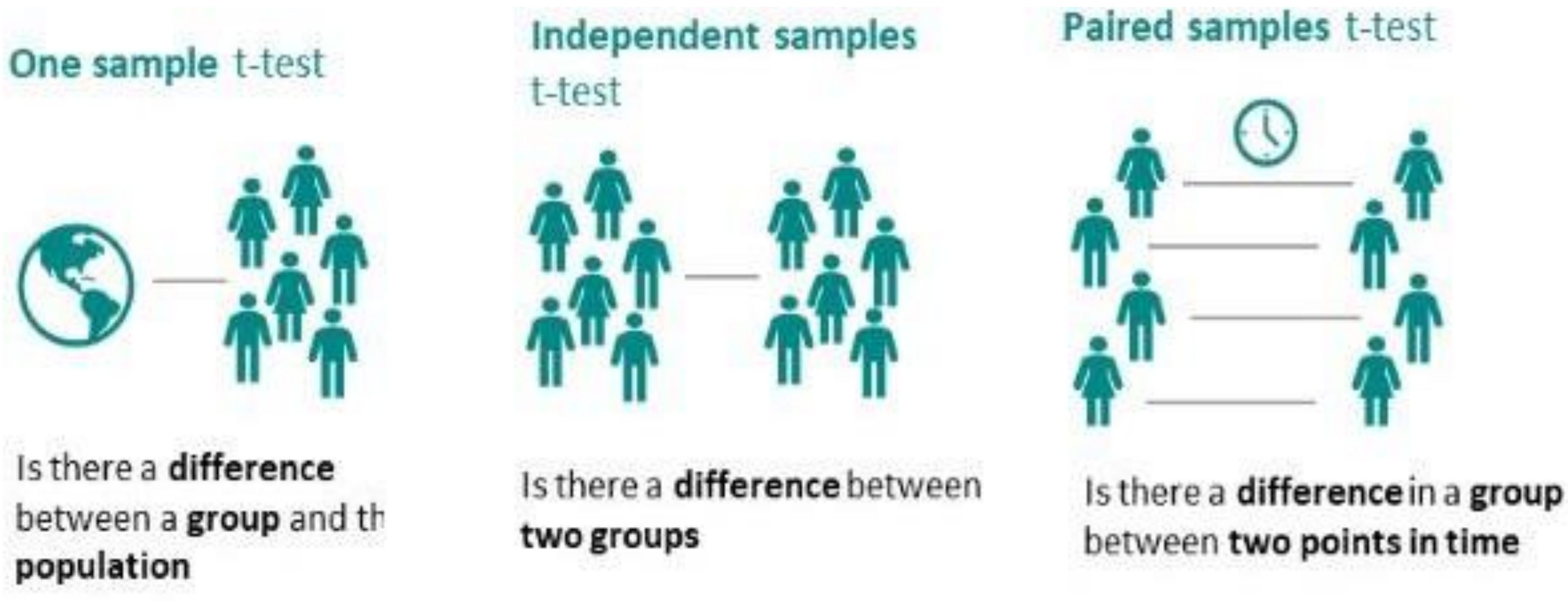
Interpretation of Results



HYPOTHESIS TESTING METHOD : 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).

HYPOTHESIS TESTING METHOD : T-TEST



HYPOTHESIS TESTING METHOD : T-TEST

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 α value and the degrees of freedom for your data.

HYPOTHESIS TESTING METHOD : T-TEST

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

HYPOTHESIS TESTING METHOD : ONE SAMPLE T-TEST

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 a perform t-test, we randomly collect the data of 10 girls with their marks

H_0: $\mu <= 600$

H_1: $\mu > 600$

HYPOTHESIS TESTING METHOD : ONE SAMPLE T-TEST

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

Girls_Score
587
602
627
610
619
622
605
608
596
592

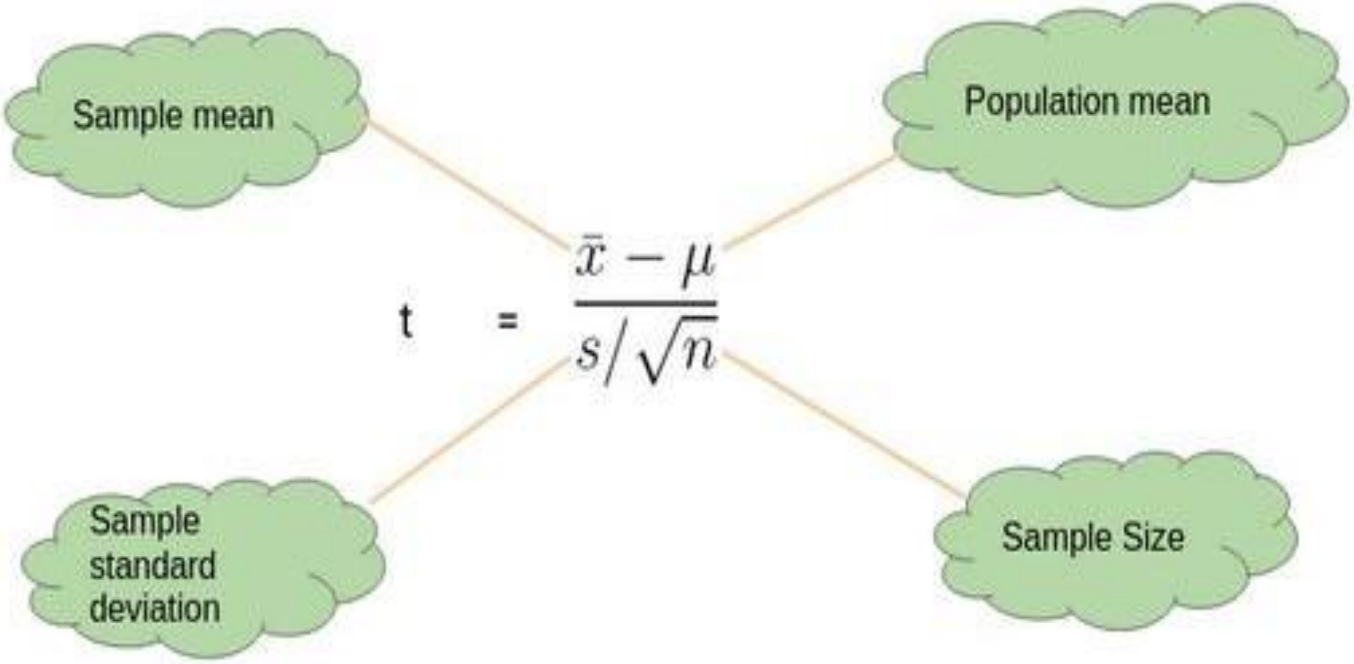
HYPOTHESIS TESTING METHOD : ONE SAMPLE T-TEST

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 α=0.05.

α=0.05.

HYPOTHESIS TESTING METHOD : ONE SAMPLE T-TEST

3. Check the data for errors.



HYPOTHESIS TESTING METHOD : ONE SAMPLE T-TEST

3. Check the data for errors.

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

HYPOTHESIS TESTING METHOD : ONE SAMPLE T-TEST

3. Check the data for errors.

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$= \frac{606.8 - 600}{13.14/\sqrt{10}}$$

$$= 1.64$$

HYPOTHESIS TESTING METHOD - CHI SQUARE GOODNESS OF fiT TEST (χ^2)

4. Check the assumptions for the test

Based on t score	t $score_{calculated}$ < t $score_{table}$	no statistically significant difference, Ho can not be rejected,
	t $score_{calculated}$ > t $score_{table}$	statistically significant difference, Ho is rejected
Based on P value	$Pvalue_{table}$ > alpha= 0.05	no statistically significant difference, Ho can not be rejected.
	$Pvalue_{table}$ < alpha= 0.05	statistically significant difference Ho is rejected

Dr. S.R.Khonde

HYPOTHESIS TESTING METHOD - ONE SAMPLE T-TEST

4. Check the assumptions for the test Degree of Freedom and P value

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

df= 10-1=9

- t score (calculated) = 1.64
- t score (table) = 1.833
- df = 9
- PValue: 0.06

Dr. S.R.Khonde

HYPOTHESIS TESTING METHOD - ONE SAMPLE T-TEST

5. Perform the test and draw your conclusion.

- The value of tscore is 1.64
- Since 1.64 <1.83
- 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
- we cannot reject the null hypothesis.

HYPOTHESIS TESTING METHOD : T-TEST

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)

HYPOTHESIS TESTING METHOD : TWO SAMPLE T-TEST

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: $\mu_1 - \mu_2 <= 15$

H_1: $\mu_1 - \mu_2 > 15$

HYPOTHESIS TESTING METHOD : TWO SAMPLE T-TEST

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

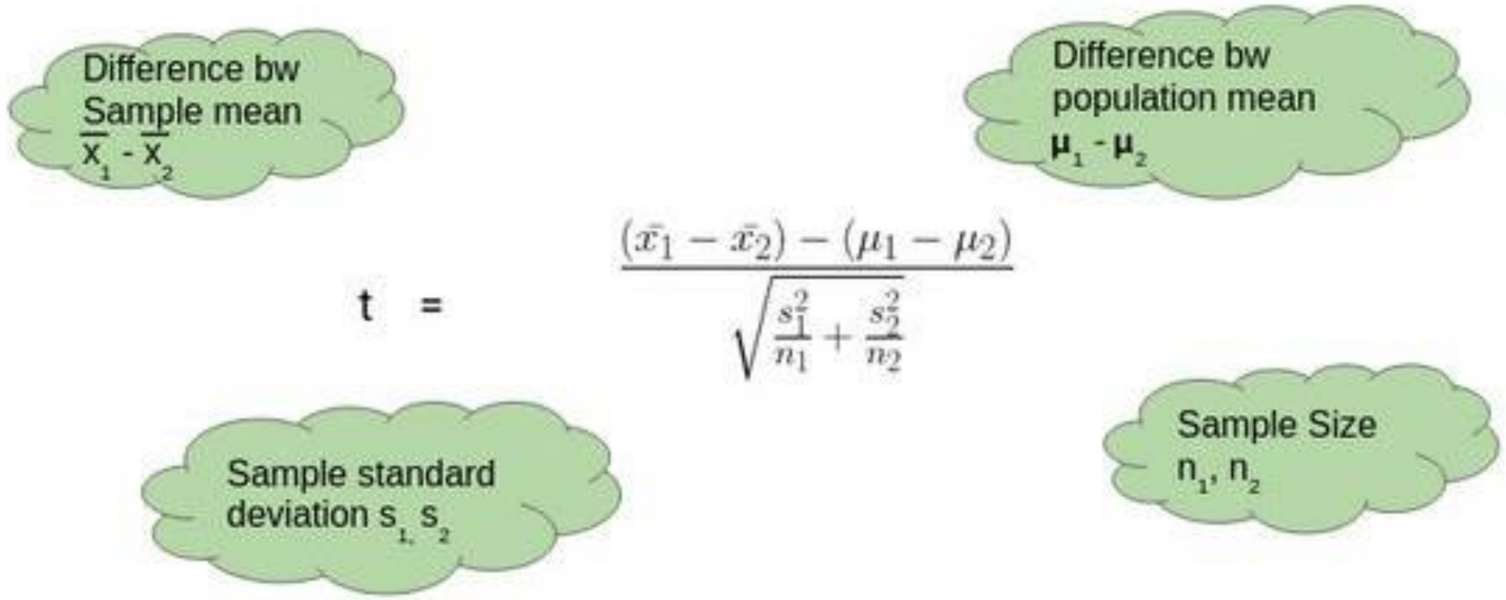
HYPOTHESIS TESTING METHOD : ONE SAMPLE T-TEST

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 α=0.05.

α=0.05.

HYPOTHESIS TESTING METHOD : TWO SAMPLE T-TEST

3. Check the data for errors.



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

HYPOTHESIS TESTING METHOD : TWO SAMPLE T-TEST

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

HYPOTHESIS TESTING METHOD : TWO SAMPLE T-TEST

3. Check the data for errors.

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

$$\frac{(630.1 - 606.8) - (15)}{\sqrt{\frac{(13.42)^2}{10} + \frac{(13.14)^2}{10}}}$$

$$= 1.833$$

Dr. S.R.Khonde

HYPOTHESIS TESTING METHOD - TWO SAMPLE T-TEST

4. Check the assumptions for the test

Based on t score	$t\ score_{calculated} < t\ score_{table}$	no statistically significant difference, Ho can not be rejected,
	$t\ score_{calculated} > t\ score_{table}$	statistically significant difference, Ho is rejected
Based on P value	$Pvalue_{table} > alpha = 0.05$	no statistically significant difference, Ho can not be rejected.
	$Pvalue_{table} < alpha = 0.05$	statistically significant difference Ho is rejected

Dr. S.R.Khonde

HYPOTHESIS TESTING METHOD - TWO SAMPLE T-TEST

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 n1– 1 and n2– 1.

df= (10-1)+(10-1)=18

- t score (calculated) = 1.833
- t score (table) = 1.73
- df= 18
- P Value: 0.041

HYPOTHESIS TESTING METHOD – TWO SAMPLE T-TEST

5. Perform the test and draw your conclusion.

- The value of tscore is 1.64
- Since 1.833 > 1.73
- 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 value < alpha= 0.05)
- We reject the null hypothesis.

Paired t Tests

H_0: $\mu_{before} = \mu_{after}$
H_a: $\mu_{before} \neq \mu_{after}$

$$t = \frac{\bar{d}}{s/\sqrt{n}}$$

Patient	Before	After	difference
1	120	122	-2
2	122	120	2
3	143	141	2
4	100	109	-9
5	109	109	0

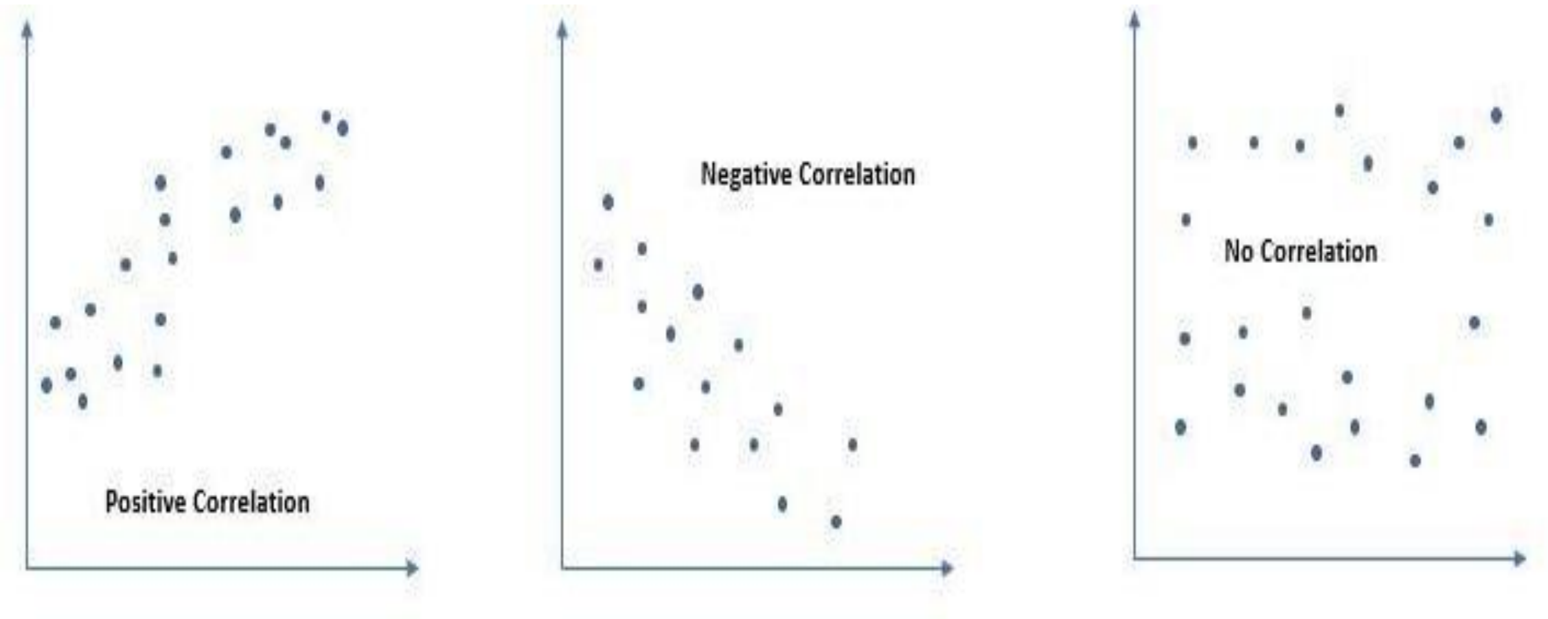
- Example: Before and after medicine BP was measured. Is there a **difference** at 95% confidence level?

- $\bar{d}$ = -1.4 , s = 4.56 , n = 5
- $t_{cal.}$ = 1.4/2.04 = - 0.69

CORRELATION

- 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 ± 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.

CORRELATION



PEARSON 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

y_i = y variable sample

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

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

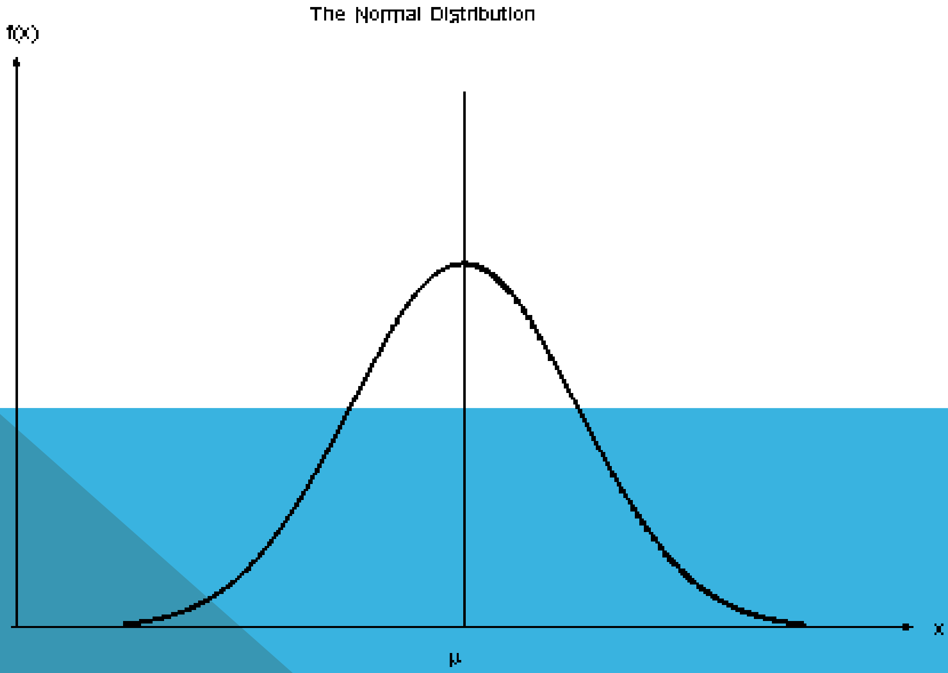
WHY CORRELATION?

- 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?

ASSUMPTIONS FOR A PEARSON CORRELATION:

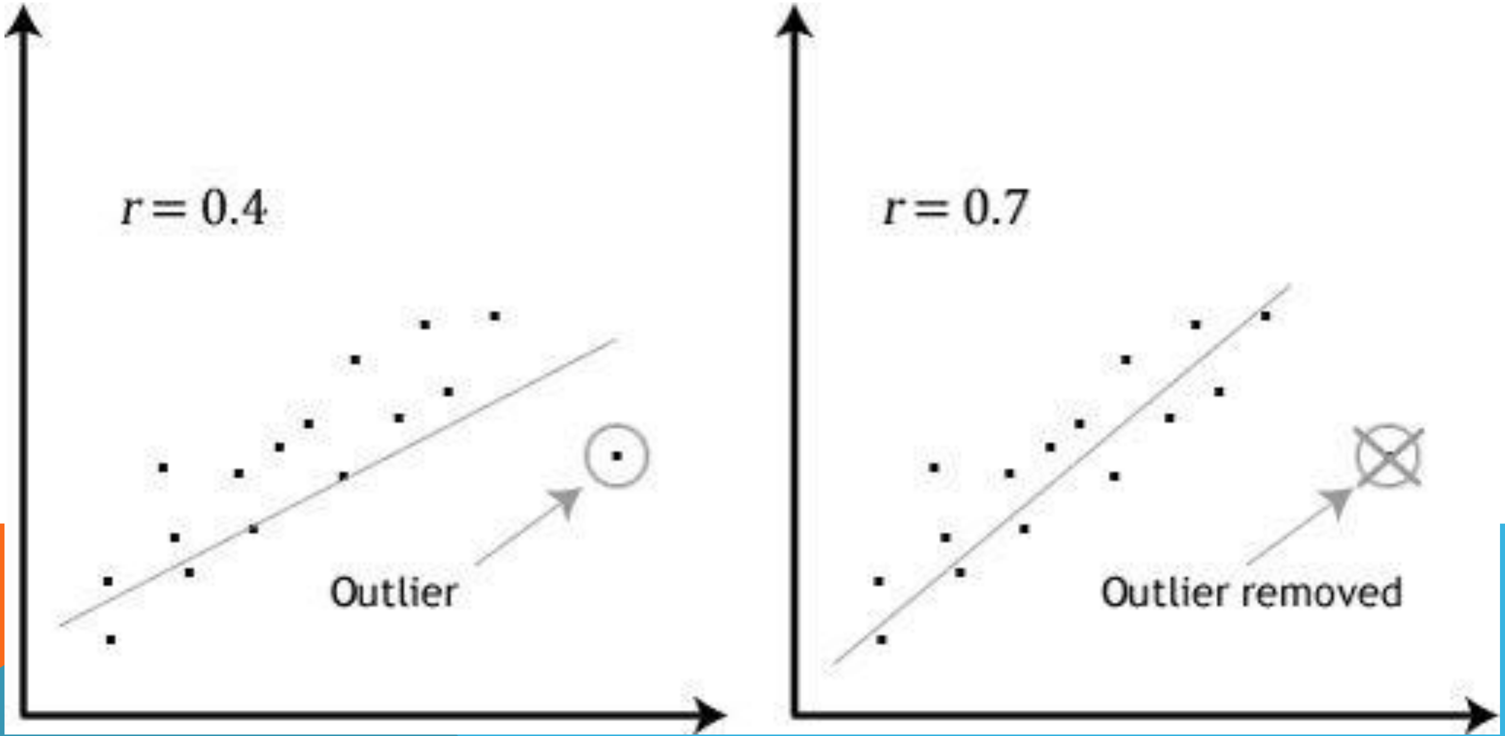
1. Both variables should be normally distributed.

   This is sometimes called the 'Bell Curve' or the 'Gaussian Curve'



ASSUMPTIONS FOR A PEARSON CORRELATION:

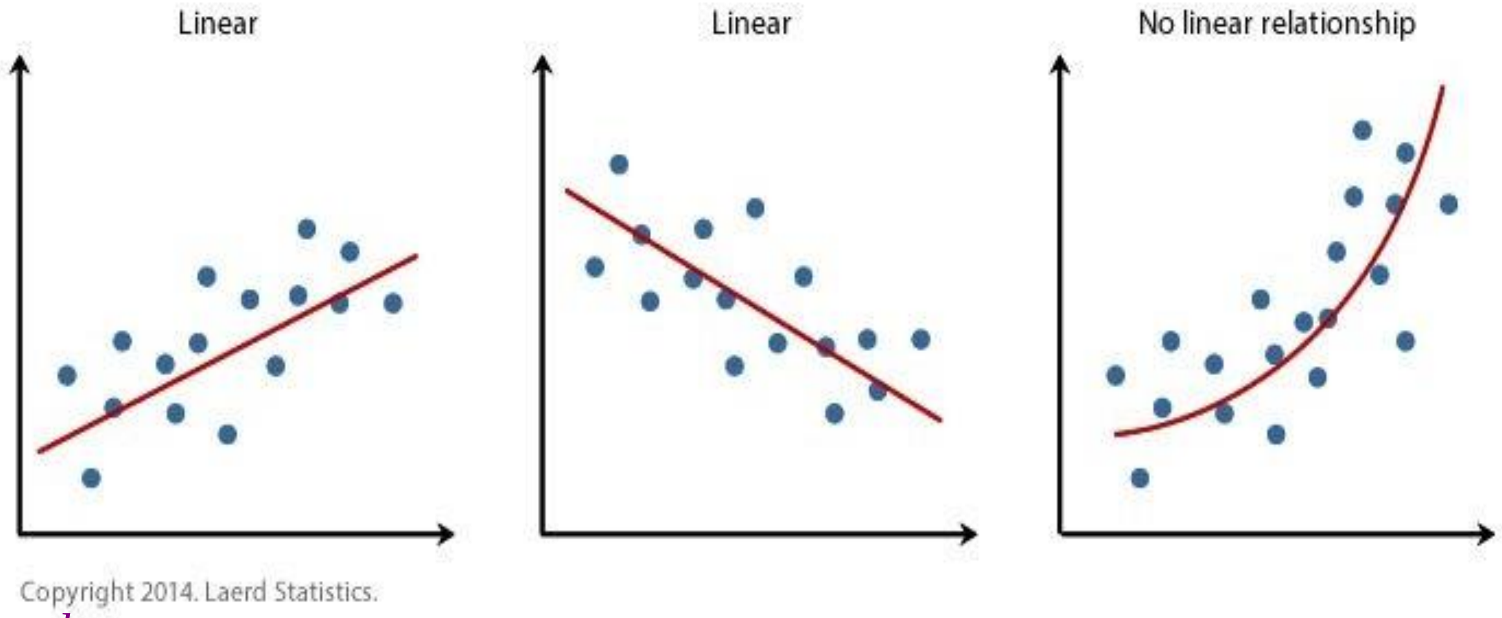
2. There should be no significant outliers



ASSUMPTIONS FOR A PEARSON CORRELATION:

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



ASSUMPTIONS FOR A PEARSON CORRELATION:

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.

PEARSON CORRELATION EXAMPLE

	A	B	C	D	E	F	G	H
5								
6	Hours Played Sport	Test Score	$x - \bar{x}$	$y - \bar{y}$	$(xi - \bar{x})^2$	$(yi - \bar{y})^2$	$(xi - \bar{x})*(yi - \bar{y})$	
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								

PEARSON CORRELATION EXAMPLE

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

PEARSON CORRELATION EXAMPLE

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

Dr. S.R.Khonde

PEARSON CORRELATION EXAMPLE

- 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.

END
of
UNIT II