Biostatistics

Measures of central tendency

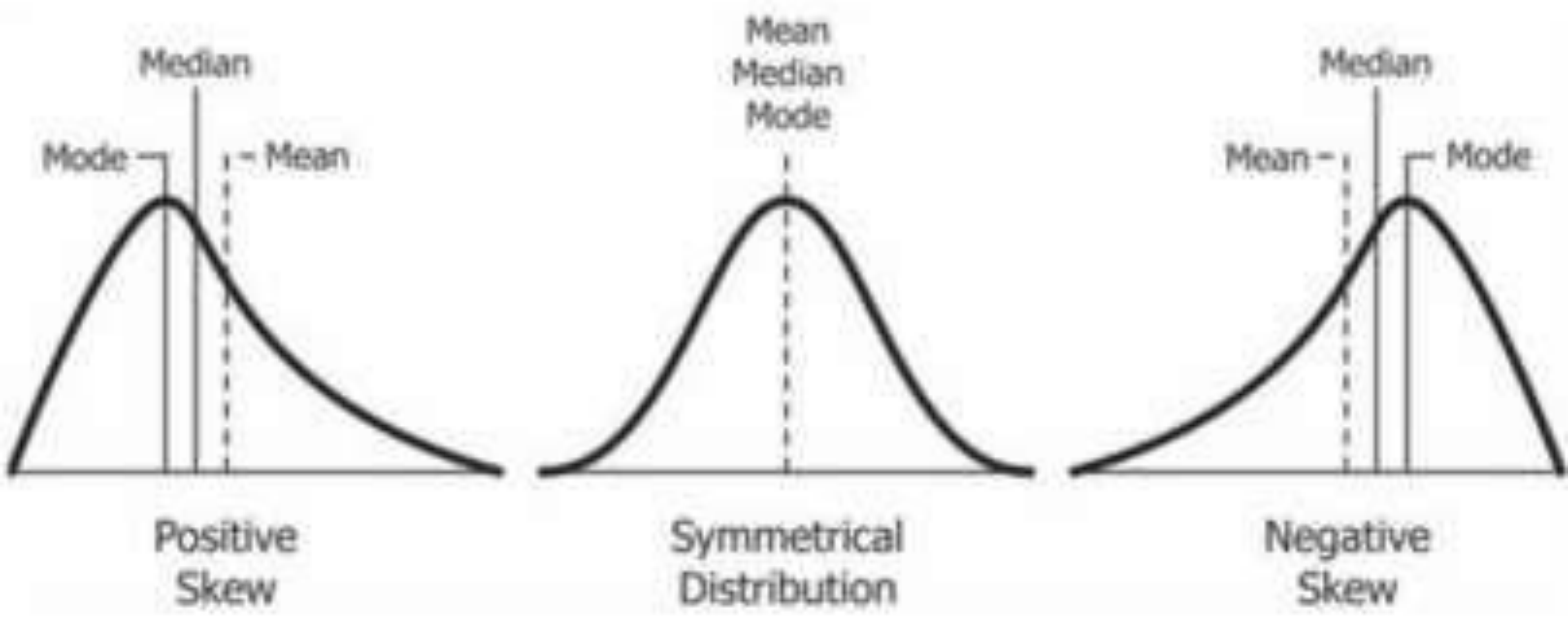
Subject Code: 22ZOOC23

WHAT IS NORMAL DISTRIBUTION?

- In a normal distribution, data is symmetrically distributed with no **skew**.
- When plotted on a graph, the data follows a bell shape.
- Most values clustering around a central region and tapering off as they go further away from the center.
- Normal distributions are also called Gaussian distributions or bell curves because of their shape.

What is Skewness ?

- **Skewness** is a measure of the asymmetry of a distribution.
- A distribution is asymmetrical when its left and right side are not mirror images.
- A distribution can have right (or positive), left (or negative), or zero skewness.
- A right-skewed distribution is longer on the right side of its peak.
- A left-skewed distribution is longer on the left side of its peak.


Median
Mode
Mean
Mean
Median
Mode
Median
Mean
Mode
Positive Skew
Symmetrical Distribution
Negative Skew

Measures of central tendency

- **Measures of central tendency** help you find the middle, or the average, of a dataset.
- The 3 most common measures of central tendency are the mode, median, and mean.
- **Mode**: the most frequent value.
- **Median**: the middle number in an ordered dataset.
- **Mean**: the sum of all values divided by the total number of values.

Mode

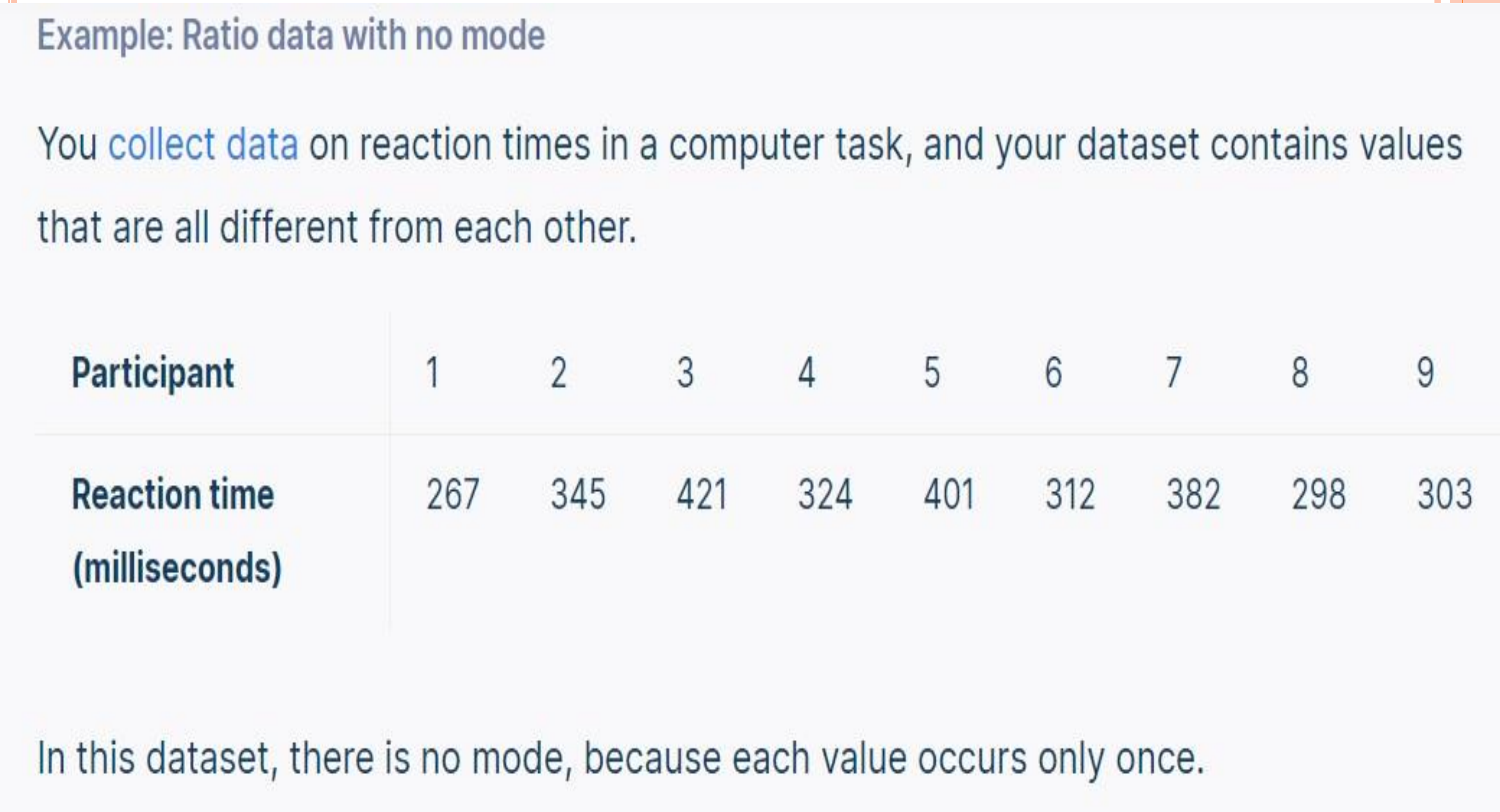
- The **mode** is the most frequently occurring value in the dataset.
- It’s possible to have no mode, one mode, or more than one mode.
- To find the mode, sort your dataset numerically or categorically.
- Select the response that occurs most frequently.
- Mode describes qualitative data.

- Example: Finding the mode in a survey, you ask 9 participants whether they identify as conservative, moderate, or liberal.
- Frequency table to count up the values for each category.

Political ideology	**Frequency**
Conservative	2
Moderate	3
Liberal	4

- **Mode: Liberal**
- The mode is most applicable to data from a nominal level of measurement.

- Nominal data is classified into mutually exclusive categories, so the mode tells you the most popular category.
- For continuous variables or ratio levels of measurement.
- The mode may not be a helpful measure of central tendency.
- There are many more possible values than there are in a nominal or ordinal level of measurement.
- It’s unlikely for a value to repeat in a ratio level of measurement.



Example: Ratio data with no mode

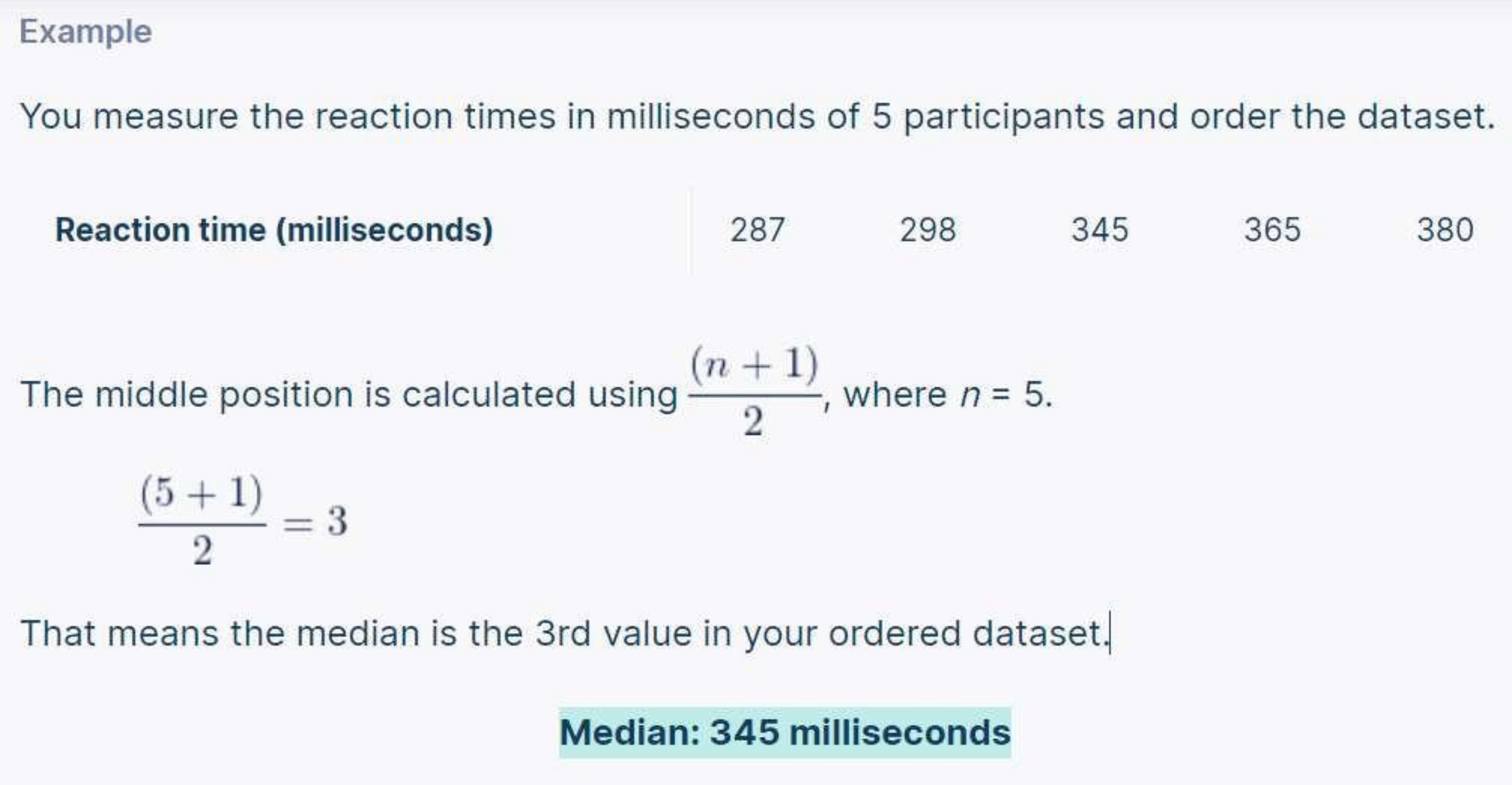
You collect data on reaction times in a computer task, and your dataset contains values that are all different from each other.

Participant	1	2	3	4	5	6	7	8	9
Reaction time (milliseconds)	267	345	421	324	401	312	382	298	303

In this dataset, there is no mode, because each value occurs only once.

Median

- The median of a dataset is the value that's **exactly in the middle** when it is ordered from low to high.
- Median of a finite set of values which divides the set into two equal parts.
- **Median of an odd-numbered dataset**
  - For an odd-numbered dataset, find the value that lies at the $\left[\frac{n+1}{2}\right]$ position, where *n* is the number of values in the dataset.



Example

You measure the reaction times in milliseconds of 5 participants and order the dataset.

Reaction time (milliseconds)	287	298	345	365	380

The middle position is calculated using $\frac{(n+1)}{2}$, where *n* = 5.

$$\frac{(5+1)}{2} = 3$$

That means the median is the 3rd value in your ordered dataset.

Median: 345 milliseconds

- **Even Number of Observations**
- If the total number of observation is even, then the median formula is:

$$Median = \frac{(\frac{n}{2})^{th}\, term + (\frac{n}{2}+1)^{th}\, term}{2}$$

- where n is the number of observations
- For a set there is only one median for data.

Example

You measure the reaction times of 6 participants and order the dataset.

Reaction time (milliseconds)	287	298	345	357	365	380

The middle positions are calculated using $\frac{n}{2}$ and $(\frac{n}{2})+1$, where *n* = 6.

$$\frac{6}{2}=3$$

$$(\frac{6}{2})+1=4$$

That means the middle values are the 3rd value, which is **345**, and the 4th value, which is **357**

To get the median, take the mean of the 2 middle values by adding them together and dividing by 2.

$$\frac{(345+357)}{2}=351$$

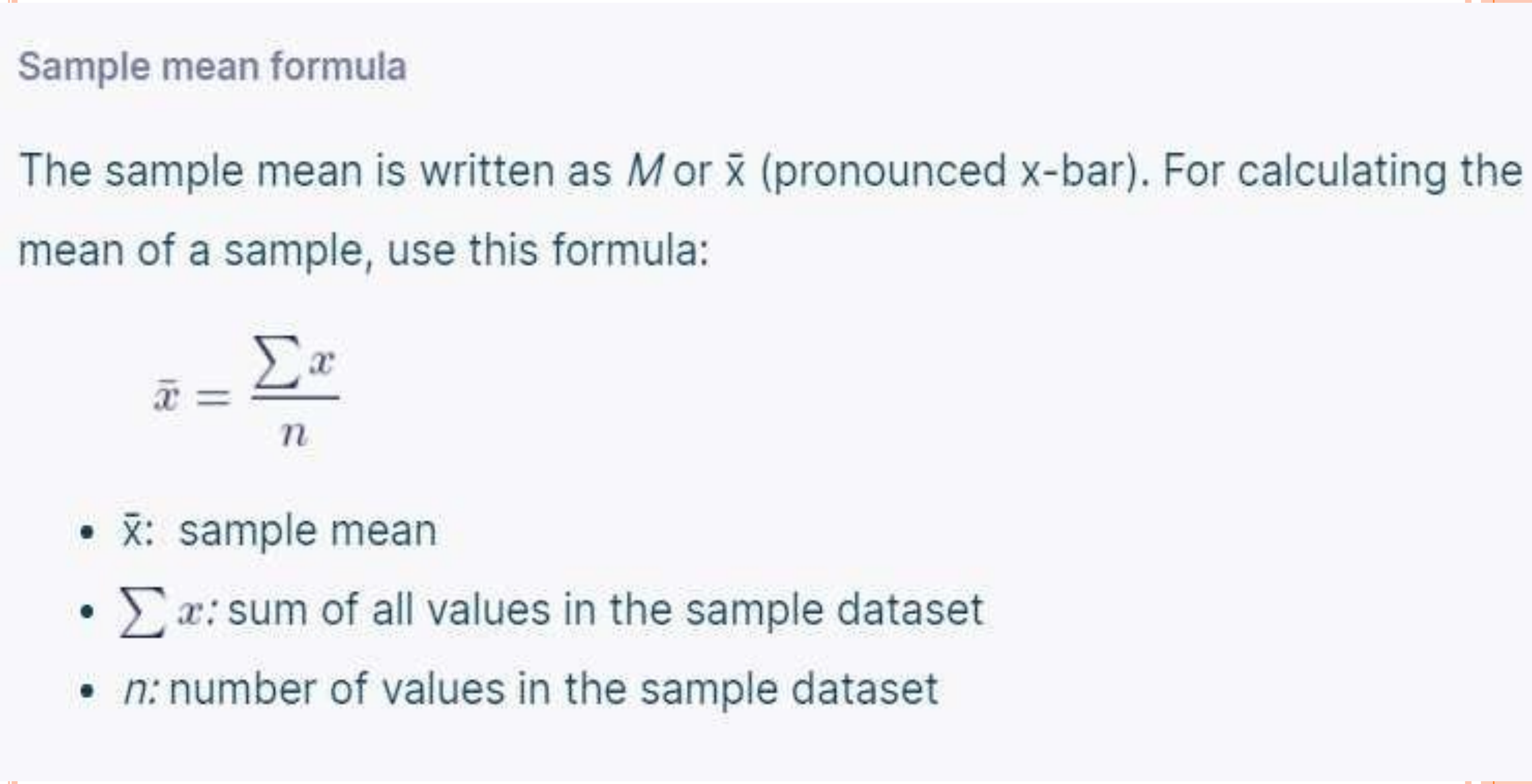
Median: 351 milliseconds

Mean

- The arithmetic mean of a dataset (which is different from the geometric mean) is the sum of all values divided by the total number of values.
- It's the most commonly used measure of central tendency because all values are used in the calculation.
- It is the descriptive measure of the average
- The formula for calculating the arithmetic mean is,
  - Arithmetic Mean ($\bar{x}$) = Sum of all observations / Number of observations

Population versus sample mean

- A dataset contains values from a sample or a population.
- A population is the entire group, while a sample is only a subset of that population.
- While data from a sample can help estimates about a population
- But only full population data can give you the complete picture.
- In statistics, the notation of a sample mean and a population mean and their formulas are different.
- But the procedures for calculating the population and sample means are the same.

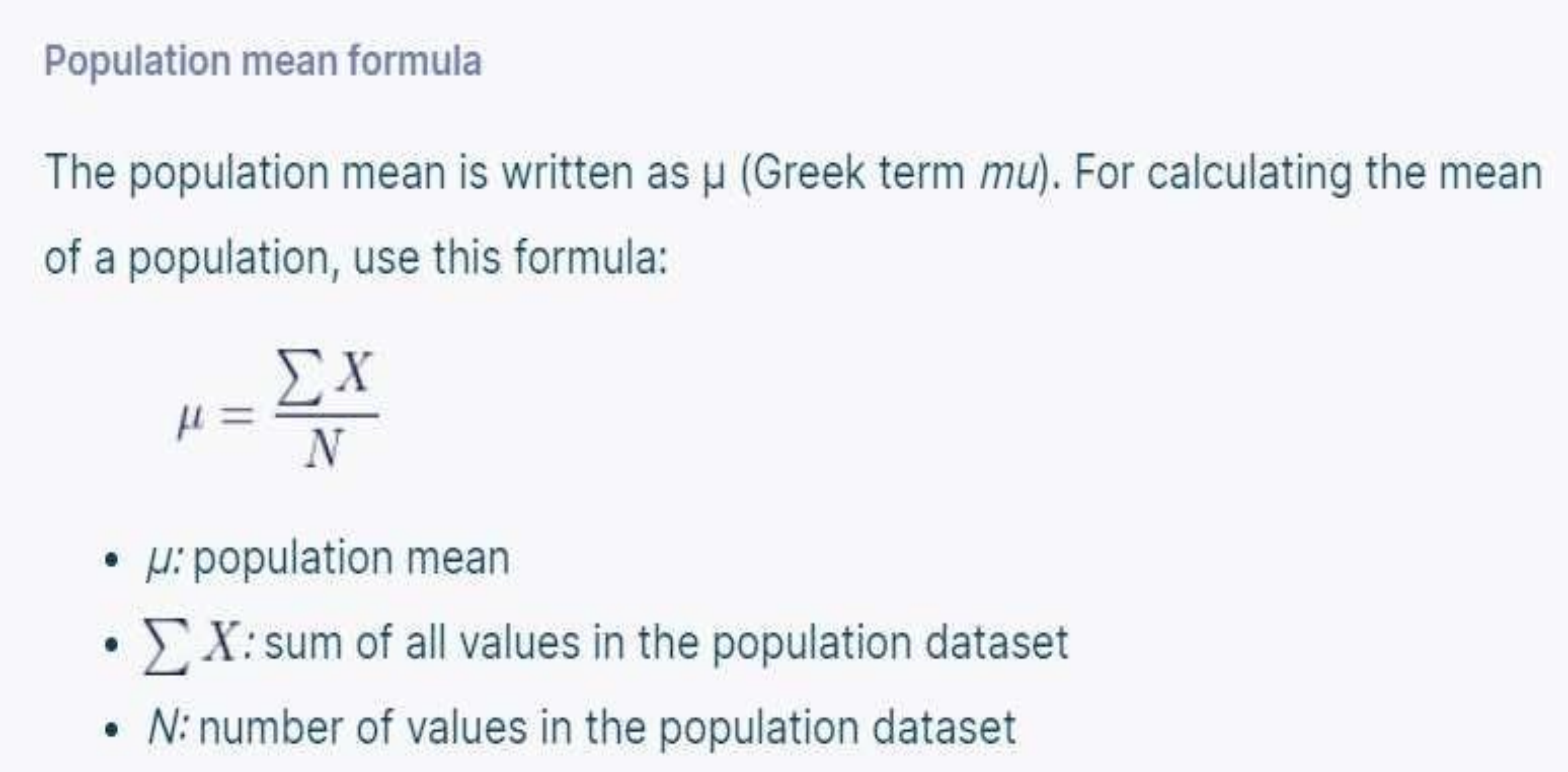


Sample mean formula

The sample mean is written as *M* or x̄ (pronounced x-bar). For calculating the mean of a sample, use this formula:

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

- x̄: sample mean
- $\sum x$: sum of all values in the sample dataset
- *n:* number of values in the sample dataset



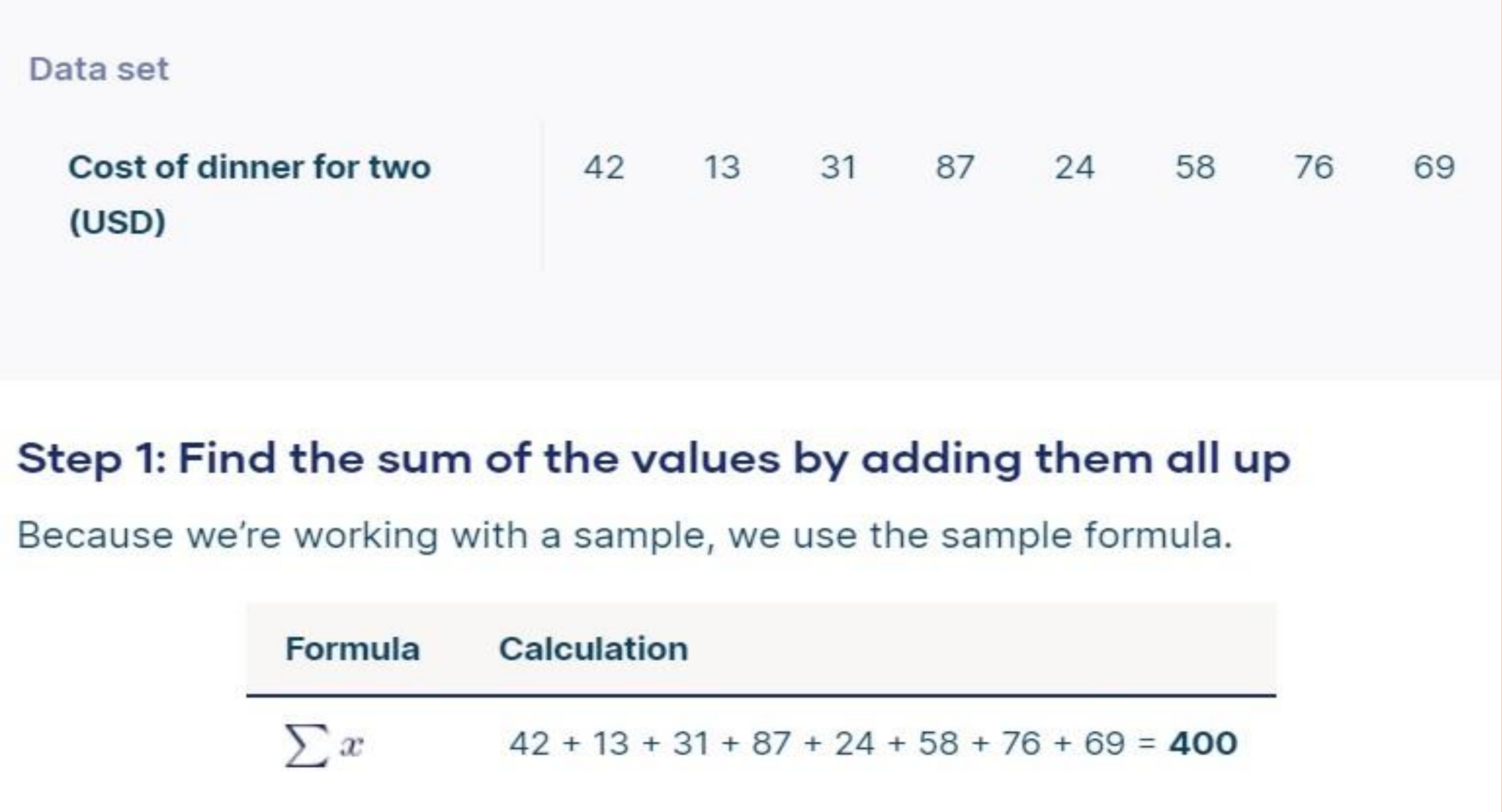
Population mean formula

The population mean is written as μ (Greek term *mu*). For calculating the mean of a population, use this formula:

$$\mu = \frac{\sum X}{N}$$

- μ: population mean
- $\sum X$: sum of all values in the population dataset
- N: number of values in the population dataset

- Sample of 8 neighbors how much they spent the last time they went out for dinner, and find the mean cost.



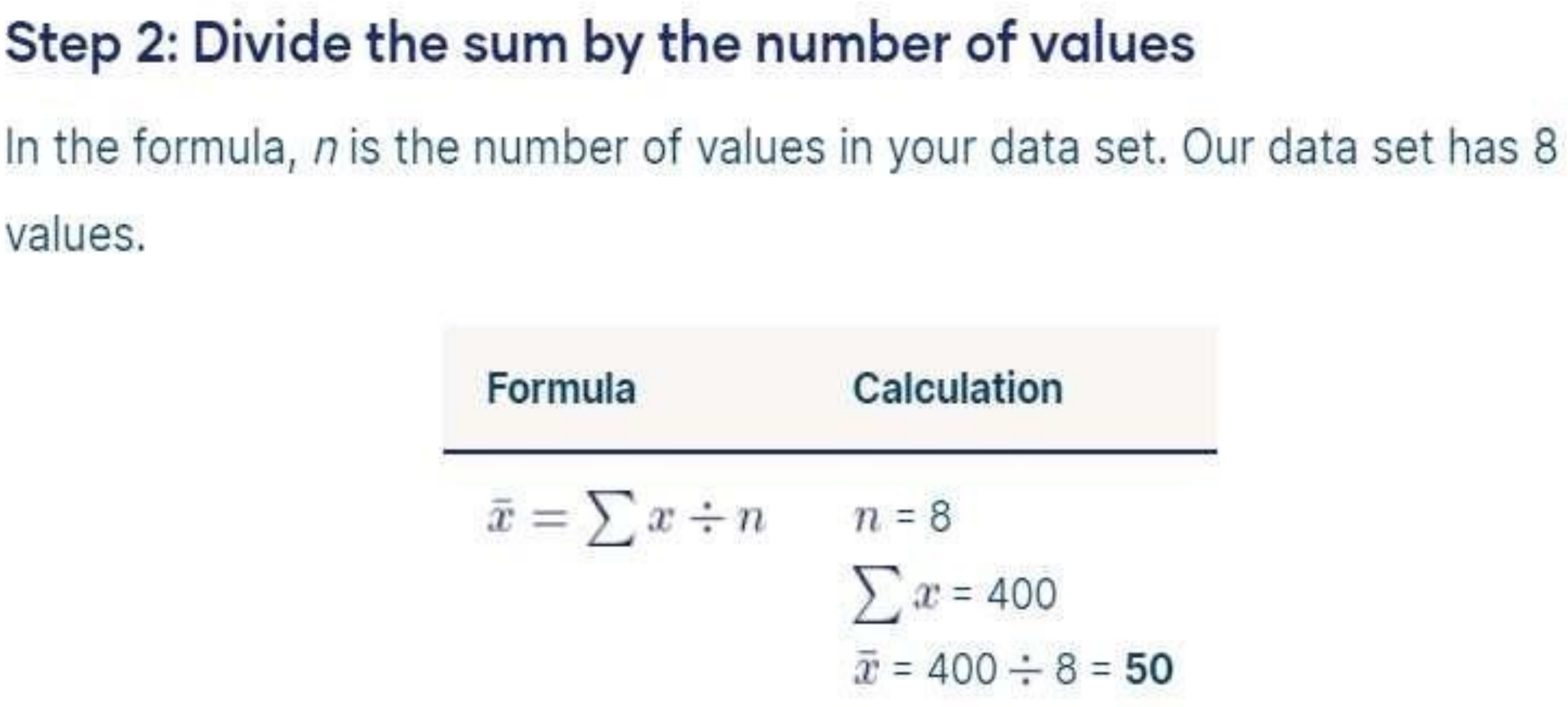
Data set

Cost of dinner for two (USD)	42	13	31	87	24	58	76	69

Step 1: Find the sum of the values by adding them all up

Because we're working with a sample, we use the sample formula.

Formula	Calculation
$\sum x$	42 + 13 + 31 + 87 + 24 + 58 + 76 + 69 = **400**



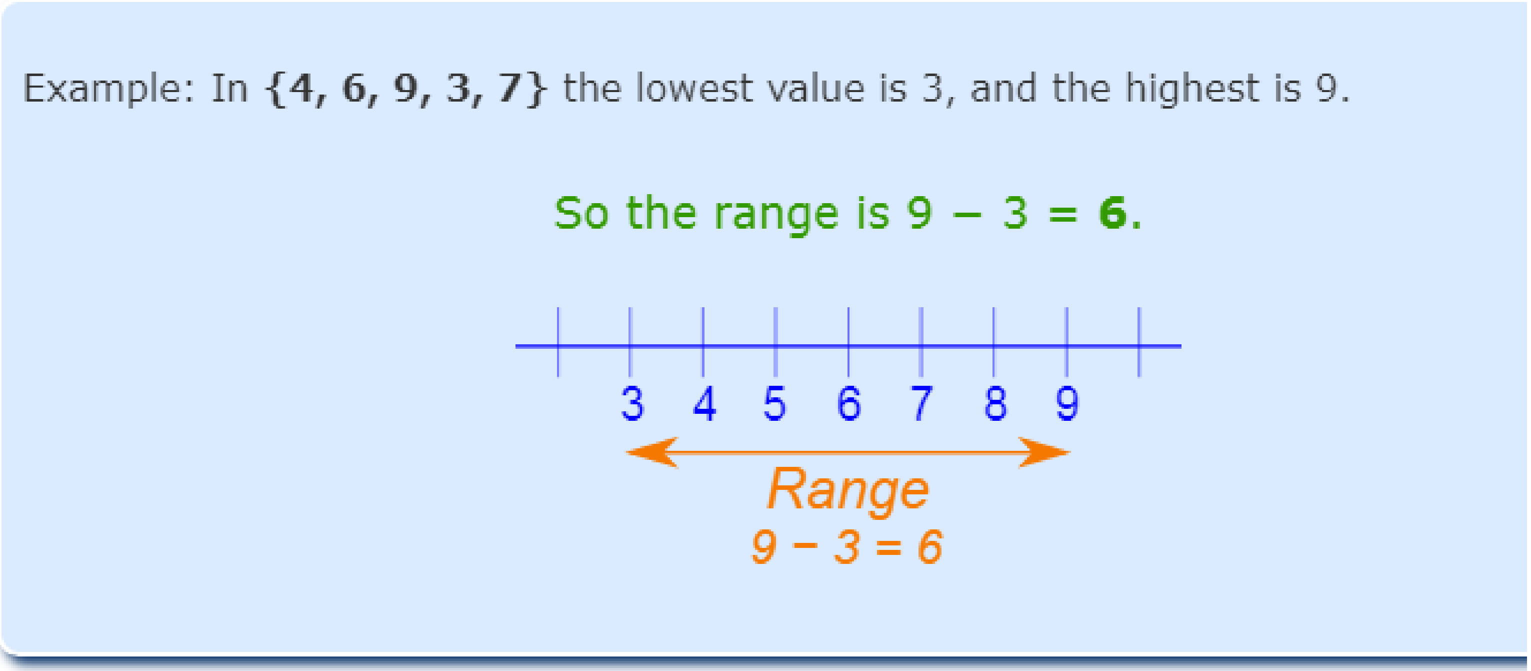
Step 2: Divide the sum by the number of values

In the formula, *n* is the number of values in your data set. Our data set has 8 values.

Formula	Calculation
$\bar{x} = \sum x \div n$	n = 8 $\sum x$ = 400 $\bar{x}$ = 400 ÷ 8 = **50**

Range

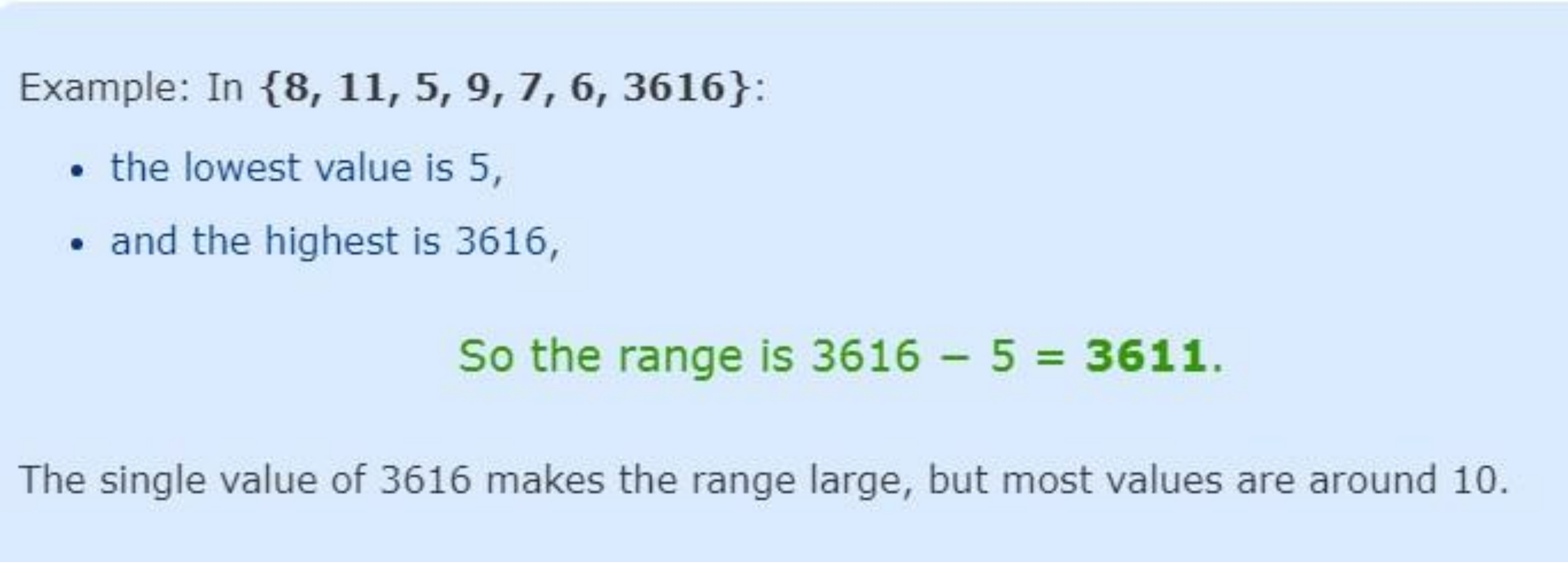
- The Range is the difference between the lowest and highest values.



Example: In **{4, 6, 9, 3, 7}** the lowest value is 3, and the highest is 9.

So the range is 9 − 3 = **6**.

- The range can sometimes be misleading when there are extremely high or low values.



Example: In **{8, 11, 5, 9, 7, 6, 3616}**:

- the lowest value is 5,
- and the highest is 3616,

So the range is 3616 − 5 = **3611**.

The single value of 3616 makes the range large, but most values are around 10.

Representation of data

Discuss on

Introduction

Data and its types

Methods of data presentation

Introduction

Data are a set of facts.

- Data - purpose of collection.
- Collected data - utilization.
- Information the data are conveying and how the data can be used.

Data are available in a raw format

—> summarized

—> organized

—> analyzed

Each data set needs to be presented in a certain way depending on what it is used for.

Planning how the data will be presented is essential before appropriately processing raw data.

Aim: Roles and appropriate use of text, tables, and graphs (graphs, plots, or charts), which are commonly used in reports, articles, posters, and presentations.

Line graph with whiskers

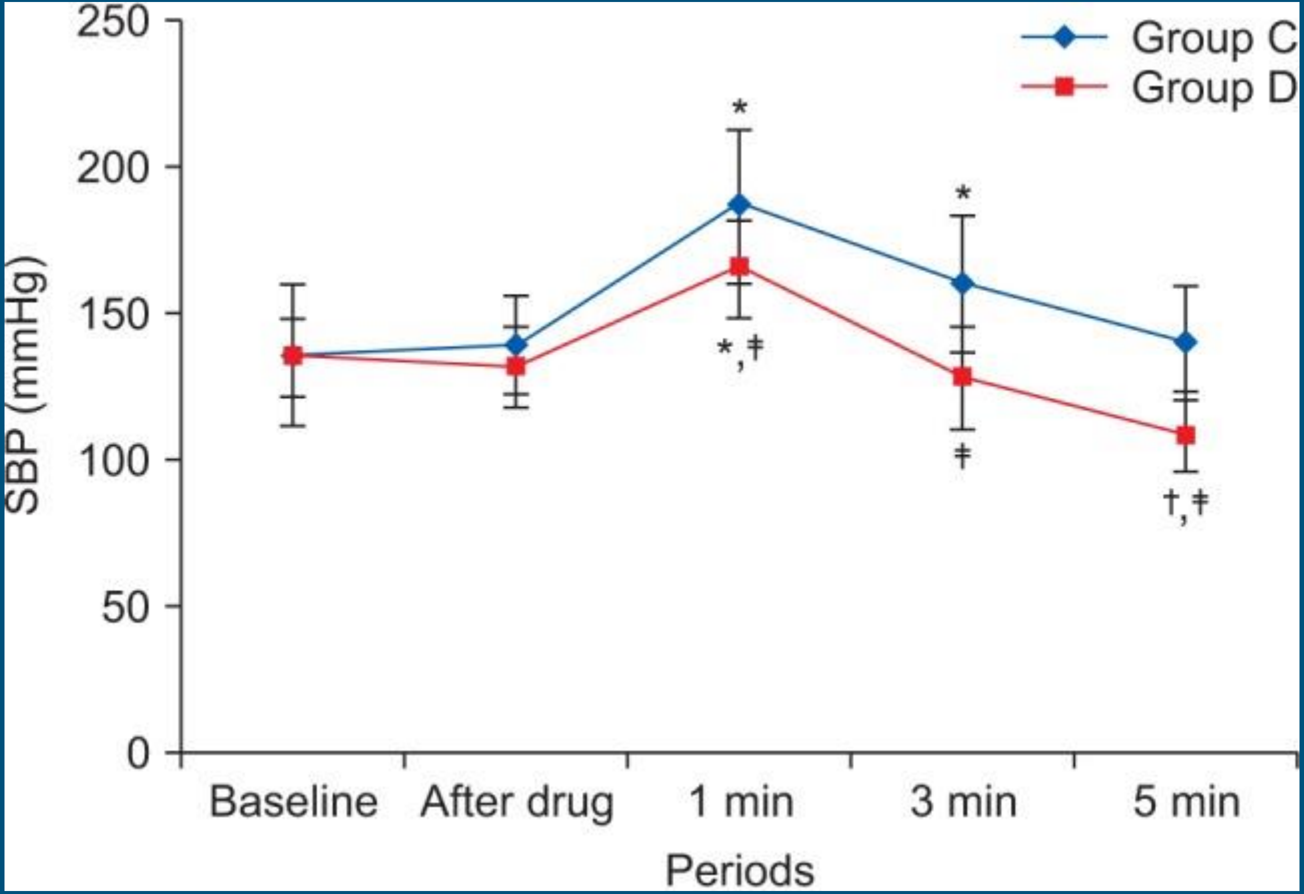
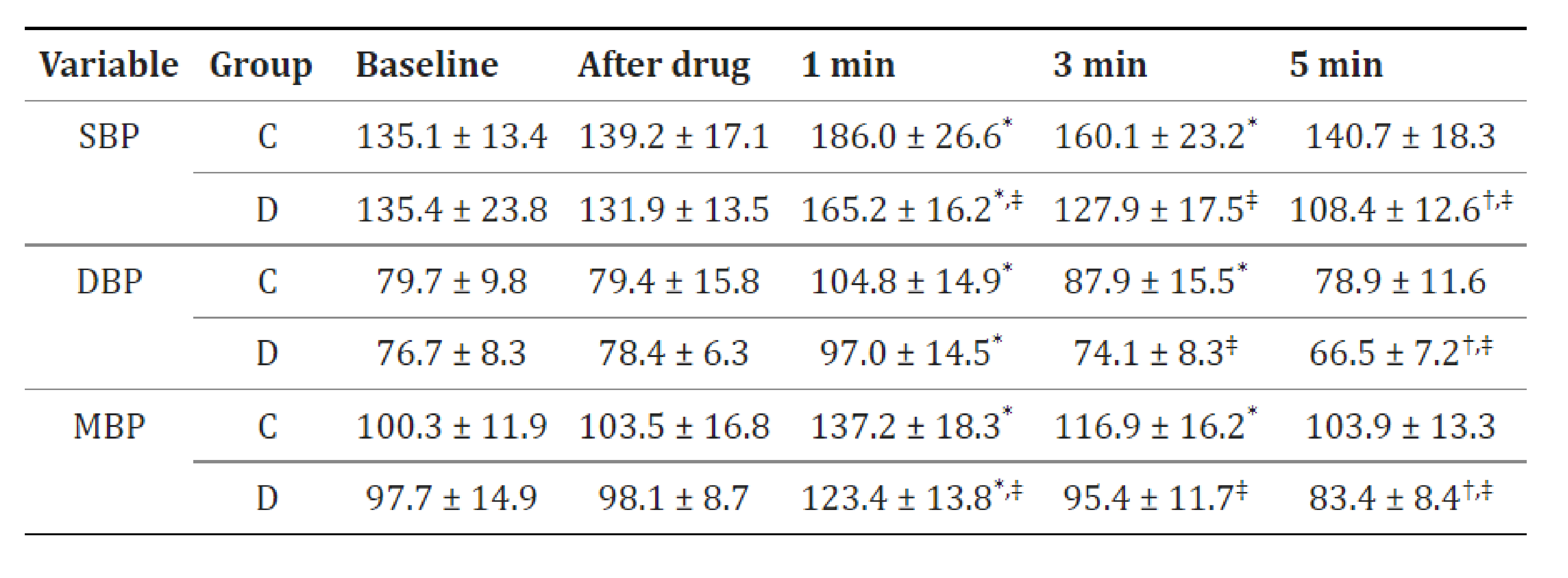


Table - Values are expressed as mean ± SD



Variable	Group	Baseline	After drug	1 min	3 min	5 min
SBP	C	135.1 ± 13.4	139.2 ± 17.1	186.0 ± 26.6*	160.1 ± 23.2*	140.7 ± 18.3
	D	135.4 ± 23.8	131.9 ± 13.5	165.2 ± 16.2*,‡	127.9 ± 17.5‡	108.4 ± 12.6†,‡
DBP	C	79.7 ± 9.8	79.4 ± 15.8	104.8 ± 14.9*	87.9 ± 15.5*	78.9 ± 11.6
	D	76.7 ± 8.3	78.4 ± 6.3	97.0 ± 14.5*	74.1 ± 8.3‡	66.5 ± 7.2†,‡
MBP	C	100.3 ± 11.9	103.5 ± 16.8	137.2 ± 18.3*	116.9 ± 16.2*	103.9 ± 13.3
	D	97.7 ± 14.9	98.1 ± 8.7	123.4 ± 13.8*,‡	95.4 ± 11.7‡	83.4 ± 8.4†,‡

Advantages of diagrams and graphs:

- Attractive and create lasting impression
- Make comparisons simple
- Forecasting using available data
- Can understand relations between variables
- Location of descriptive statistical measures is possible.

Types of diagrams:

—> Bar diagrams (Qualitative data) - Simple, Multiple, Subdivided, percentage

—> Pie diagram

—> Pictograms

—> Cartograms

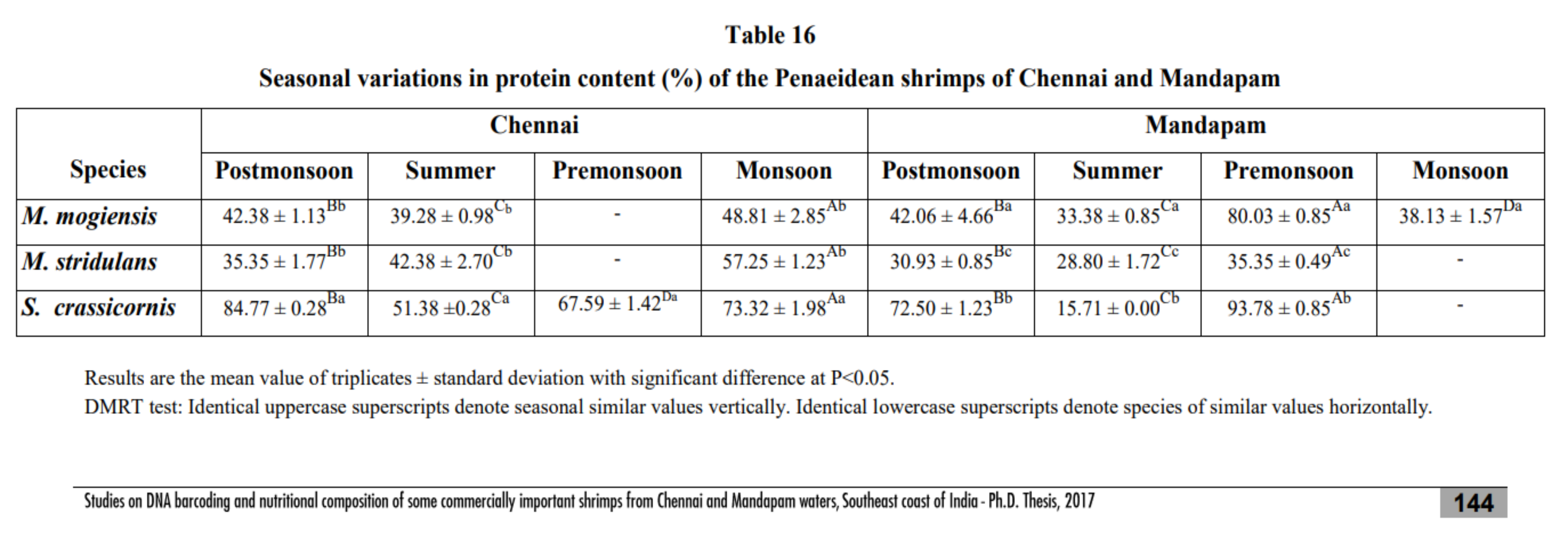


Table 16

Seasonal variations in protein content (%) of the Penaeidean shrimps of Chennai and Mandapam

Species	Chennai				Mandapam			
	Postmonsoon	Summer	Premonsoon	Monsoon	Postmonsoon	Summer	Premonsoon	Monsoon
M. mogiensis	42.38 ± 1.13^{Bb}	39.28 ± 0.98^{Cb}	-	48.81 ± 2.85^{Ab}	42.06 ± 4.66^{Ba}	33.38 ± 0.85^{Ca}	80.03 ± 0.85^{Aa}	38.13 ± 1.57^{Da}
M. stridulans	35.35 ± 1.77^{Bb}	42.38 ± 2.70^{Cb}	-	57.25 ± 1.23^{Ab}	30.93 ± 0.85^{Bc}	28.80 ± 1.72^{Cc}	35.35 ± 0.49^{Ac}	-
S. crassicornis	84.77 ± 0.28^{Ba}	51.38 ±0.28^{Ca}	67.59 ± 1.42^{Da}	73.32 ± 1.98^{Aa}	72.50 ± 1.23^{Bb}	15.71 ± 0.00^{Cb}	93.78 ± 0.85^{Ab}	-

Results are the mean value of triplicates ± standard deviation with significant difference at $P<0.05$.

DMRT test: Identical uppercase superscripts denote seasonal similar values vertically. Identical lowercase superscripts denote species of similar values horizontally.

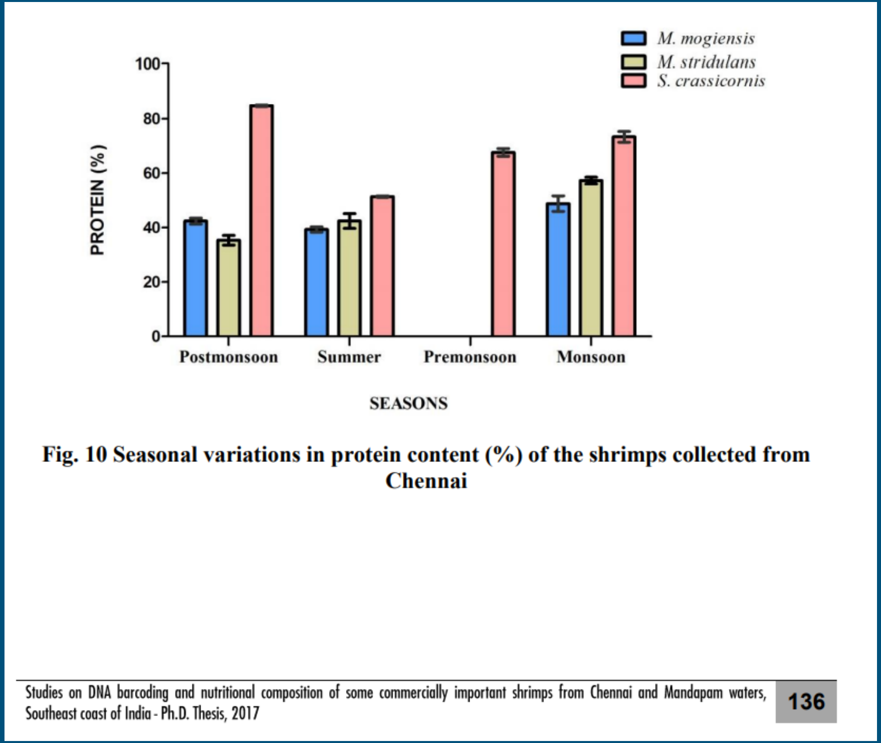


Fig. 10 Seasonal variations in protein content (%) of the shrimps collected from Chennai

Studies on DNA barcoding and nutritional composition of some commercially important shrimps from Chennai and Mandapam waters, Southeast coast of India - Ph.D. Thesis, 2017 136

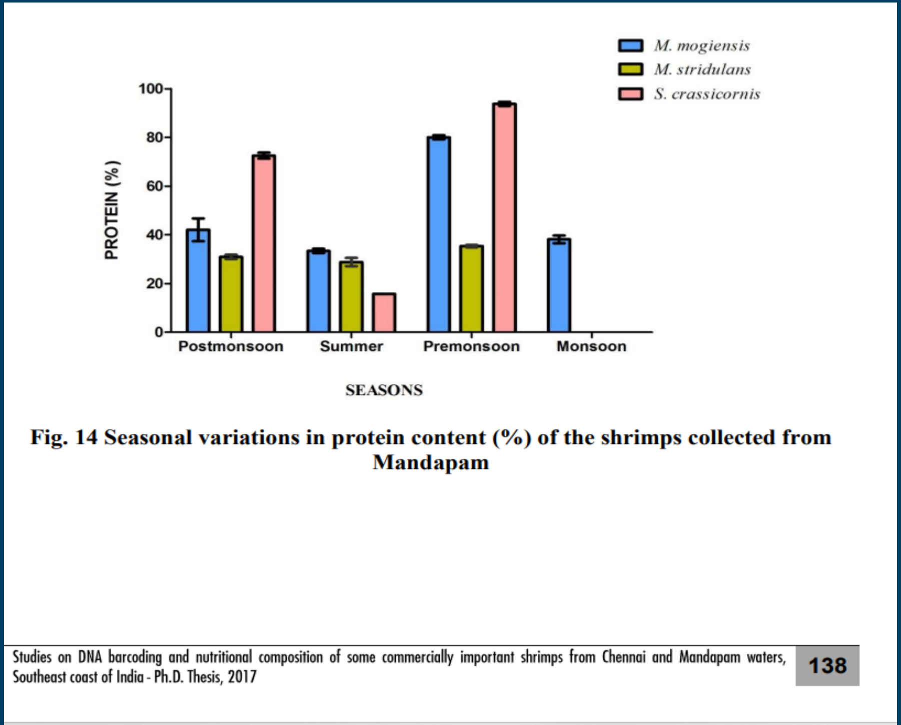
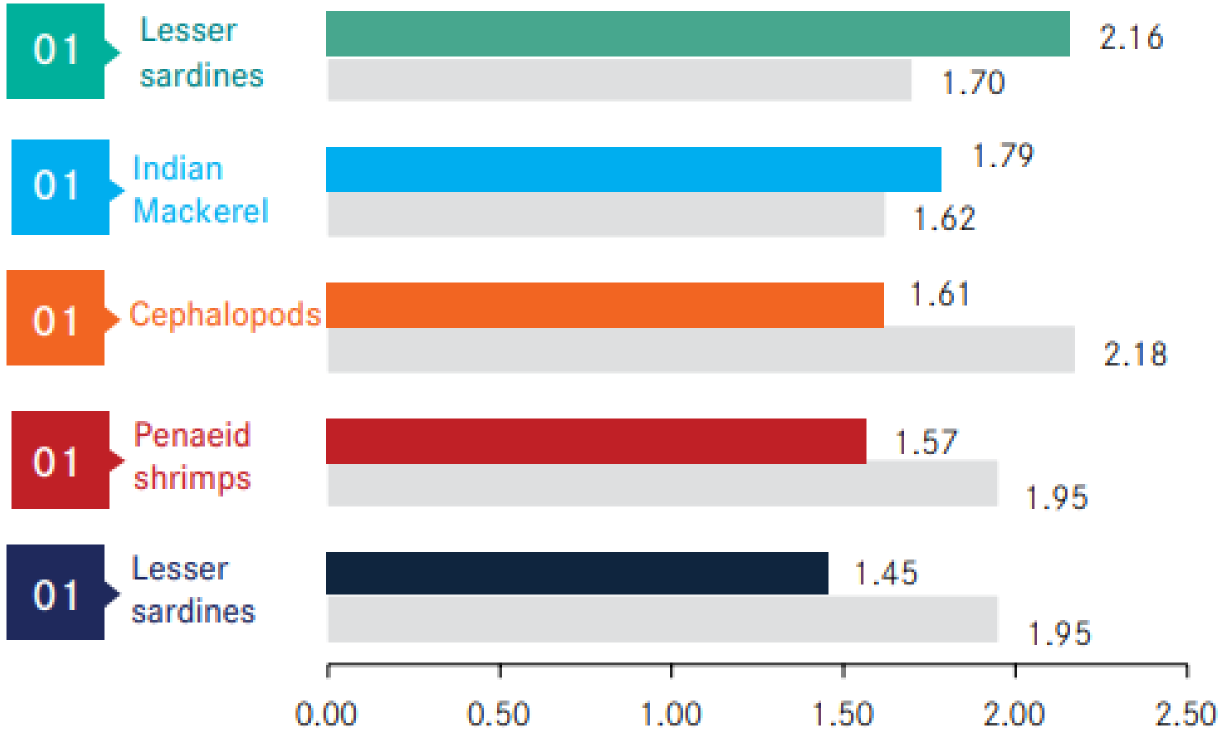
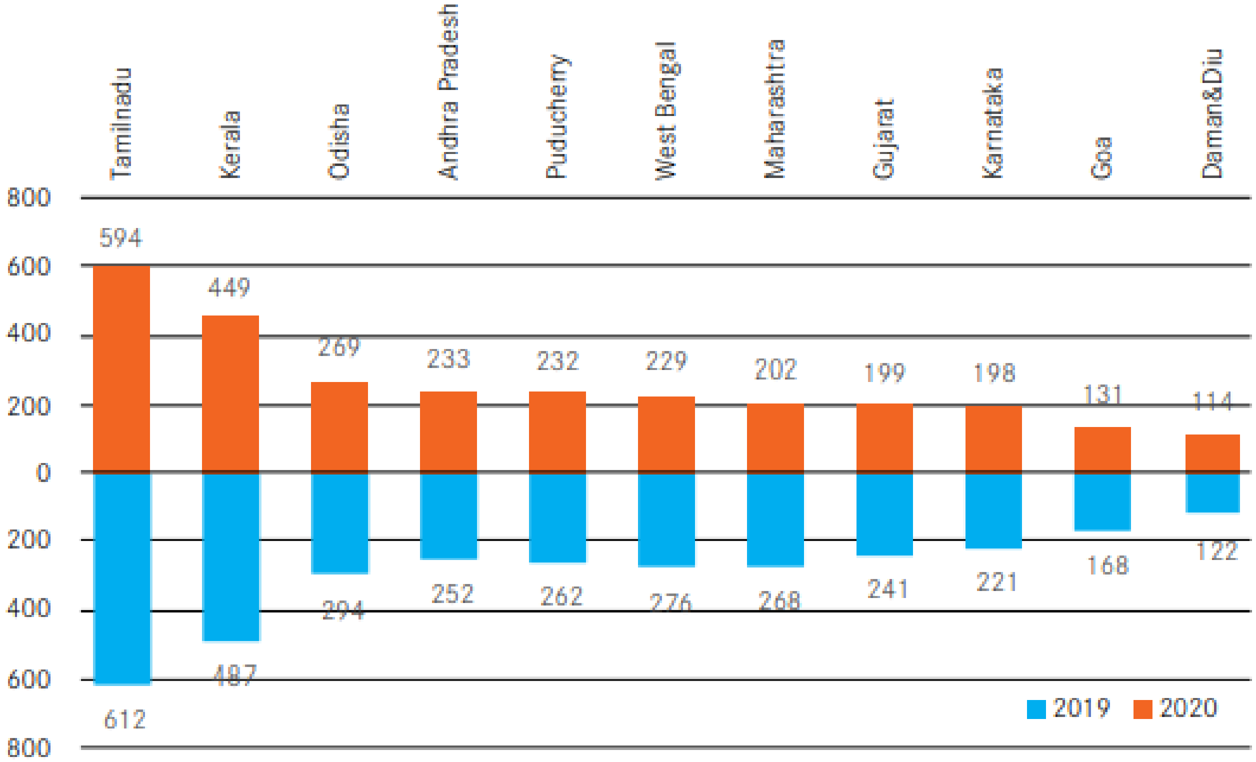


Fig. 14 Seasonal variations in protein content (%) of the shrimps collected from Mandapam

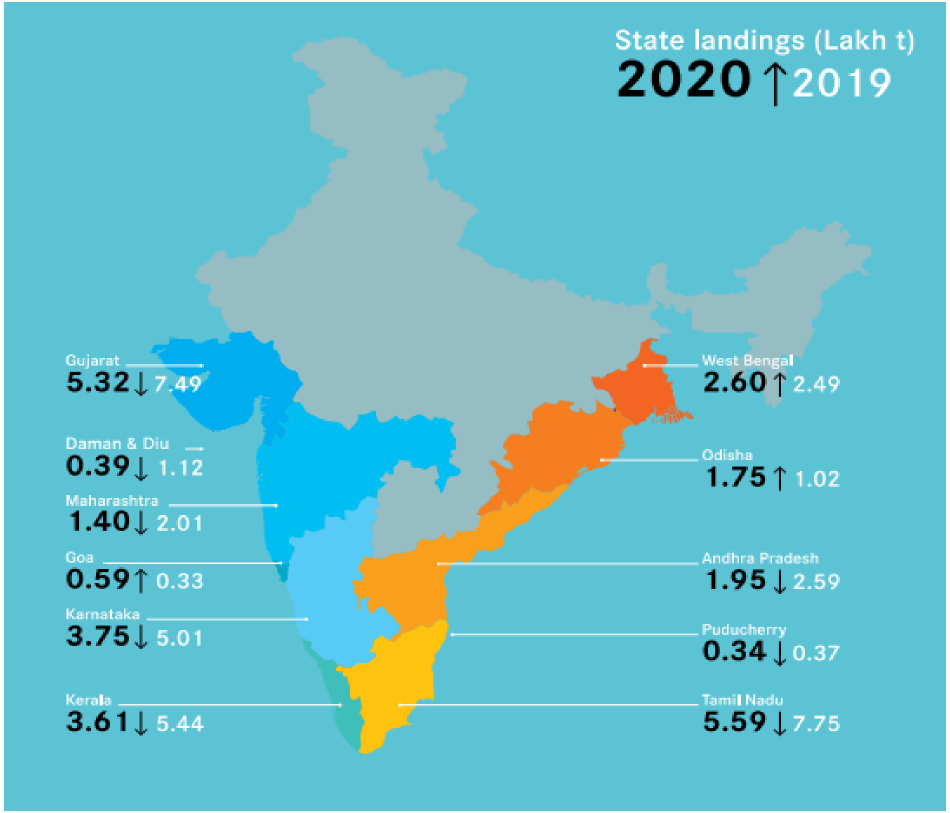
Studies on DNA barcoding and nutritional composition of some commercially important shrimps from Chennai and Mandapam waters, Southeast coast of India - Ph.D. Thesis, 2017 138

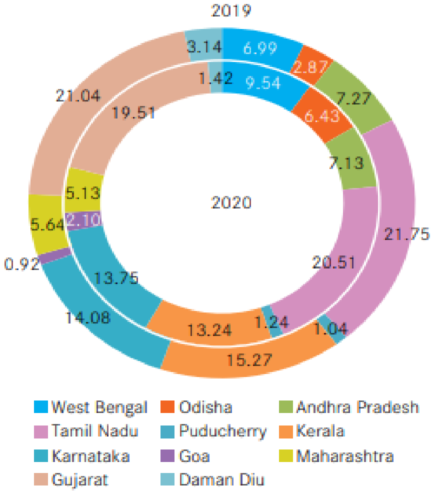


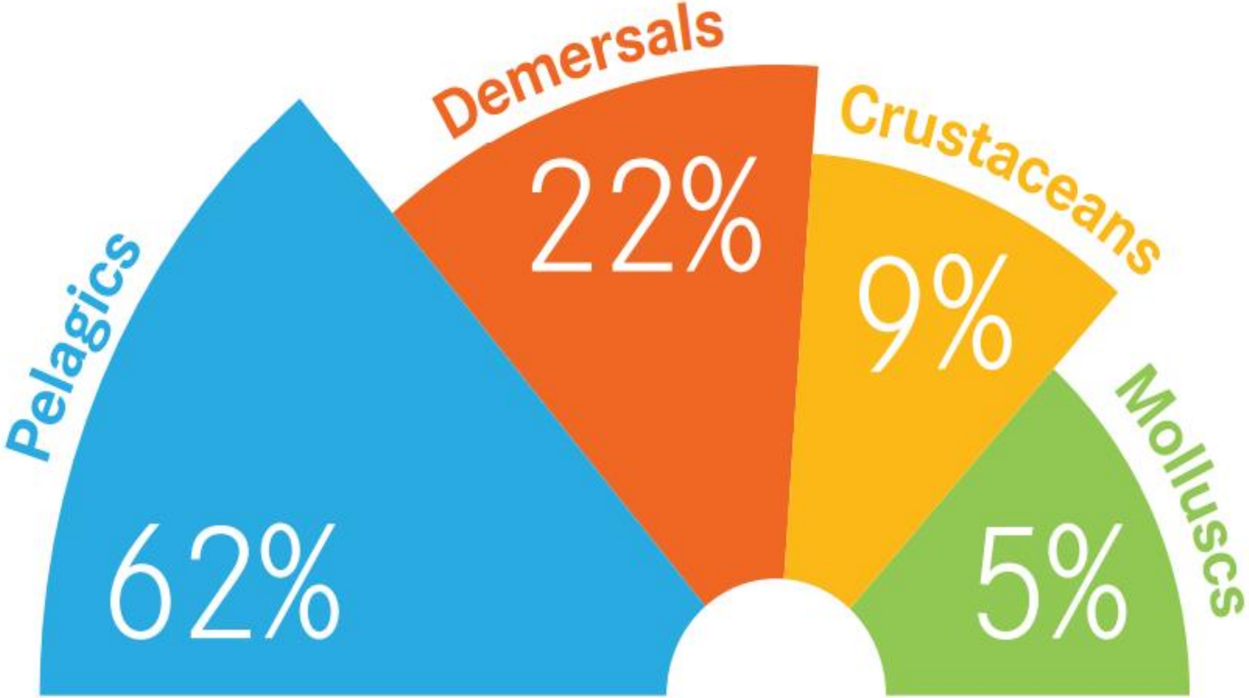
Major five speceis/groups with their contribution (lakh tonnes) towards total marine fish landings in India (2019 & 2020)



1.062 million tonnes which is 29.82% in 2019); 0.435 million tonnes in northeast region contributing 15.97% to national total (0.351 million tonnes which is 9.85% in 2019).

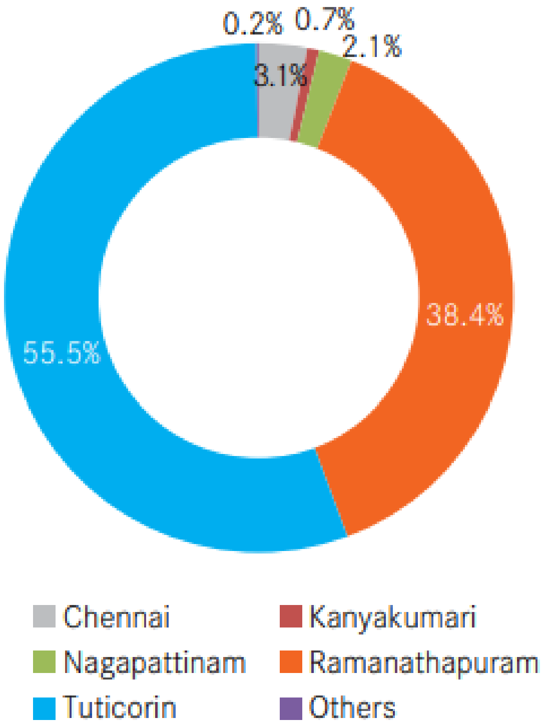

State landings (Lakh t)
2020 ↑ 2019
Gujarat
5.32 ↓ 7.49
Daman & Diu
0.39 ↓ 1.12
Maharashtra
1.40 ↓ 2.01
Goa
0.59 ↑ 0.33
Karnataka
3.75 ↓ 5.01
Kerala
3.61 ↓ 5.44
West Bengal
2.60 ↑ 2.49
Odisha
1.75 ↑ 1.02
Andhra Pradesh
1.95 ↓ 2.59
Puducherry
0.34 ↓ 0.37
Tamil Nadu
5.59 ↓ 7.75


2019
2020
3.14
6.99
2.87
7.27
21.75
1.04
15.27
14.08
0.92
5.64
21.04
1.42
9.54
6.43
7.13
20.51
1.24
13.24
13.75
2.10
5.13
19.51
West Bengal
Odisha
Andhra Pradesh
Tamil Nadu
Puducherry
Kerala
Karnataka
Goa
Maharashtra
Gujarat
Daman Diu



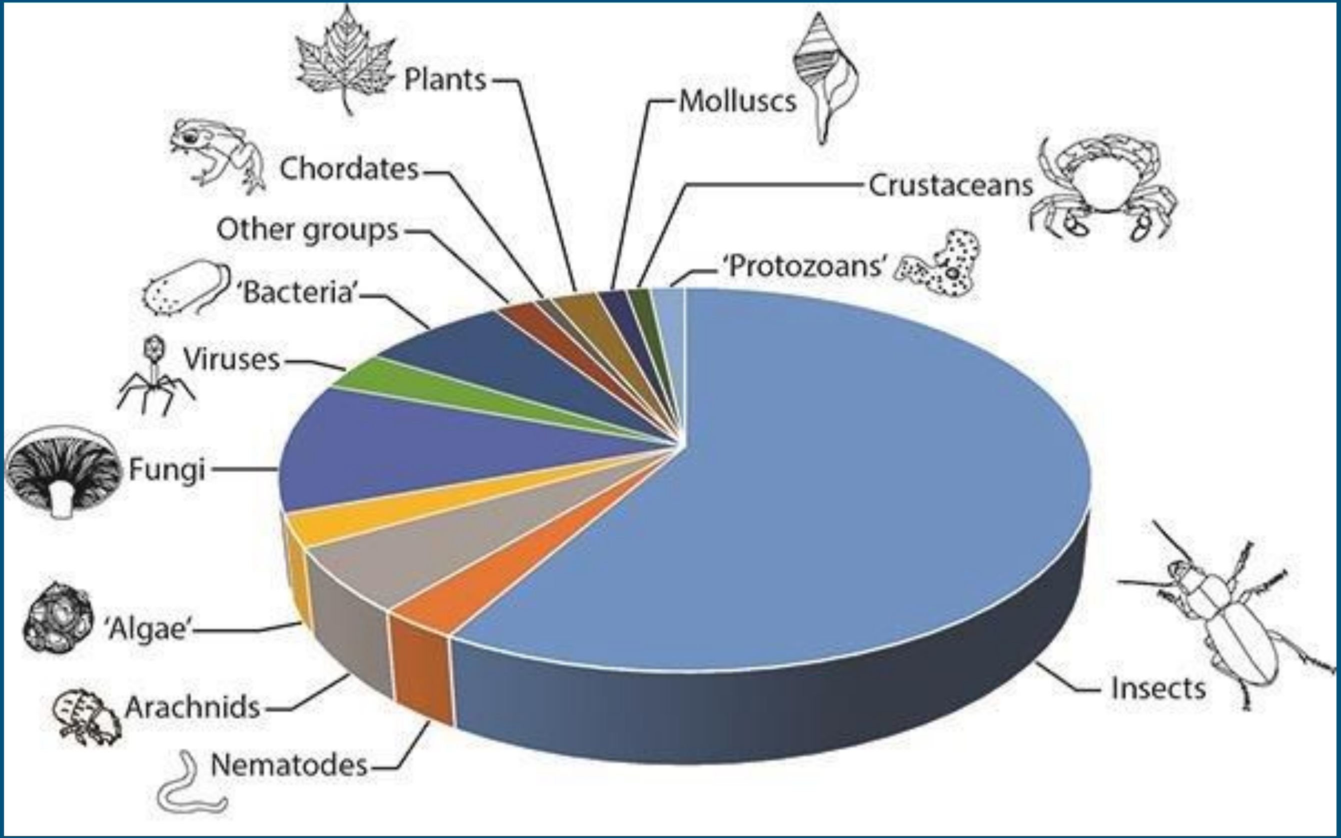
Contribution of different groups to the marine fish landing of Kerala during 2020.

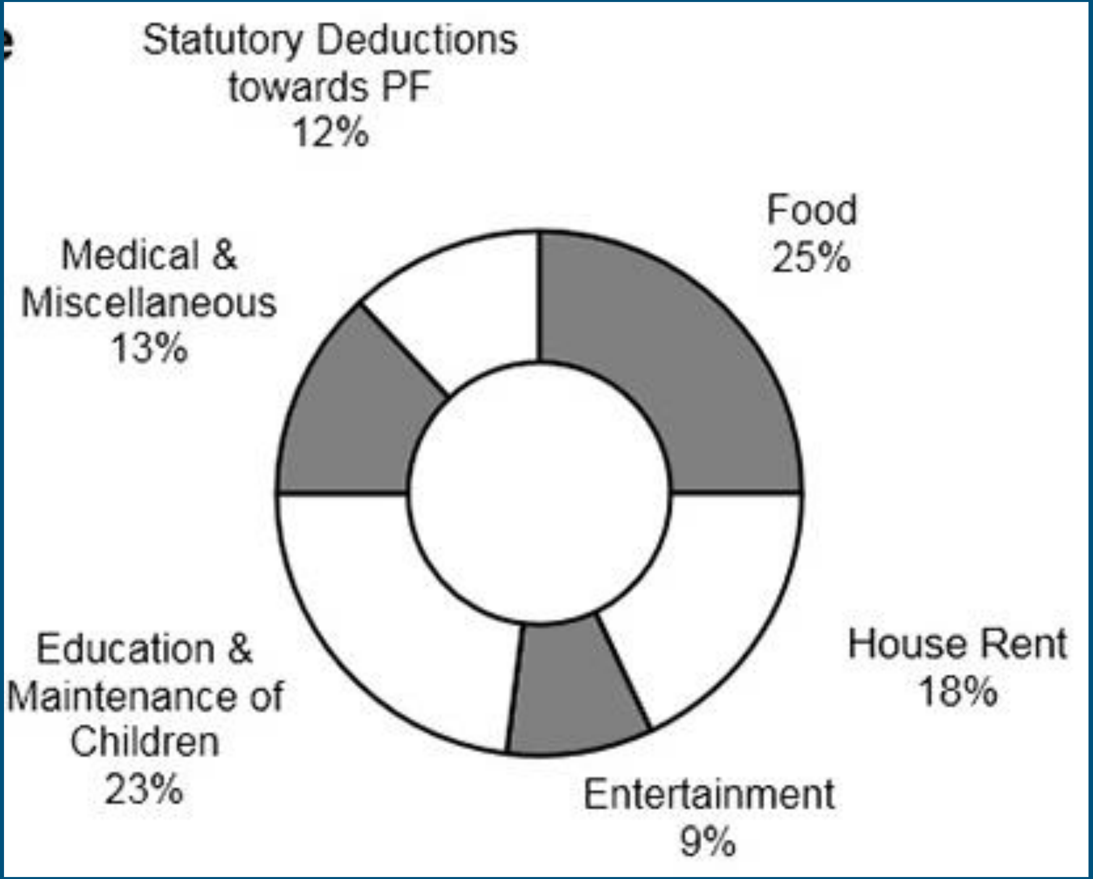
Chicoreus ramosus landed by skin diving



Percentage composition of gastropod catch at different Districts in Tamil Nadu

Pie diagram


Plants
Molluscs
Chordates
Crustaceans
Other groups
'Protozoans'
'Bacteria'
Viruses
Fungi
'Algae'
Arachnids
Nematodes
Insects



The total earnings of person X are Rs. 3,600 per month basic, plus 10% as transport and meals allowance on the monthly salary.

Q:

Calculate the amount of expenditure on education and maintenance per month, if a person X pays 23 % of its total earnings as Education and Maintenance of children?

Sol:

Total Emoluments = 3600 + 10% of 3600 = Rs. 3960

So 23% of 3960 = Rs. 910.8

FREQUENCY DISTRIBUTION

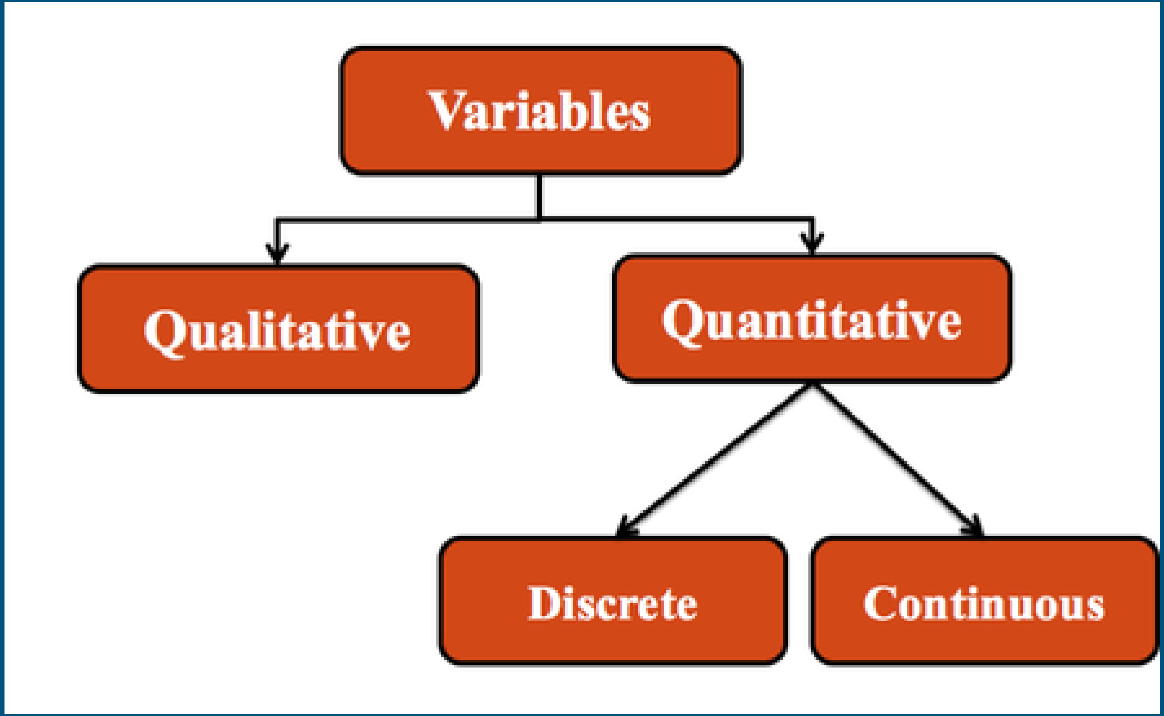
“Classification of a random variable into a no. of classes or CI indicating frequency the C or representatives of CI occur in the data”

Classify qualitative and quantitative data - frequency table

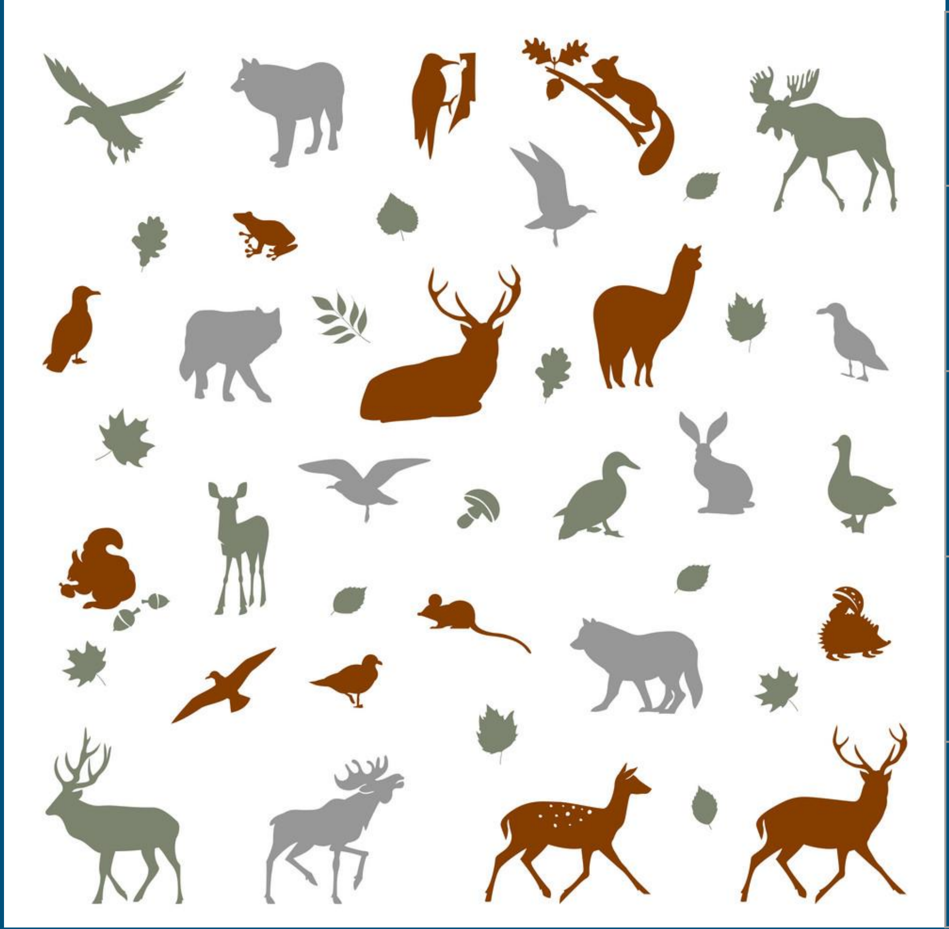
Graphical representation of a frequency distribution → Cumulative frequency graphs / Ogives

Types: Relative and percent relative frequency

Frequency distribution



Relative frequency - Discrete data



Species	Frequency (f)

Calculate….

Cumulative frequency distribution

Frequency (f)	Cumulative frequency	% cumulative Rf = Rf /Σf x 100
8	8	
2		
10		
7		
1		
Σf =		

Frequency distribution

Calculate….

Classes	Frequency (f)	Relative frequency $Rf = f/\Sigma f$	% Rf $= Rf \times 100$
Crustaceans	105		
Molluscs	5		
Penaeid shrimps	84		
Non-penaeid shrimps	73		
Teleosts	91		
	$\Sigma f = 358$		

Discrete frequency distribution

Calculate….

Class interval (No. of fin rays)	Midpoint 11+21 / 2	Frequency (f) (No. of fins)	Relative frequency Rf = f/Σf	% Rf = Rf x 100
11 - 21	16	5		
22 - 32		3		
33 - 43		4		
44 - 54		2		
55 - 65		1		
		Σf =		

Tally option…..in a data pool

PROBABILITY DISTRIBUTION

•••

- Observed frequency distribution (Experimental data)

- Theoretical frequency distribution (Assumptions; hypothetical computation)

  1. Binomial 2. Poisson 3. Normal

Probability distribution

Chance (p) - The no. of times (N) an event occurs in a very large number of trials (X).

$P = X / N$ (p can vary from 0 to 1);

No. of individuals is limited or less individuals, denote " f ";

No. of individuals are large, denote "ρ"

If $X = 0$, then $f = 0$ and $\rho = 0$

If $\rho = 1$, then $P = X / N$ will be $1 = X / N$. So $X = N$ (Probability of a win or lose).

Calculating relative frequency or Empirical probability P(A):

P(A) = Frequency of occurrence / No of trials

Basic laws of probability:

1. If probability of occurrence of an event is 1, event will occur certainly.
   a. If closer to 1, event will likely occur.

1. If probability of occurrence of an event is 0, event will never occur.
   a. If closer to 0, less likely the event will occur.

1. Probability of any event must assume a value between 0 and 1.

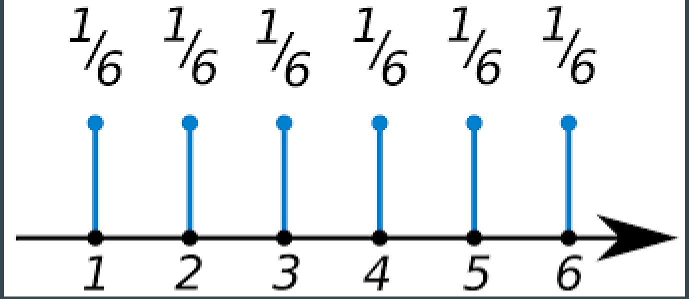
1. Probability of sample space in any experiment is always 1.

Probability distribution (P):

__________The numerical quantity (value) obtained by the outcome of a random experiment is random variable.

This random variable have various values or sets of values with different probabilities under different conditions or from different sample size.

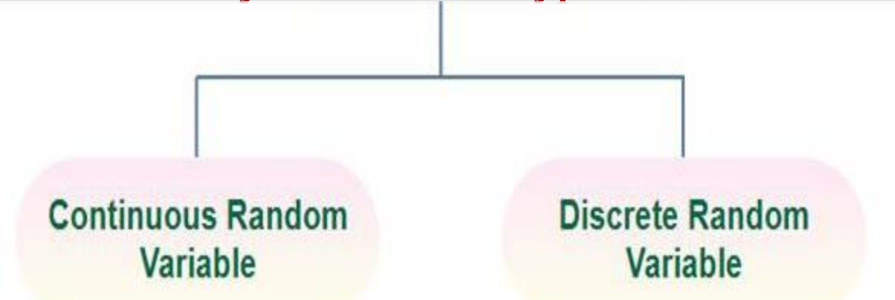
If different sample size "n" in a population is considered, the "mean calculation" is considered the random variable.



The probability of occurrence of outcomes → P

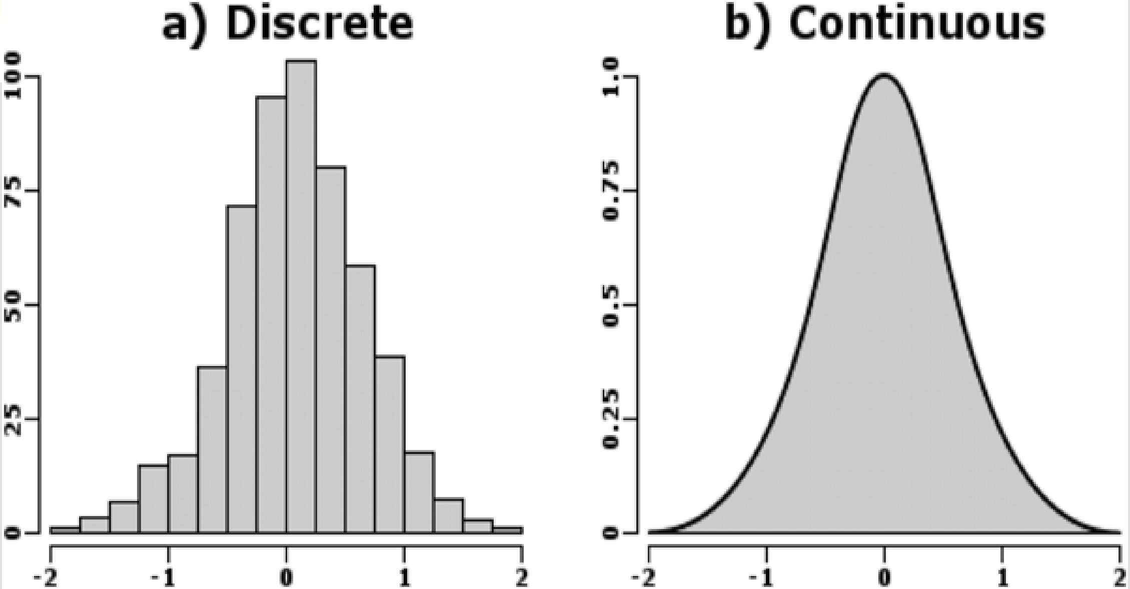
Sum of all probabilities in a pd is " 1 "

Probability distribution types (2):



Continuous Random Variable

Discrete Random Variable



Binomial distribution

Standard Normal

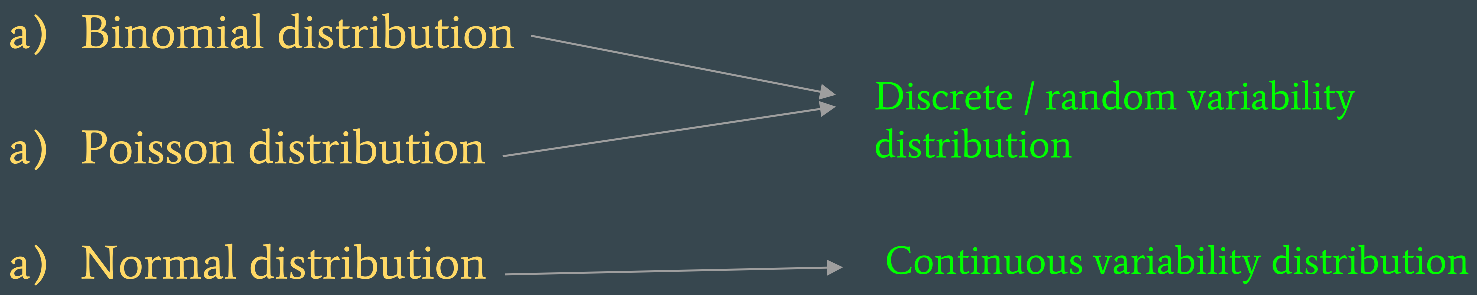
distribution

Frequency distribution

Frequency distribution of a variable or CI is of two types:

1. Observed frequency distribution = Actual data of an experiment.

1. Theoretical frequency distribution = On assumptions; hypothetical populations; ideal and reasonable distributions.



a) Binomial distribution

a) Poisson distribution

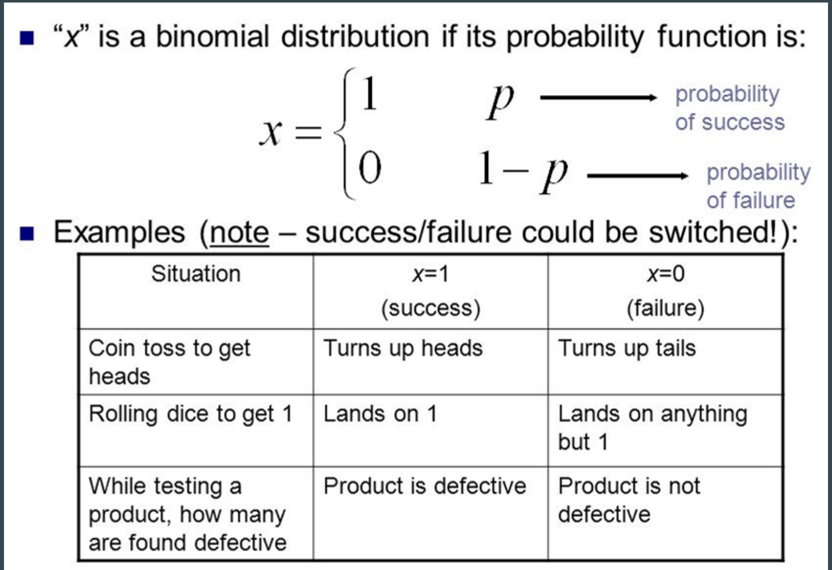
a) Normal distribution

Discrete / random variability distribution

Continuous variability distribution

Binomial distribution

- **Random** variables and its distribution
- Only **two mutual outcomes** each with a known probability



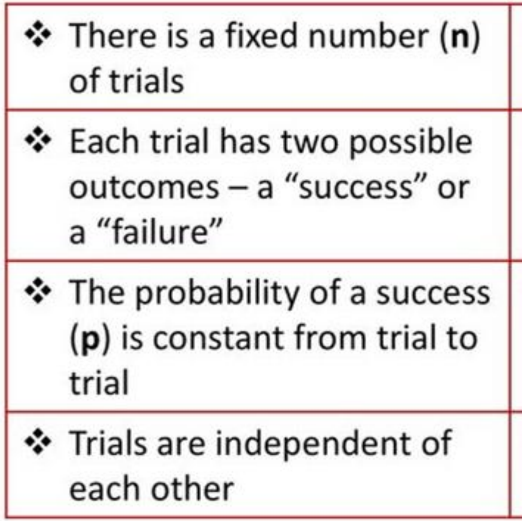
- "*x*" is a binomial distribution if its probability function is:

$$x = \begin{cases} 1 & p \longrightarrow \text{probability of success} \\ 0 & 1-p \longrightarrow \text{probability of failure} \end{cases}$$

- Examples (note – success/failure could be switched!):

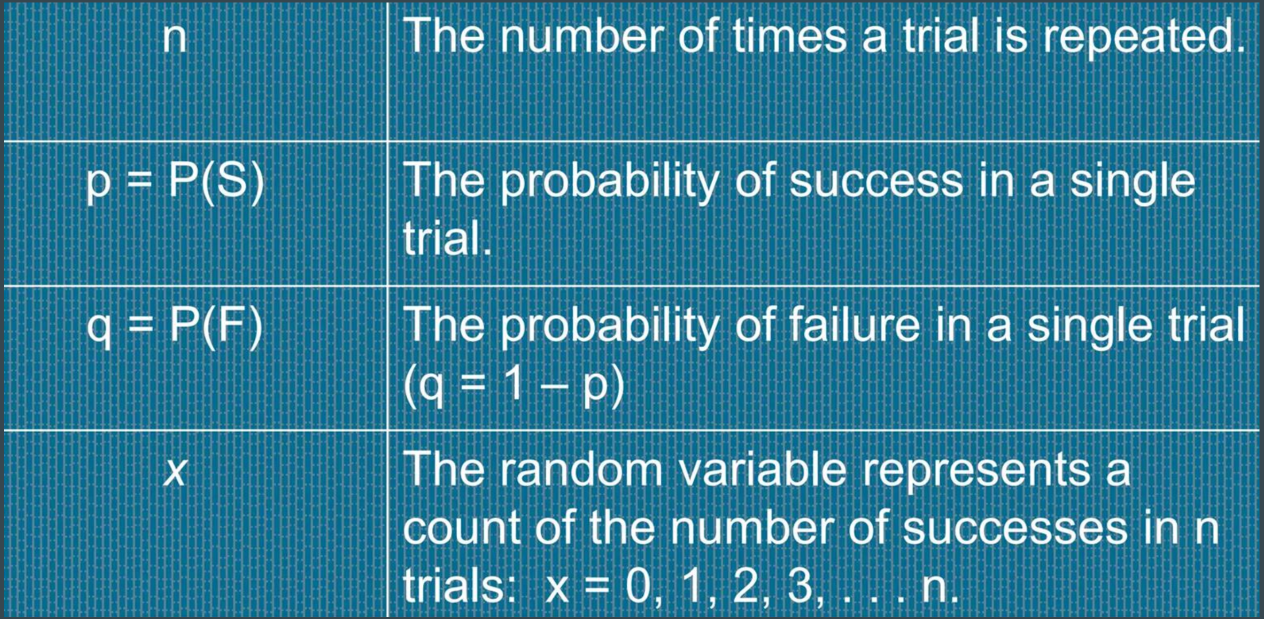
Situation	*x*=1 (success)	*x*=0 (failure)
Coin toss to get heads	Turns up heads	Turns up tails
Rolling dice to get 1	Lands on 1	Lands on anything but 1
While testing a product, how many are found defective	Product is defective	Product is not defective

Dichotomous classification

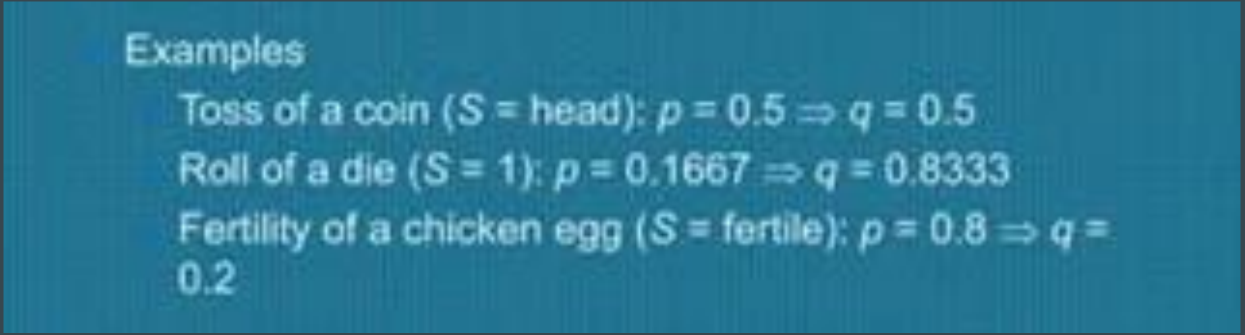


- ❖ There is a fixed number (**n**) of trials
- ❖ Each trial has two possible outcomes – a "success" or a "failure"
- ❖ The probability of a success (**p**) is constant from trial to trial
- ❖ Trials are independent of each other

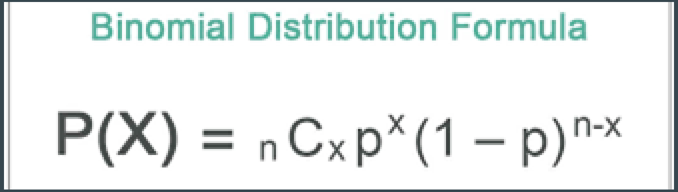
Ref only...



n	The number of times a trial is repeated.
p = P(S)	The probability of success in a single trial.
q = P(F)	The probability of failure in a single trial (q = 1 – p)
x	The random variable represents a count of the number of successes in n trials: x = 0, 1, 2, 3, . . . n.

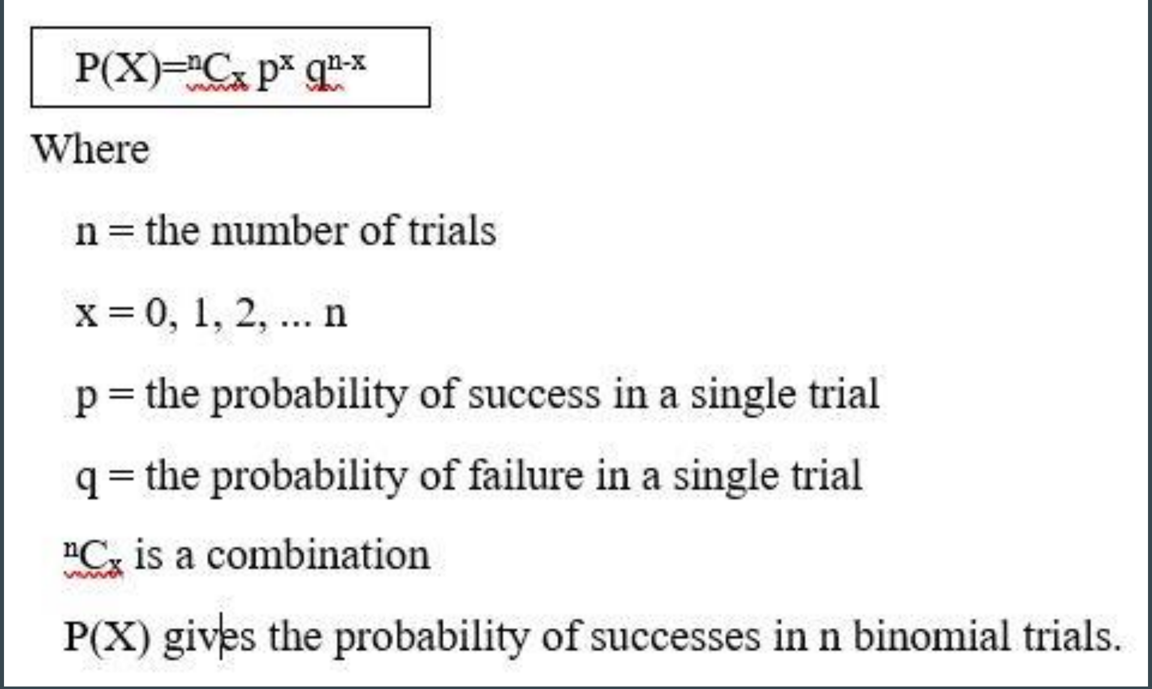


- Examples
  - Toss of a coin (S = head): $p = 0.5 \Rightarrow q = 0.5$
  - Roll of a die (S = 1): $p = 0.1667 \Rightarrow q = 0.8333$
  - Fertility of a chicken egg (S = fertile): $p = 0.8 \Rightarrow q = 0.2$



Binomial Distribution Formula

$$P(X) = {}_{n}C_{x}\, p^{x}\, (1 - p)^{n-x}$$



$$\boxed{P(X)={}^{n}C_{x}\, p^{x}\, q^{n-x}}$$

Where

n = the number of trials

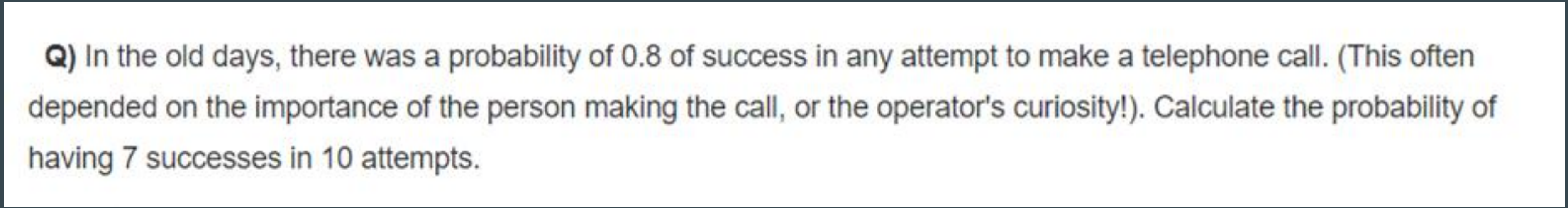
x = 0, 1, 2, ... n

p = the probability of success in a single trial

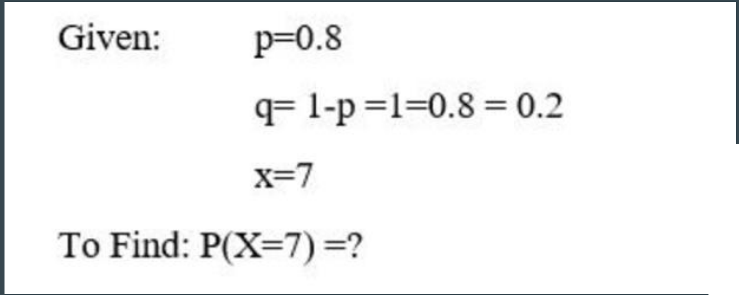
q = the probability of failure in a single trial

${}^{n}C_{x}$ is a combination

P(X) gives the probability of successes in n binomial trials.



Q) In the old days, there was a probability of 0.8 of success in any attempt to make a telephone call. (This often depended on the importance of the person making the call, or the operator's curiosity!). Calculate the probability of having 7 successes in 10 attempts.

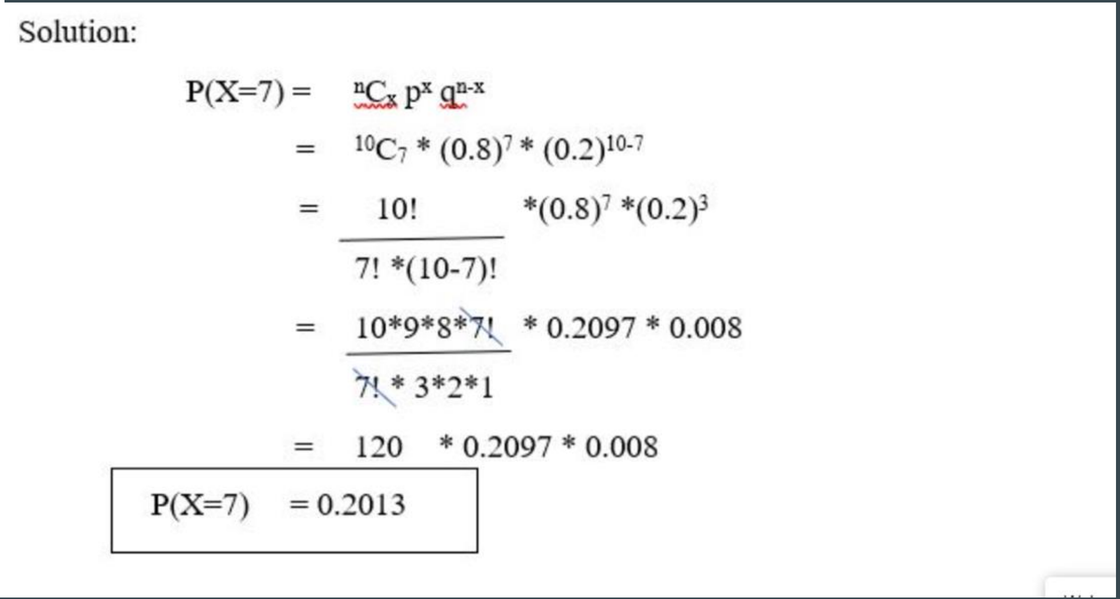


Given: $p=0.8$

$q= 1-p =1=0.8 = 0.2$

$x=7$

To Find: $P(X=7) =?$

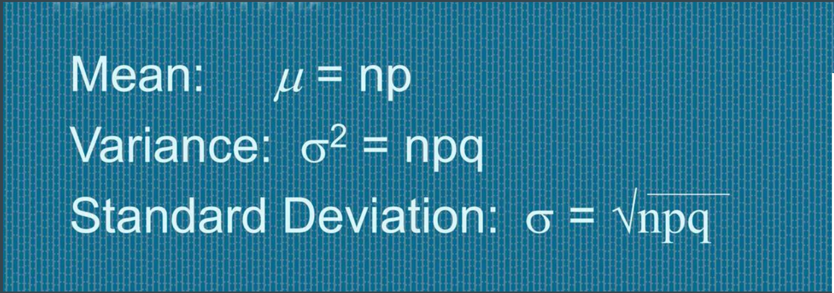


Solution:

$$\begin{aligned} P(X=7) &= {}^{n}C_{x}\, p^{x}\, q^{n-x} \\ &= {}^{10}C_{7} * (0.8)^{7} * (0.2)^{10-7} \\ &= \frac{10!}{7! *(10-7)!} *(0.8)^{7} *(0.2)^{3} \\ &= \frac{10*9*8*\cancel{7!}}{\cancel{7!} * 3*2*1} * 0.2097 * 0.008 \\ &= 120 \quad * 0.2097 * 0.008 \end{aligned}$$

$$\boxed{P(X=7) \quad = 0.2013}$$

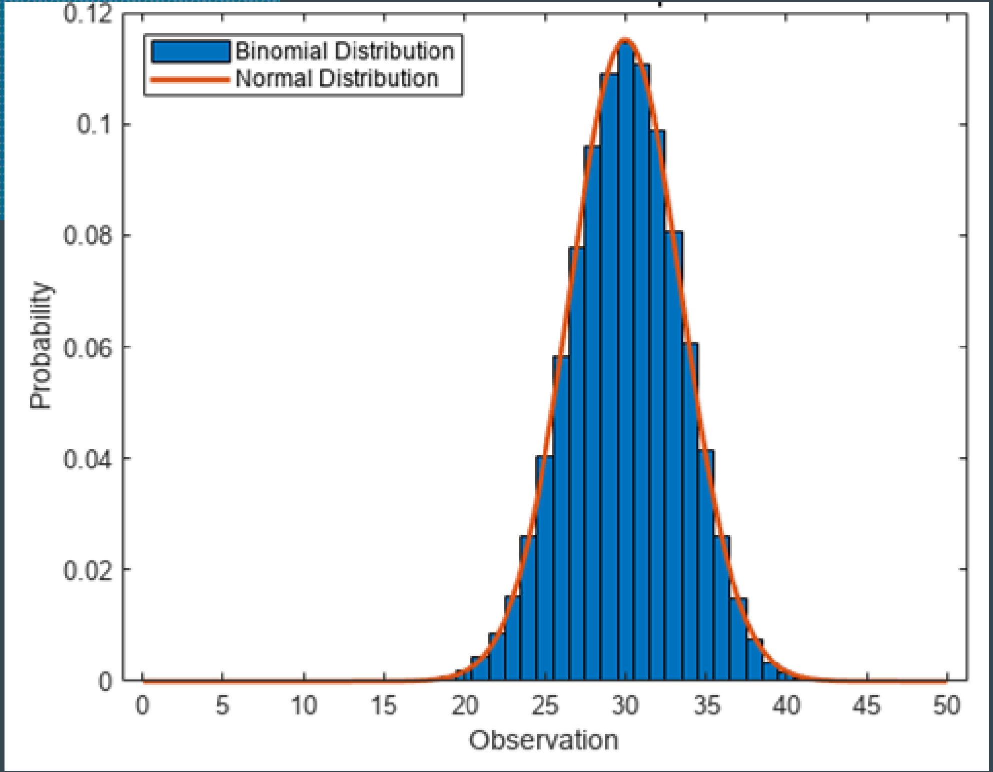
Population parameters of a binomial distribution:



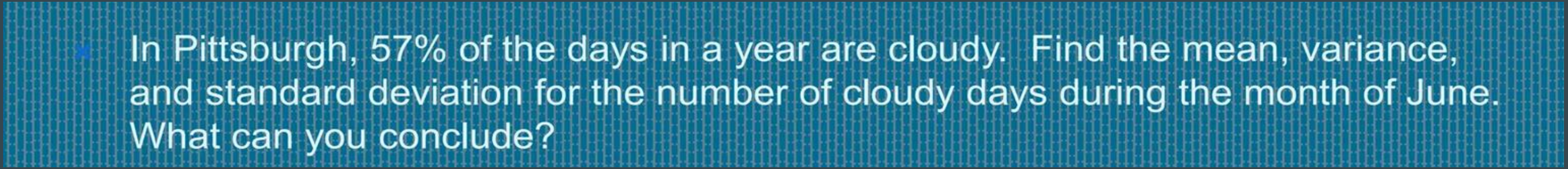
Mean: $\mu = np$

Variance: $\sigma^2 = npq$

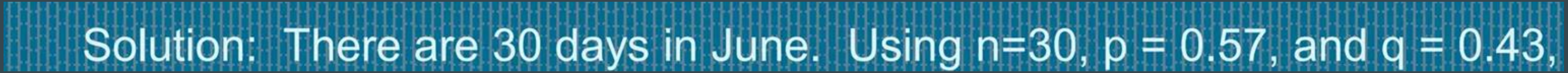
Standard Deviation: $\sigma = \sqrt{npq}$



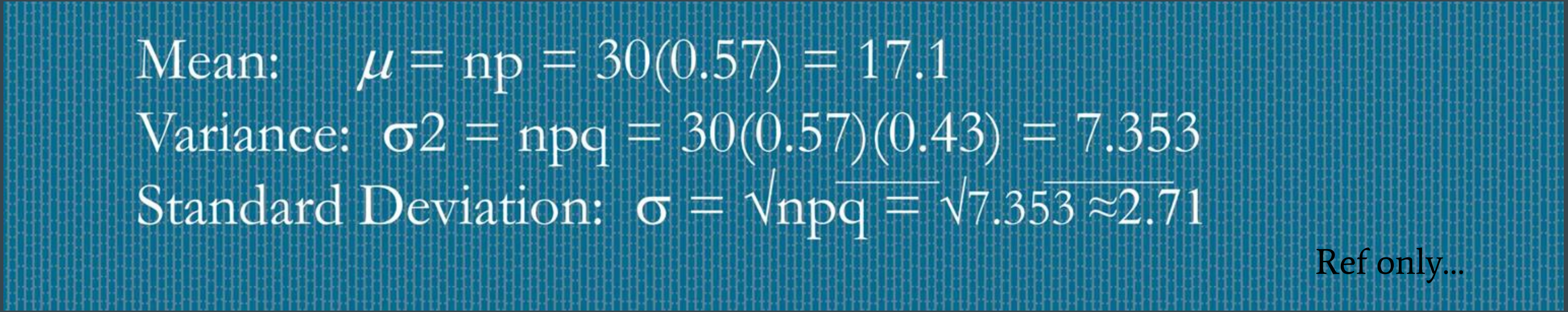
Example:



- In Pittsburgh, 57% of the days in a year are cloudy. Find the mean, variance, and standard deviation for the number of cloudy days during the month of June. What can you conclude?



Solution: There are 30 days in June. Using n=30, p = 0.57, and q = 0.43,



Mean: $\mu = np = 30(0.57) = 17.1$

Variance: $\sigma 2 = npq = 30(0.57)(0.43) = 7.353$

Standard Deviation: $\sigma = \sqrt{npq} = \sqrt{7.353} \approx 2.71$

Ref only...

Poisson distribution

- It is a discrete probability distribution.
- In binomial distribution, both outcomes → Success and failures are known.
- But in some situations,

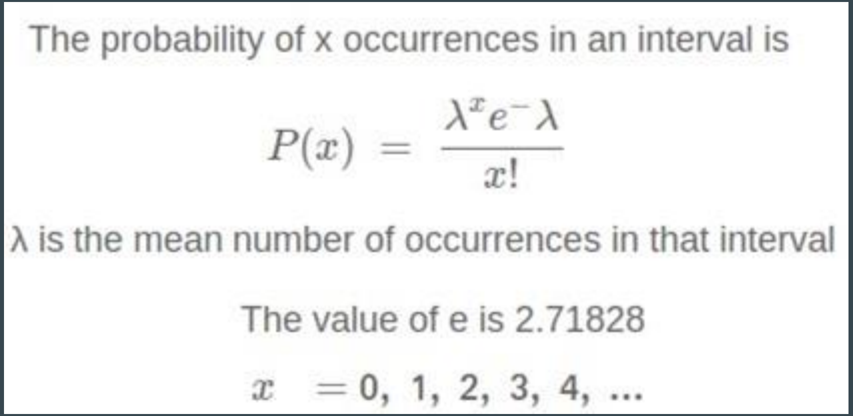
  successes are known (no. of times an event does occur);
  failures are not known or cannot be predicted or calculated,

This type of distribution → Poisson distribution (Simeon Denis Poisson)

- Poisson distribution represents the probability distribution of rare events, prob. of occurrence is very small.
- Here,
  - No. of trials / events (n) is very large to infinity, but the probability of success (P) is small for each event.

- So value of

n.p = “m” = constant “ mean of distribution “ Poisson distribution



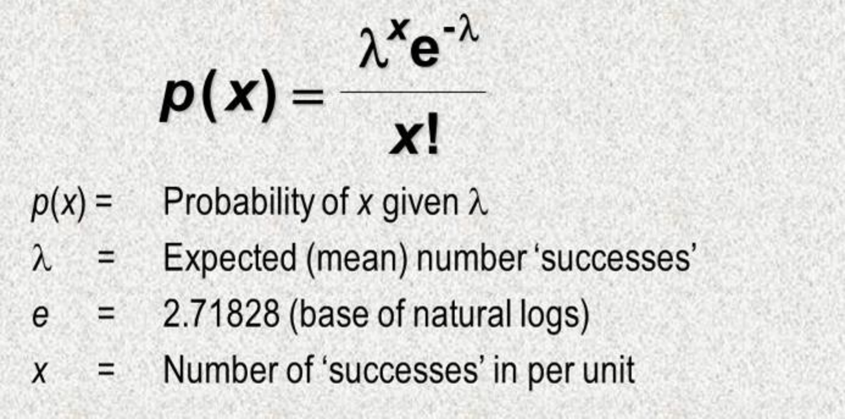
The probability of x occurrences in an interval is

$$P(x) = \frac{\lambda^x e^{-}\lambda}{x!}$$

λ is the mean number of occurrences in that interval

The value of e is 2.71828

$x = 0, 1, 2, 3, 4, \ldots$



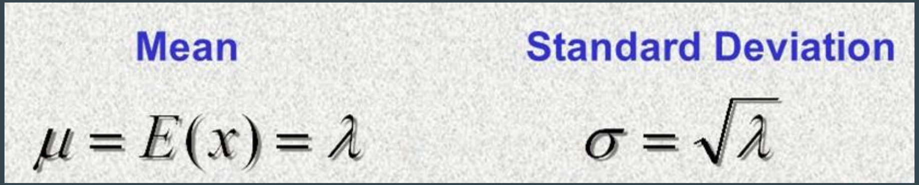
$$p(x) = \frac{\lambda^{x} e^{-\lambda}}{x!}$$

$p(x)$ = Probability of x given λ

λ = Expected (mean) number 'successes'

e = 2.71828 (base of natural logs)

x = Number of 'successes' in per unit



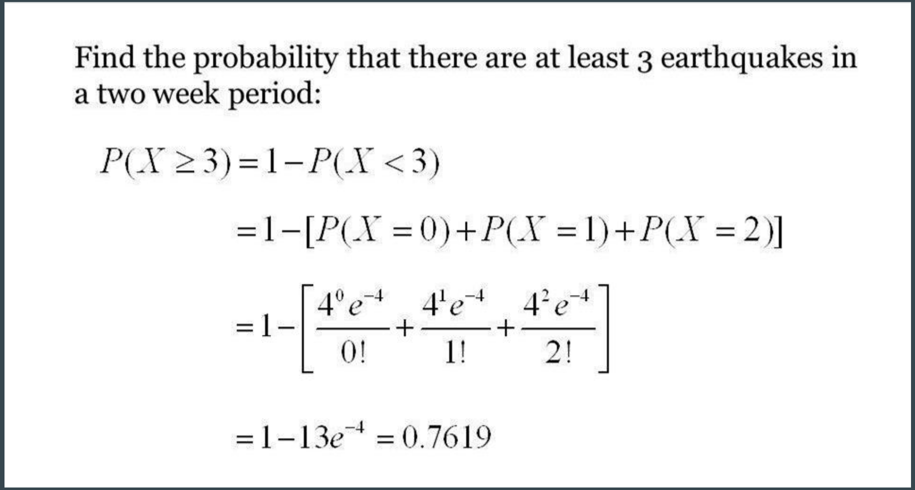
Mean

$$\mu = E(x) = \lambda$$

Standard Deviation

$$\sigma = \sqrt{\lambda}$$

Example:



Find the probability that there are at least 3 earthquakes in a two week period:

$$P(X \geq 3) = 1 - P(X < 3)$$

$$= 1 - [P(X = 0) + P(X = 1) + P(X = 2)]$$

$$= 1 - \left[\frac{4^{0} e^{-4}}{0!} + \frac{4^{1} e^{-4}}{1!} + \frac{4^{2} e^{-4}}{2!}\right]$$

$$= 1 - 13e^{-4} = 0.7619$$

Normal distribution / Gaussian distribution

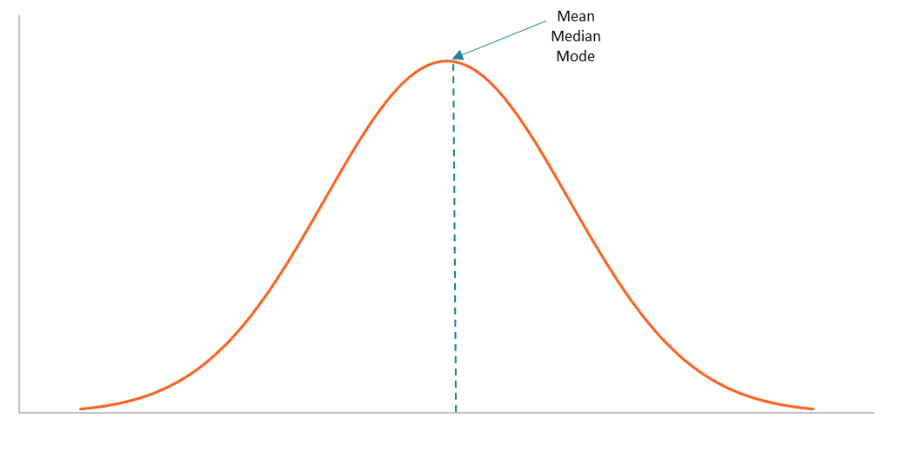
- It is a continuous probability distribution.
- Frequencies of variables are concentrated closely around center and gradually fall towards two ends.

  - Abraham De Moivre (1667-1754)
  - Significance -

    Gaussian distribution - Carl Gauss (1777-1855)

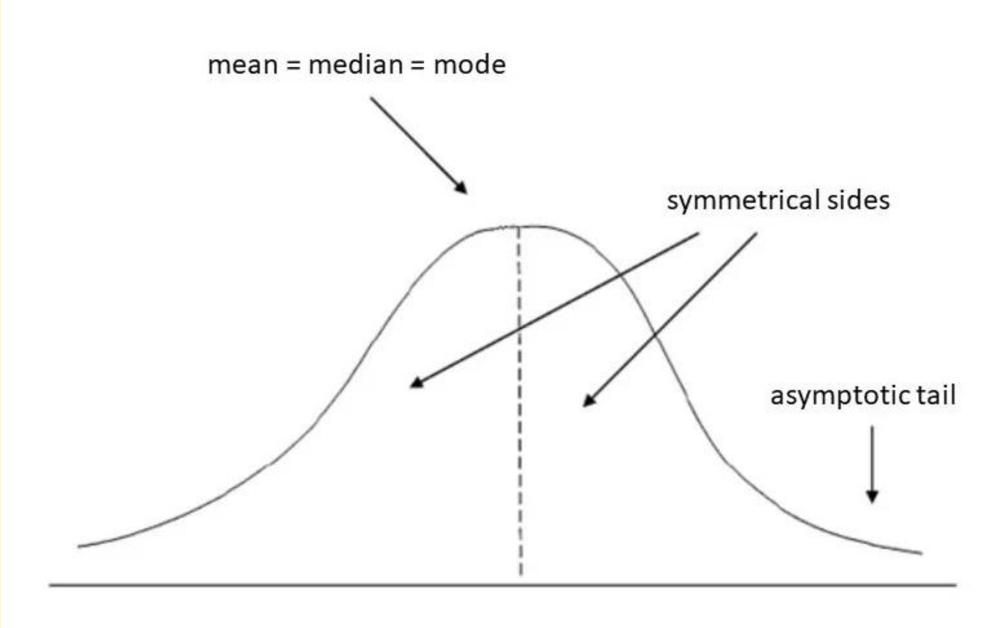
Graphical representation of Normal distribution (Gaussian curve):

- Symmetrical, bell-shaped curve
- Mean value of variable lies at the peak of curve
- Largest no. of observations (values) lie at mean and close to it
- **Md** coincides with **μ** (ordinate divides the area under normal curve into two equal parts)
- Variables of lower value - Left
- Variables of higher value - Right



Properties of Normal distribution (Gaussian curve):

- Common probability distribution of frequencies of a random continuous variable.
- Unimodal, perfectly symmetrical and continuous.
- A bell-shaped curve, asymptotic - touches at infinity (never meet the baseline).

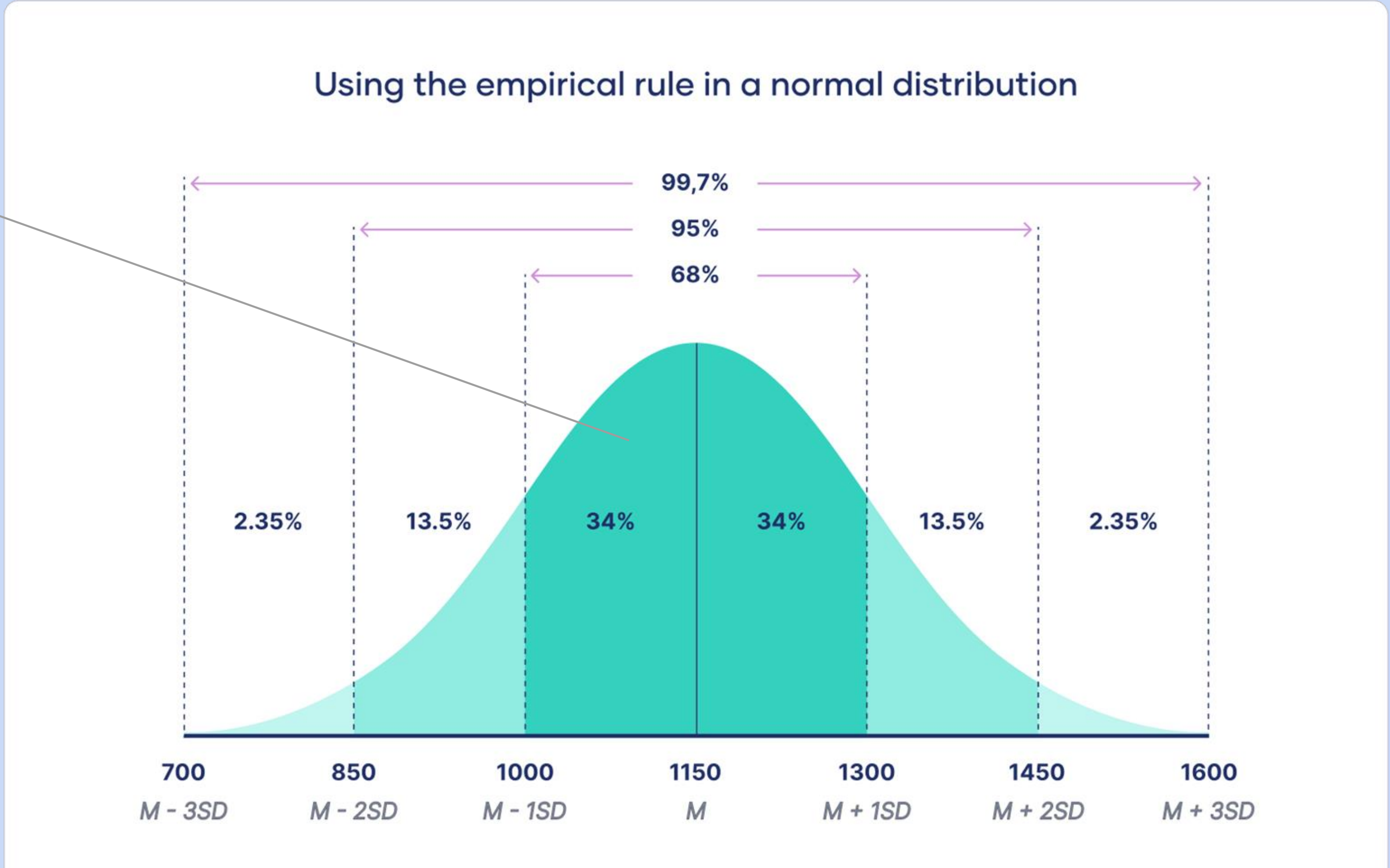


- For a normal distribution, all measures of central tendency are equal (**mean = median = mode**).
- Maximum ordinate of normal curve is at mean (m) and value of maximum ordinate is $\frac{1}{\sigma\sqrt{2\pi}}$ which divides the normal curve into two equal parts (L / R).
- Observations are clustered around mean; few observations at the extremes.
- Total area covered by curve = 1 (**unity**).

Normal distribution area and SD:

Values on a normal curve lie between two limits **μ and 3σ.** ie., (μ ± 3σ).

Effective range

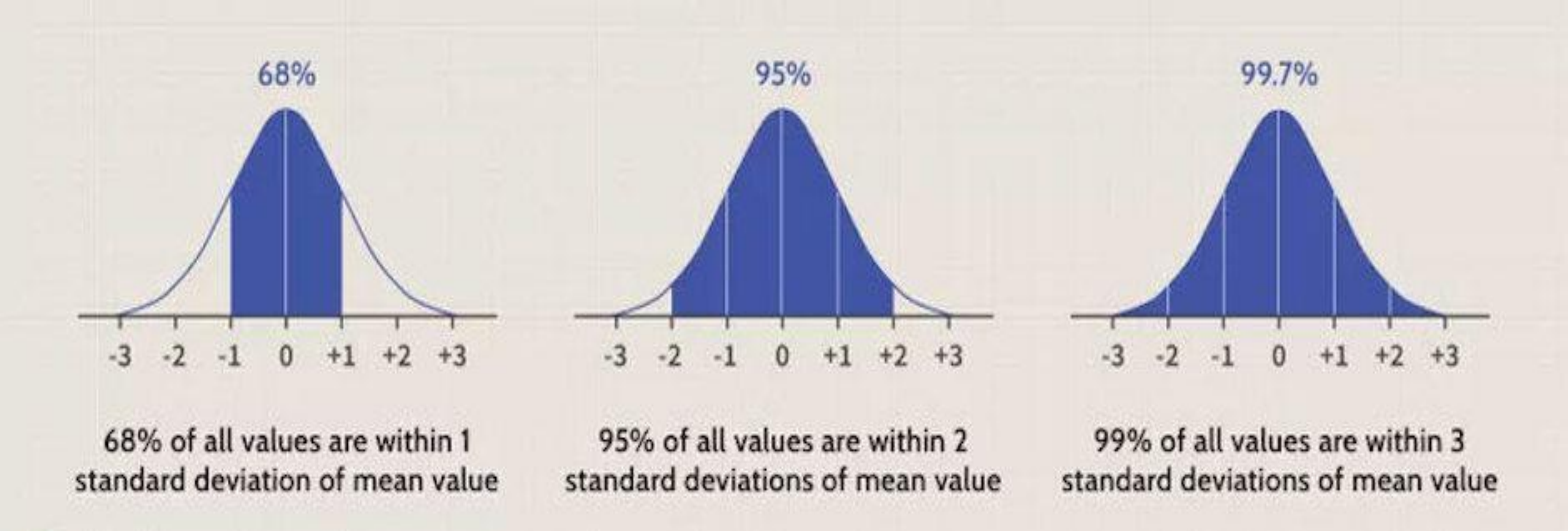


The empirical rule predicts that in normal distributions,

68.27 % of observations (Area = 34.135 %) fall within the first standard deviation (μ ± σ),

95.45 % within the (Area = 47.725 %) first two standard deviations (μ ± 2σ),

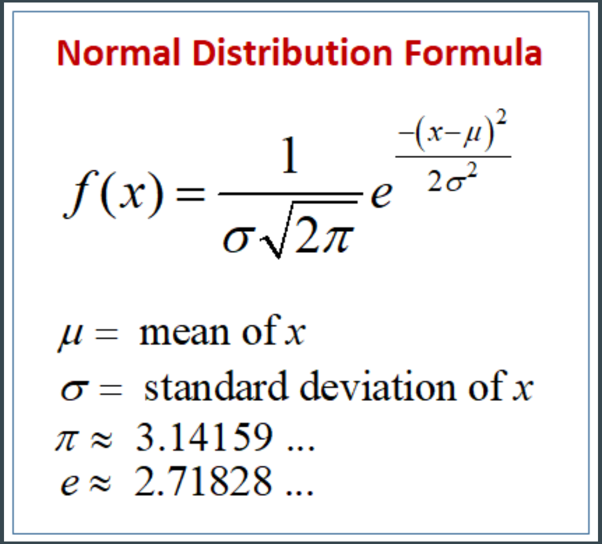
99.73 % within the (Area = 49.865 %) first three standard deviations (μ ± 3σ) of the



68% of all values are within 1 standard deviation of mean value

95% of all values are within 2 standard deviations of mean value

99% of all values are within 3 standard deviations of mean value



Normal Distribution Formula

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$\mu =$ mean of x

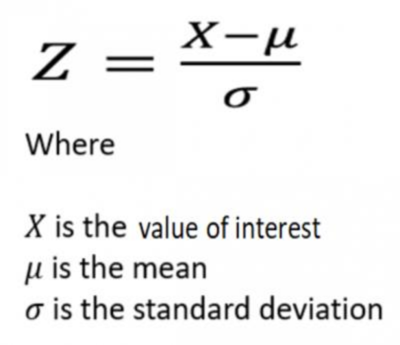
$\sigma =$ standard deviation of x

$\pi \approx$ 3.14159 ...

$e \approx$ 2.71828 ...

z-score is an example of a standardized score.

A z-score measures how many standard deviations a data point is from the mean in a distribution.



$$Z = \frac{X - \mu}{\sigma}$$

Where

X is the value of interest
μ is the mean
σ is the standard deviation

X = Value of continuous random variable

Eg: Ht, Wt, age, fecundity, Resp. rate, Hb %, IQ.

Applications of Normal distribution:

- Applicable to most biological phenomena, bcz of continuous frequency distribution and normal distribution is based on continuous frequency of variables.
- Other sampling methods (t, F, χ^2) can be approximated by normal distribution.
- Sampling theories and its applications.
- Population studies.
- Statistical hypothesis and testing the level of significance.

Standard Normal distribution - curve (based on Z scale):

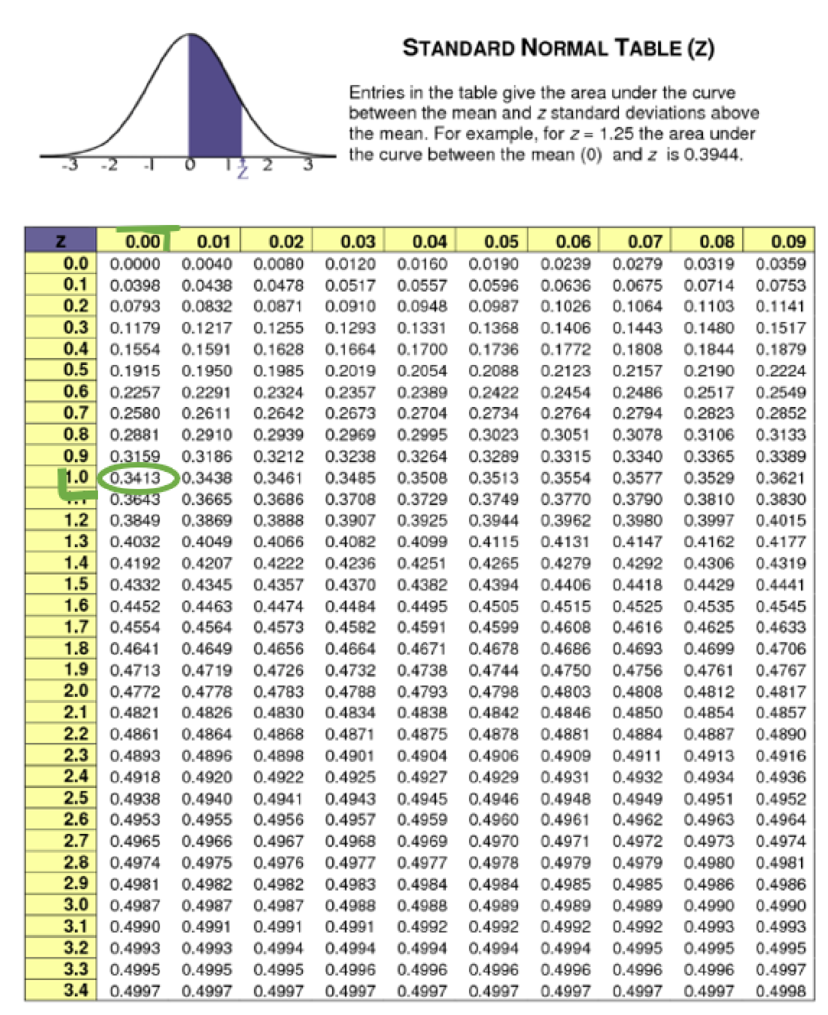
Normal distribution depends on values of **μ** and **σ** and it varies to different data. So normal distribution curves are standardized and converted into one **Standard** normal distribution curve.

The **mean (μ) of Standard Normal distribution curve = 0 and standard deviation (σ) = 1**

To change normal curve → standard curve **x - scale is changed into z - scale.**

(Since) in X-scale, μ = σ

Z-scale, μ = 0 and σ = 1



STANDARD NORMAL TABLE (Z)

Entries in the table give the area under the curve between the mean and *z* standard deviations above the mean. For example, for *z* = 1.25 the area under the curve between the mean (0) and *z* is 0.3944.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0190	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2969	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3513	0.3554	0.3577	0.3529	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998

Measures of deviation from the normal distribution:

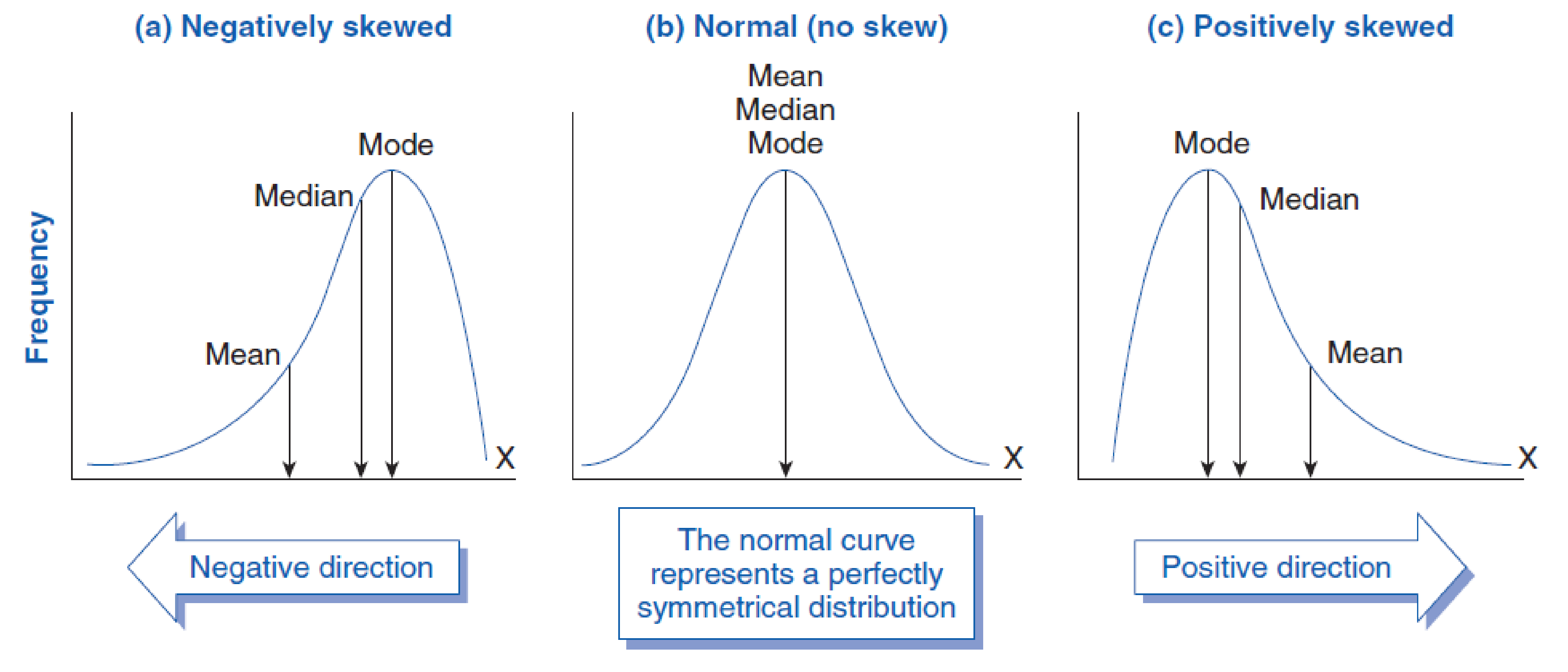
Diverge from normal curve can be studied - **Skewness** and **Kurtosis.**

SKEWNESS:

- When frequency distribution is **asymmetrical**, the distribution is **skewed (L / R)**.
- Mean, median and mode **do not coincide**.
- First quartile (Q_1) and third quartile (Q_3) of frequency curve are **not equidistant from median**.

$$(Q_3 - M) \neq (M - Q_1)$$

Skewness is a measure of the asymmetry



Measures of Skewness:

Measured by **differences between mean and mode**.

1. Absolute skewness
2. Relative skewness
3. Standard skewness
4. Karl-Pearson’s coefficient of skewness
5. Bowley’s coefficient of skewness

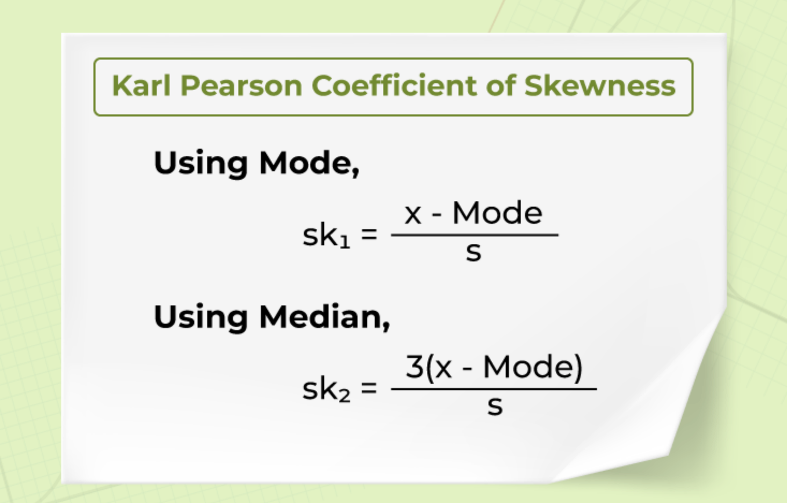
1. Absolute skewness = Mean - Mode

In symmetrical distribution, **Absolute skewness = 0** (Mean = Mode)

Positively skewed distribution, **Absolute skewness = Positive** (Mean > Mode)

Negatively skewed distribution, **Absolute skewness = Negative** (Mean < Mode)

2. Relative skewness / Coefficient of skewness



Karl Pearson Coefficient of Skewness

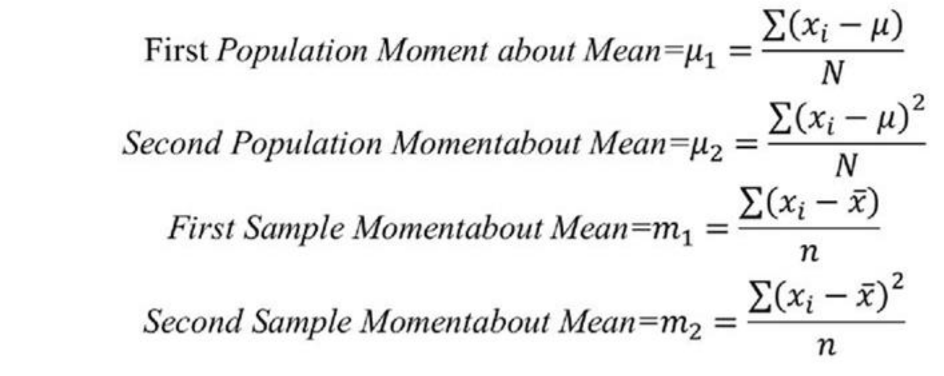
Using Mode,

$$sk_1 = \frac{x - \text{Mode}}{s}$$

Using Median,

$$sk_2 = \frac{3(x - \text{Mode})}{s}$$

3. Standardized skewness measure (moment of distribution):



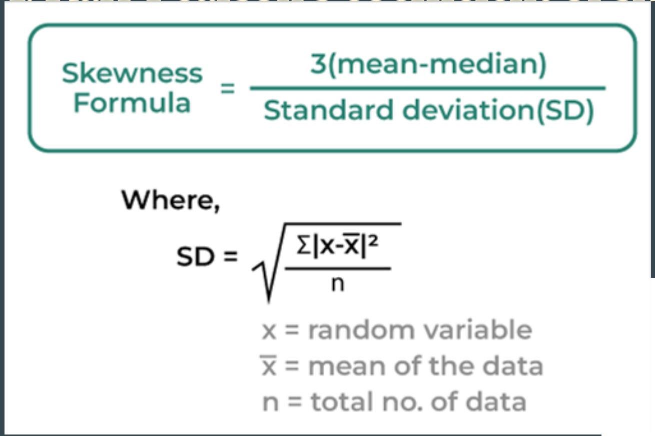
$$\text{First } \textit{Population Moment about Mean}{=}\mu_1 = \frac{\sum(x_i - \mu)}{N}$$

$$\textit{Second Population Momentabout Mean}{=}\mu_2 = \frac{\sum(x_i - \mu)^2}{N}$$

$$\textit{First Sample Momentabout Mean}{=}m_1 = \frac{\sum(x_i - \bar{x})}{n}$$

$$\textit{Second Sample Momentabout Mean}{=}m_2 = \frac{\sum(x_i - \bar{x})^2}{n}$$

4. Karl-Pearson’s coefficient of skewness



$$\text{Skewness Formula} = \frac{3(\text{mean-median})}{\text{Standard deviation(SD)}}$$

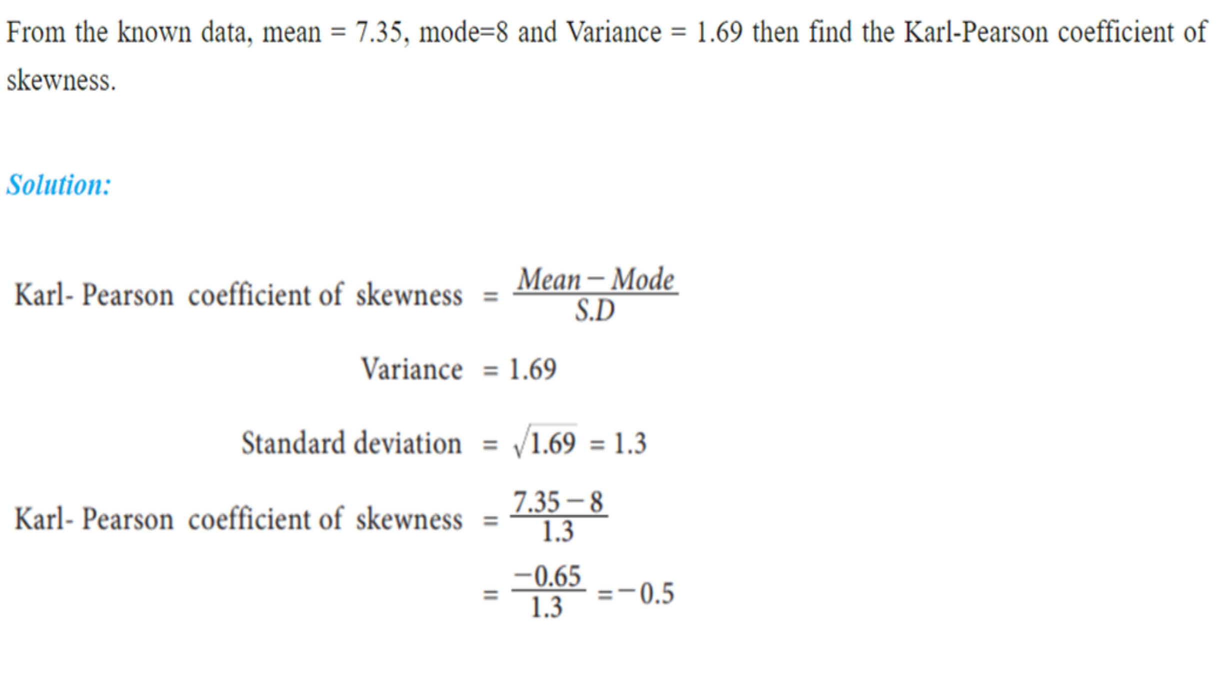
Where,

$$SD = \sqrt{\frac{\Sigma|x-\bar{x}|^2}{n}}$$

x = random variable

$\bar{x}$ = mean of the data

n = total no. of data



From the known data, mean = 7.35, mode=8 and Variance = 1.69 then find the Karl-Pearson coefficient of skewness.

Solution:

$$\text{Karl- Pearson coefficient of skewness} = \frac{Mean - Mode}{S.D}$$

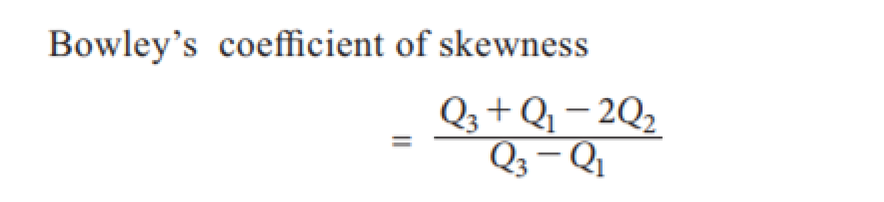
$$\text{Variance} = 1.69$$

$$\text{Standard deviation} = \sqrt{1.69} = 1.3$$

$$\text{Karl- Pearson coefficient of skewness} = \frac{7.35-8}{1.3}$$

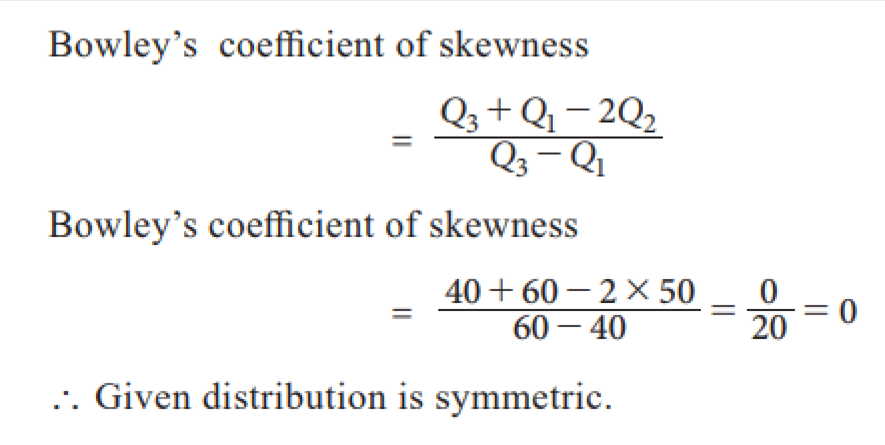
$$= \frac{-0.65}{1.3} = -0.5$$

5. Bowley's coefficient of skewness - based on quartiles



Bowley's coefficient of skewness

$$= \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$



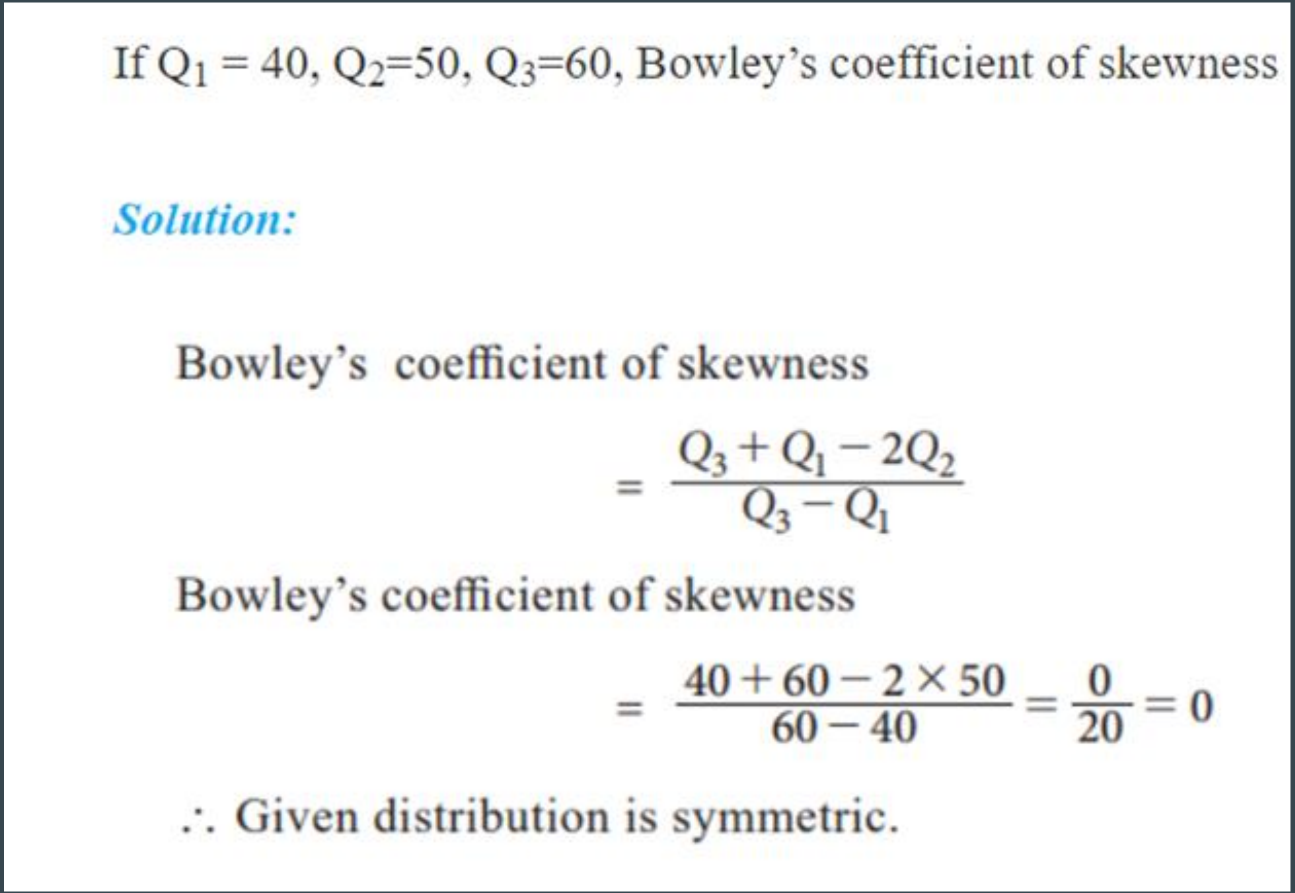
Bowley's coefficient of skewness

$$= \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

Bowley's coefficient of skewness

$$= \frac{40 + 60 - 2 \times 50}{60 - 40} = \frac{0}{20} = 0$$

$\therefore$ Given distribution is symmetric.



If $Q_1 = 40$, $Q_2=50$, $Q_3=60$, Bowley's coefficient of skewness

Solution:

Bowley's coefficient of skewness

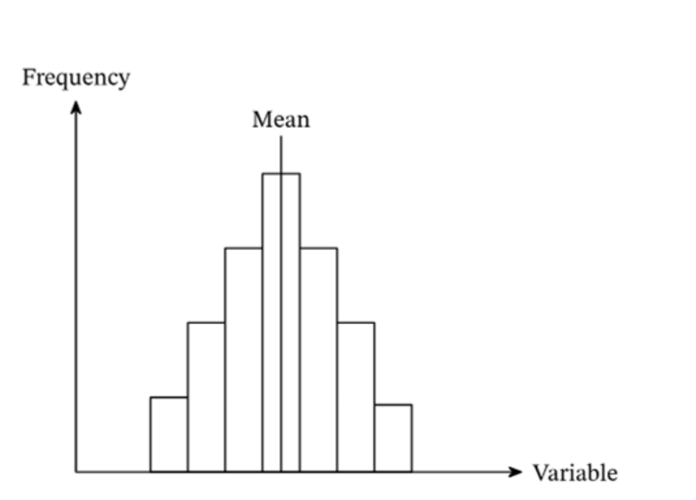
$$= \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

Bowley's coefficient of skewness

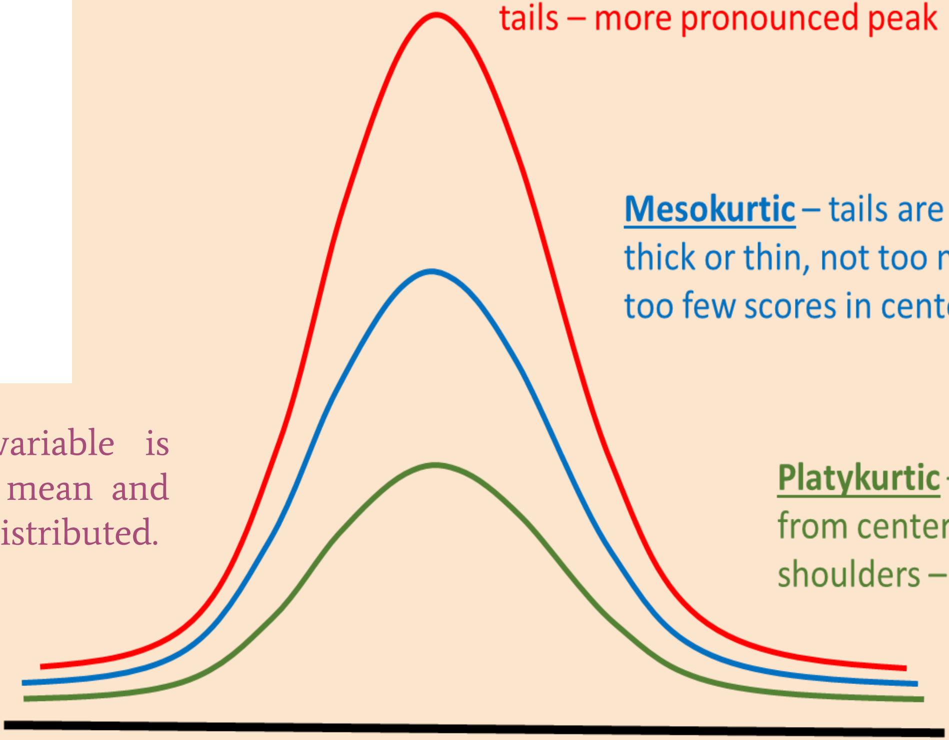
$$= \frac{40 + 60 - 2 \times 50}{60 - 40} = \frac{0}{20} = 0$$

$\therefore$ Given distribution is symmetric.

kurtosis is a measure of 'peakedness' of a distribution



The distribution of a continuous variable is symmetric and concentrated near the mean and the variable is approximately normally distributed.



Leptokurtic – scores move from shoulders into center and to bit to tails – more pronounced peak >3

Mesokurtic – tails are not too thick or thin, not too many or too few scores in center ~3

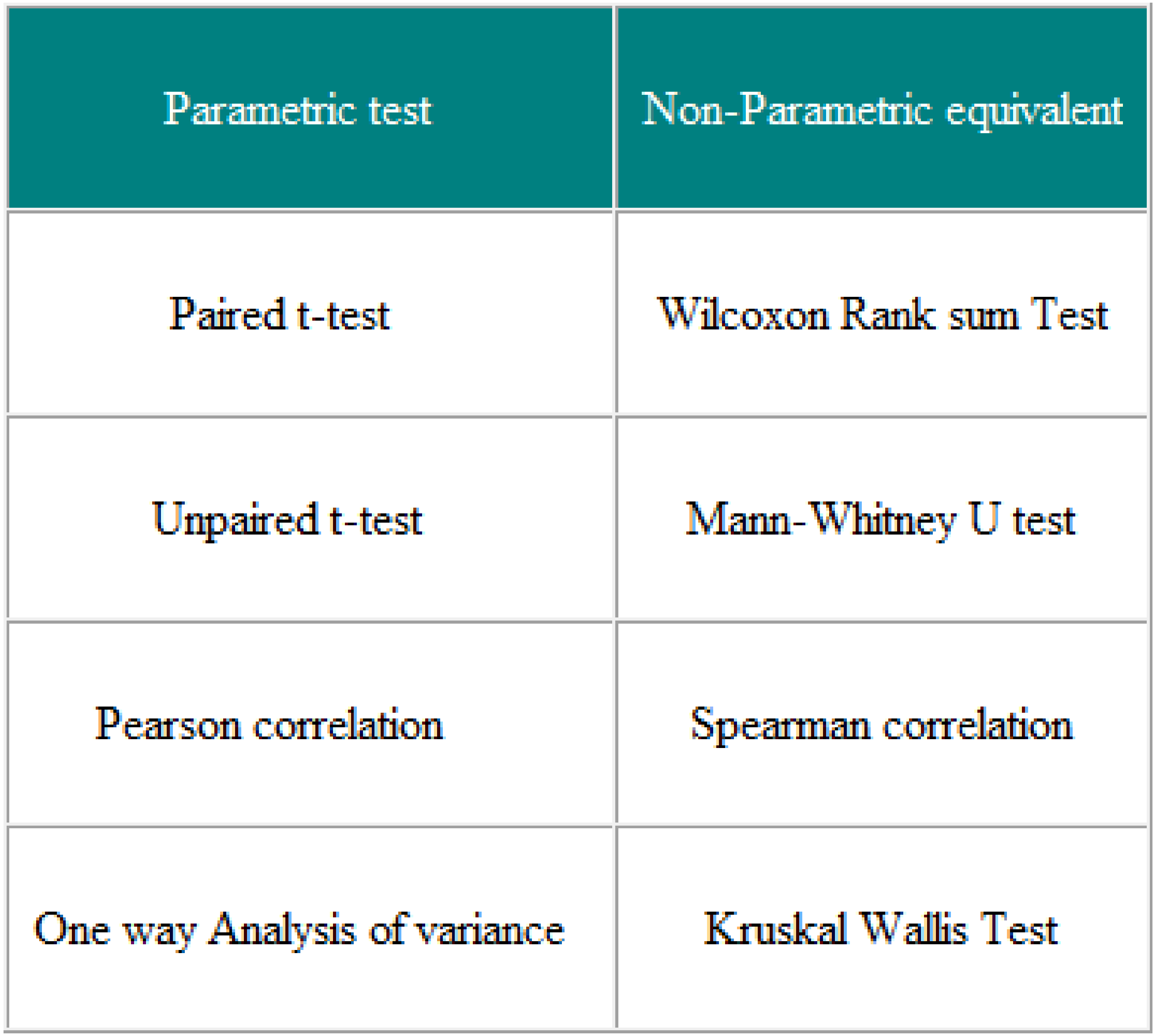
Platykurtic – scores move from center and tails into shoulders – flatter distribution <3

Parametric statistics

Subject Code: 22ZOOC23

Parametric statistics

- Parametric statistics are **based on assumptions about the distribution of population** from which the sample was taken.



Parametric test	Non-Parametric equivalent
Paired t-test	Wilcoxon Rank sum Test
Unpaired t-test	Mann-Whitney U test
Pearson correlation	Spearman correlation
One way Analysis of variance	Kruskal Wallis Test

Population

- Any collection of individuals under study is said to be Population (Universe).
- The individuals are often called the members or the units of the population may be may be physical objects or measurements expressed numerically or otherwise.

Sample

- A part or small section selected from the population is called a sample and process of such selection is called sampling.
- (The fundamental object of sampling is to get as much information as possible of the whole universe by examining only a part of it.
- An attempt is thus made through sampling to give the maximum information about parent universe with the minimum effort).

Parameters

- Statistical measurements such as Mean, Variance etc. of the population are called parameters.

Statistic

- It a statistical measure computed from sample observations alone. The theoretical distribution of a statistics is called its sampling distribution. Standard deviation of the sampling distribution of a statistic is called Standard Error.

Hypothesis

- Statement given by an individual.
- Usually it is required to make decisions about populations on the basis of sample information.
- Such decisions are called **Statistical Decisions**.
- In attempting to reach decisions it is often necessary to make assumption about population involved. Such assumptions, which are not necessarily true, are called **statistical hypothesis**.

Parametric Hypothesis

- A statistical hypothesis which refers only to values of unknown parameters of population is usually called a **parametric hypothesis**.

Null Hypothesis

- A hypothesis which is tested under the assumption that it is true is called a null hypothesis and is denoted by H0.
- Thus a hypothesis which is tested for possible rejection under the assumption that it is true is known as Null Hypothesis.

Alternative Hypothesis

- The hypothesis which differs from the given Null Hypothesis H0 and is accepted when H0 is rejected is called an alternative hypothesis and is denoted by H1 (The hypothesis against which we test the null hypothesis, is an alternative hypothesis

Simple and Composite Hypothesis

- A parametric hypothesis which describes a distribution completely is called a simple hypothesis otherwise it is called **composite hypothesis**.
- For example; In case of **Normal Distribution N** (μ, σ 2), μ = 5, σ = 3 is simple hypothesis whereas μ = 5 is a composite hypothesis as nothing have been said about σ. Similarly, μ < 5, σ = 3 is a composite hypothesis. Let H0: μ = 5 be the null hypothesis, then H1: μ ≠ 5 is two sided composite alternative hypothesis. H1: μ < 5 is one sided (Left) composite alternative hypothesis. H1: μ > 5 is one sided (Right) composite alternative hypothesis.

Test

- Test is a rule through we test the null hypothesis against the given alternative hypothesis.

Tests of Significance

- Procedure which enables us to decide, on the basis of sample information whether to accept or reject the hypothesis or to determine whether observed sampling results differ significantly from expected results are called tests of significance, rules of decisions or tests of hypothesis.

Level of Significance

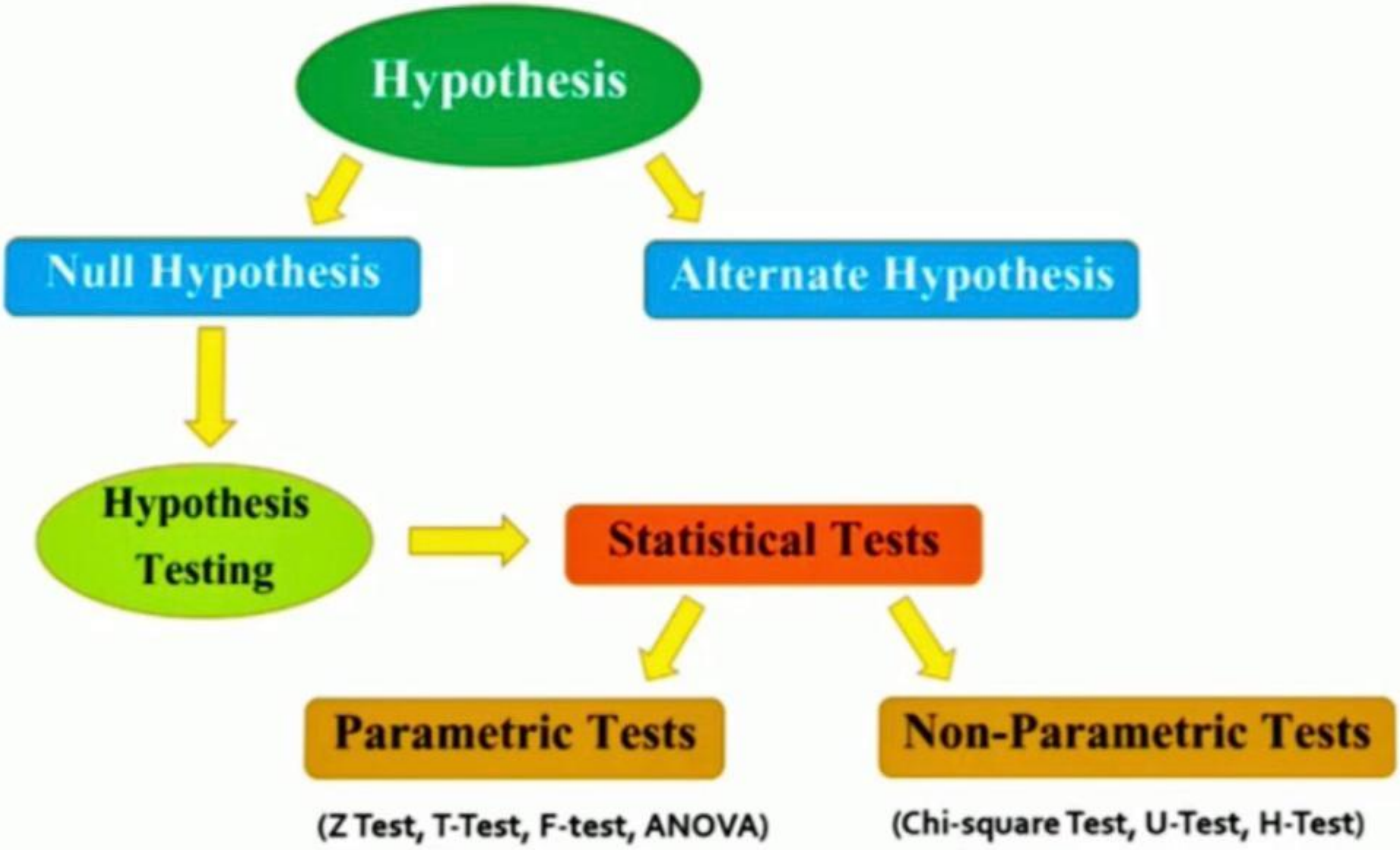
- The probability level below which we reject the hypothesis is called level of significance. The levels of significance usually employed in testing of hypothesis are 5% and 1%.

P-Value

- The p-value is the level of marginal significance within a statistical hypothesis test representing the probability of the occurrence of a given event.
- The p-value is used as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected.

Hypothesis building

- Like other tests, there are two kinds of hypotheses;
  - Null hypothesis and
  - Alternative hypothesis.
- The **alternative hypothesis** assumes that there is a statistically significant difference exists between the means, whereas the **null hypothesis** assumes that there is **no statistically significant** difference exists between the means.



Hypothesis

Null Hypothesis

Alternate Hypothesis

Hypothesis Testing

Statistical Tests

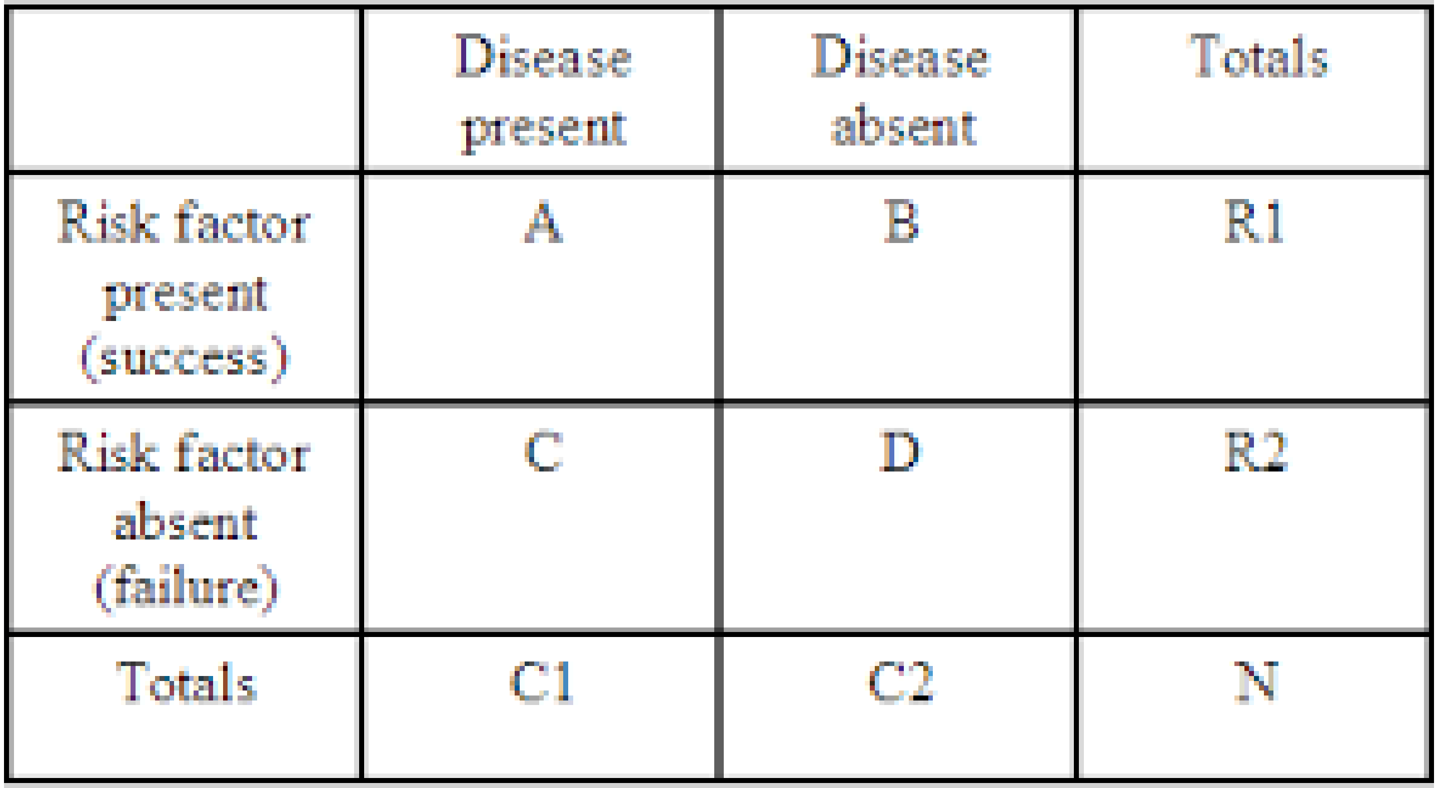
Parametric Tests

(Z Test, T-Test, F-test, ANOVA)

Non-Parametric Tests

(Chi-square Test, U-Test, H-Test)

2X2 contingency table



	Disease present	Disease absent	Totals
Risk factor present (success)	A	B	R1
Risk factor absent (failure)	C	D	R2
Totals	C1	C2	N

		Event	
		Yes	No
Exposure	Yes	a	b
	No	c	d

$$\text{Odds Ratio} = \frac{\text{odds of the event in exposed group}}{\text{odds of the event in non-exposed group}}$$

$$\text{Odds Ratio} = \frac{a/b}{c/d} = \frac{ad}{bc}$$

$$\text{Upper 95\% CI} = e^{[\ln(OR) + 1.96 \sqrt{(1/a) + (1/b) + (1/c) + (1/d)}]}$$

$$\text{Lower 95\% CI} = e^{[\ln(OR) - 1.96 \sqrt{(1/a) + (1/b) + (1/c) + (1/d)}]}$$

$$\text{Relative Risk} = \frac{\text{Probability of Getting Disease if exposed}}{\text{Probability of Getting disease if not exposed}}$$

$$\text{Relative Risk} = \frac{\text{Probability of Outcome if on drug}}{\text{Probability of Outcome if on placebo}}$$

$$\text{Odds Ratio} = \frac{\text{Odds that the diseased were exposed}}{\text{Odds the controls were exposed}}$$

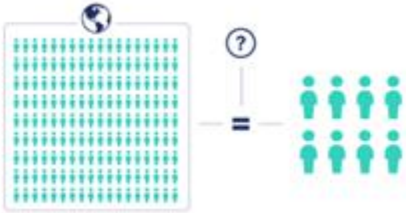
Paired-samples t test

Investigate whether there's a difference within a group between two points in time (within-subjects).

Independent-samples t test

Investigate whether there's a difference between two groups (between-subjects).

One-sample t test



Investigate whether there's a difference between a group and a standard value or whether a subgroup belongs to a population.

Parametric Test

- Parametric tests are those that **make assumptions about the parameters of the population distribution** from which the sample is drawn.
- This is often the assumption that the population data are normally distributed.
- Most of the statistical tests we perform are based on a set of assumptions. When these assumptions are violated the results of the analysis can be misleading or completely erroneous.

- Typical assumptions are:
- Normality: Data have a normal distribution (or at least is symmetric)
- Homogeneity of variances: Data from multiple groups have the same variance
- Linearity: Data have a linear relationship
- Independence: Data are independent

Types of Parametric Test

- T – test
- Z – test
- F – test
- ANOVA

Types of Parametric Test

- T – test
  - **One Sample T-test:** To compare a sample mean with that of the population mean.
- Z – test
  - **One Sample Z-test:** To compare a sample mean with that of the population mean.
- F – test
  - It is a parametric test of hypothesis testing based on **Snedecor F-distribution**.
- ANOVA
  - Also called as **Analysis of variance**, it is a parametric test of hypothesis testing.

Types of Parametric Test

- T – test
  - **One Sample T-test:** To compare a sample mean with that of the population mean.
  - It is essentially, testing the significance of the difference of the mean values when the sample size is small (i.e, less than 30) and when the population standard deviation is not available.
- Assumptions of this test:
  - Population distribution is normal, and
  - Samples are random and independent
  - The sample size is small.
  - Population standard deviation is not known.
- Mann-Whitney 'U' test is a non-parametric counterpart of the T-test.

Types of Parametric Test

- Z – test
  - **One Sample Z-test:** To compare a sample mean with that of the population mean.
  - It is used to determine whether the means are different when the population variance is known and the sample size is large (i.e, greater than 30).
- Assumptions of this test:
  - Population distribution is normal
  - Samples are random and independent.
  - The sample size is large.
  - Population standard deviation is known.
-

Types of Parametric Test

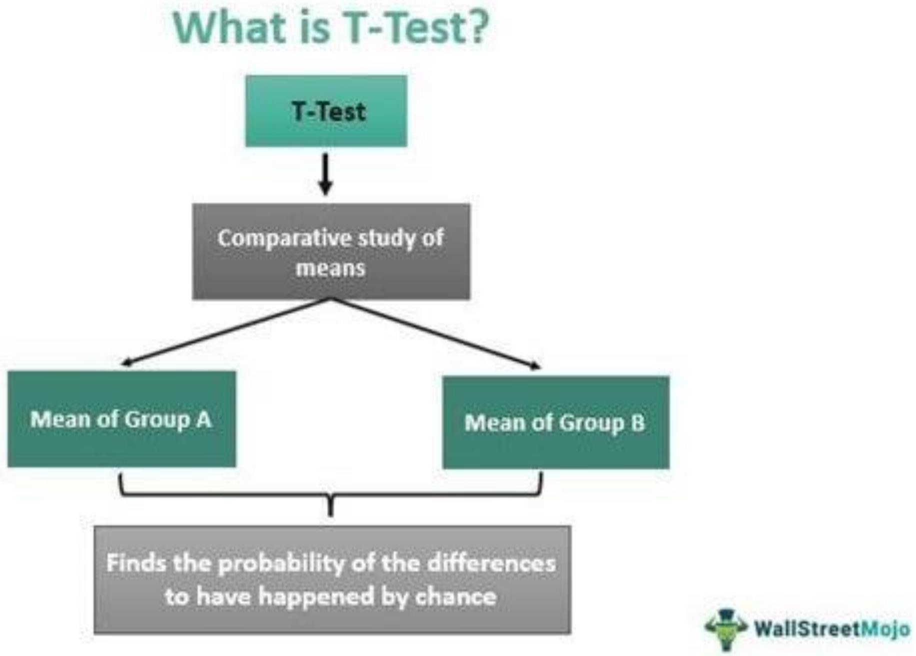
- F – test
  - It is a test for the null hypothesis that two normal populations have the same variance.
  - An F-test is regarded as a comparison of equality of sample variances.
  - F-statistic is simply a ratio of two variances.
- By changing the variance in the ratio, F-test has become a very flexible test. It can then be used to:
  - Test the overall significance for a regression model.
  - To compare the fits of different models and
  - To test the equality of means.
- Assumptions of this test:
  - Population distribution is normal, and
  - Samples are drawn randomly and independently.

Types of Parametric Test

- F – test
  - It is a test for the null hypothesis that two normal populations have the same variance.
  - An F-test is regarded as a comparison of equality of sample variances.
  - F-statistic is simply a ratio of two variances.
- By changing the variance in the ratio, F-test has become a very flexible test. It can then be used to:
  - Test the overall significance for a regression model.
  - To compare the fits of different models and
  - To test the equality of means.
- Assumptions of this test:
  - Population distribution is normal, and
  - Samples are drawn randomly and independently.

Types of Parametric Test

- It is an extension of the T-Test and Z-test.
- It is used to test the significance of the differences in the mean values among more than two sample groups.
- It uses F-test to statistically test the equality of means and the relative variance between them.
- Assumptions of this test:
  - Population distribution is normal, and
  - Samples are random and independent.
  - Homogeneity of sample variance.
- One-way ANOVA and Two-way ANOVA are is types.
- **F-statistic = variance between the sample means/variance within the sample**



What is T-Test?

T-Test

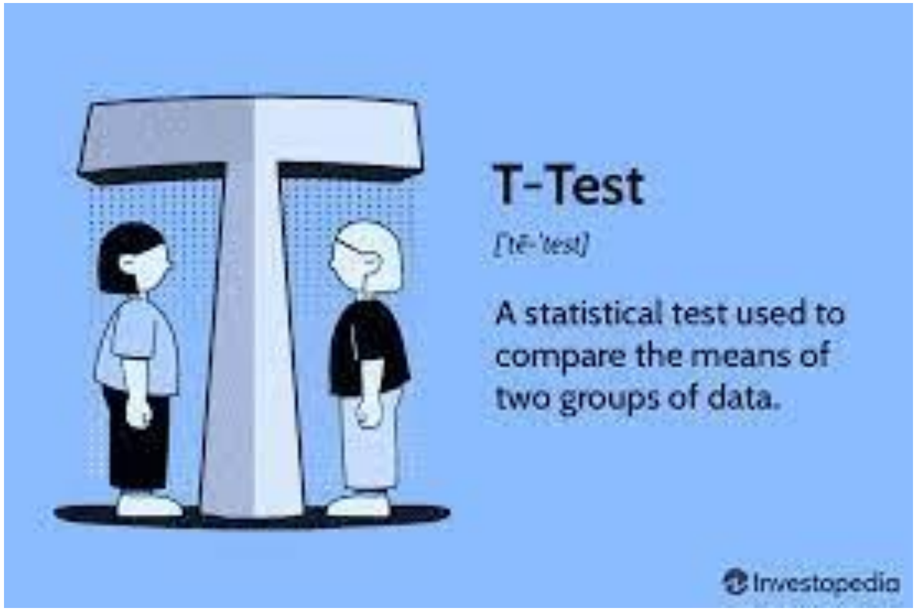
Comparative study of means

Mean of Group A

Mean of Group B

Finds the probability of the differences to have happened by chance

The Student's t test (also called T test) is used to compare the means between two groups

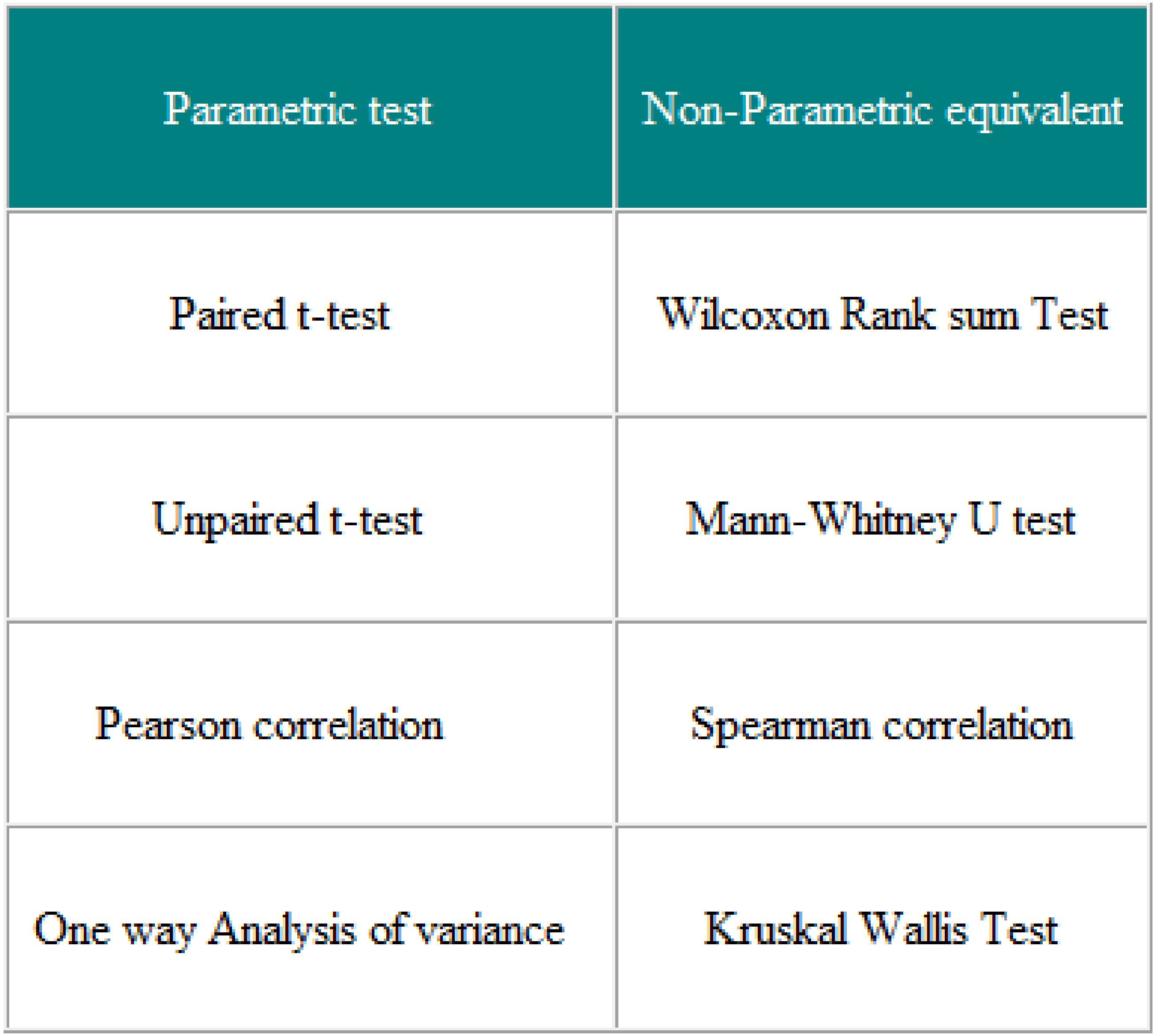


A statistical test used to compare the means of two groups of data.

Non-Parametric statistics
Subject Code: 22ZOOC23

Non-Parametric statistics

- Nonparametric statistics are **not based on assumptions,** that is, the data can be collected from a sample that does not follow a specific distribution.
- For two-group comparisons, either the Mann-Whitney U test (also known as the Wilcoxon rank sum test) is used for independent data or the Wilcoxon signed rank test is used for paired data.



Parametric test	Non-Parametric equivalent
Paired t-test	Wilcoxon Rank sum Test
Unpaired t-test	Mann-Whitney U test
Pearson correlation	Spearman correlation
One way Analysis of variance	Kruskal Wallis Test

Non Parametric Test

- Non-parametric tests are “distribution-free” and, as such, can be used for non-Normal variables.

Nonparametric tests

- Hypotheses are not about population parameters (e.g., $\mu=50$ or $\mu_1=\mu_2$).
- Instead, the null hypothesis is more general.
- For example, when comparing two independent groups in terms of a continuous outcome, the null hypothesis in a parametric test is H_0: $\mu_1 = \mu_2$.

Types of Non-Parametric Tests

- Chi Square Test
  - It is a non-parametric test of hypothesis testing.
  - As a non-parametric test, chi-square can be used:
    - » test of goodness of fit.
    - » as a test of independence of two variables.
- Mann Whitney U Test
  - It is a non-parametric test of hypothesis testing.
- Kruskal Wallis H Test
  - It is a non-parametric test of hypothesis testing.

Chi Square Test

- It helps in assessing the goodness of fit between a set of observed and those expected theoretically.
- It makes a comparison between the expected frequencies and the observed frequencies.
- Greater the difference, the greater is the value of chi-square.
- If there is no difference between the expected and observed frequencies, then the value of chi-square is equal to zero.
- It is also known as the **"Goodness of fit test"** which determines whether a particular distribution fits the observed data or not.

Calculation of Chi Square Test

- Chi-square is also used to test the independence of two variables.
- Conditions for chi-square test:
  - Randomly collect and record the Observations.
  - In the sample, all the entities must be independent.
  - No one of the groups should contain very few items, say less than 10.
  - The reasonably large overall number of items. Normally, it should be at least 50, however small the number of groups may be.
- Chi-square as a parametric test is used as a test for population variance based on sample variance.
  - If we take each one of a collection of sample variances, divide them by the known population variance and multiply these quotients by (n-1), where n means the number of items in the sample, we get the values of chi-square.

Mann Whitney U Test

- This test is used to investigate whether two independent samples were selected from a population having the same distribution.
- It is a true non-parametric counterpart of the T-test and gives the most accurate estimates of significance especially when sample sizes are small and the population is not normally distributed.
- It is based on the comparison of every observation in the first sample with every observation in the other sample.
- The test statistic used here is "U".
- Maximum value of "U" is 'n_1*n_2' and the minimum value is zero.
- It is also known as:
- Mann-Whitney Wilcoxon Test / Mann-Whitney Wilcoxon Rank Test.

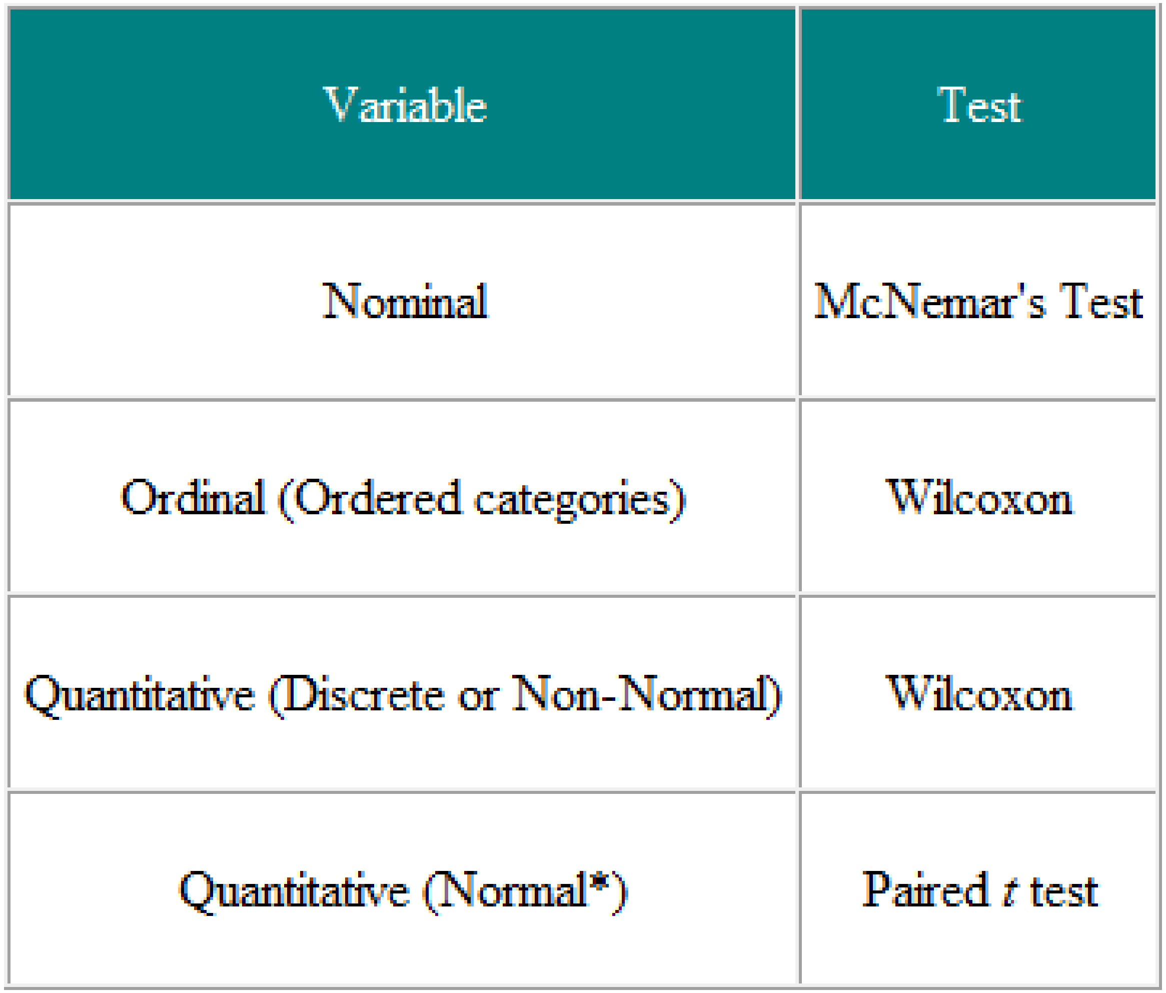
Calculation

- Mathematically, U is given by:
- **$U_1 = R_1 – n_1(n_1+1)/2$**
- where n_1 is the sample size for sample 1, and R_1 is the sum of ranks in Sample 1.
- **$U_2 = R_2 – n_2(n_2+1)/2$**
- When consulting the significance tables, the smaller values of U_1 and U_2 are used. The sum of two values is given by,
- **$U_1 + U_2 = \{ R_1 – n_1(n_1+1)/2 \} + \{ R_2 – n_2(n_2+1)/2 \}$**
- Knowing that $R_1+R_2 = N(N+1)/2$ and $N=n_1+n_2$, and doing some algebra, we find that the sum is:
- **$U_1 + U_2 = n_1*n_2$**

Kruskal Wallis H Test

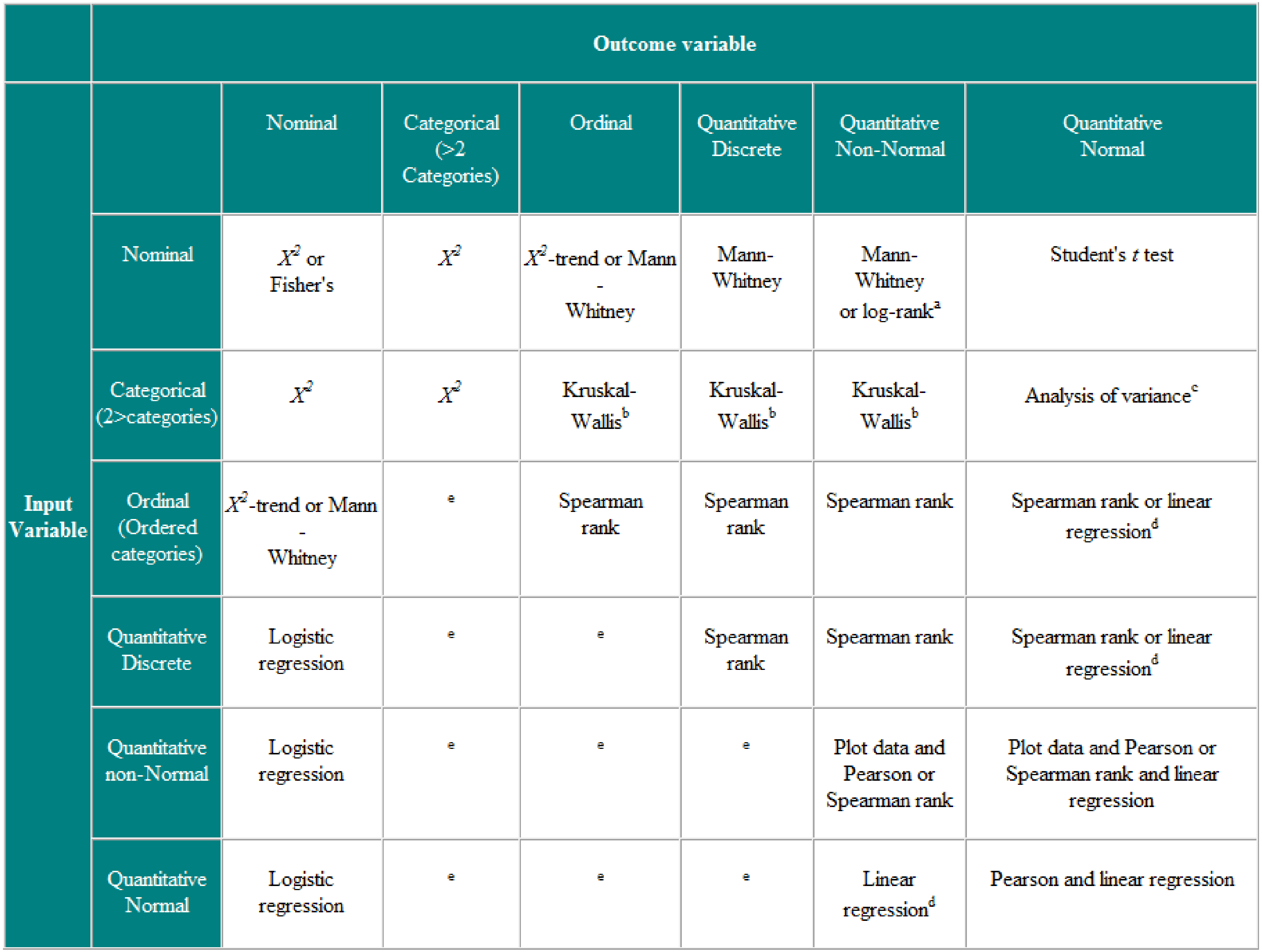
- This test is used for comparing two or more independent samples of equal or different sample sizes.
- It extends the Mann-Whitney-U-Test which is used to comparing only two groups.
- One-Way ANOVA is the parametric equivalent of this test. And that's why it is also known as '**One-Way ANOVA on ranks**.
- It uses ranks instead of actual data.
- It does not assume the population to be normally distributed.
- The test statistic used here is "H".

Choice of statistical test from paired or matched observation



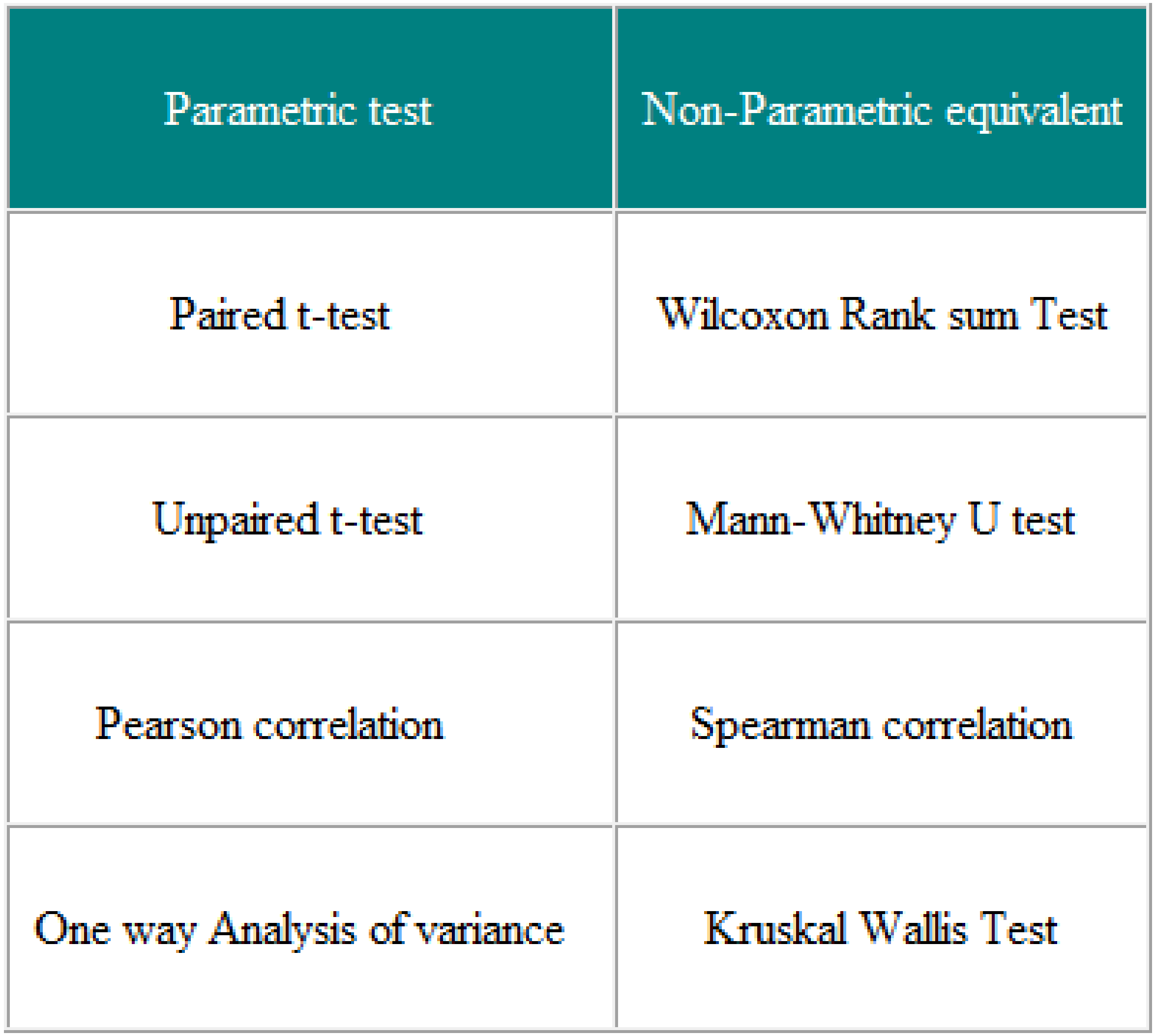
Variable	Test
Nominal	McNemar's Test
Ordinal (Ordered categories)	Wilcoxon
Quantitative (Discrete or Non-Normal)	Wilcoxon
Quantitative (Normal*)	Paired *t* test

Choice of statistical test for independent observations



		Outcome variable					
		Nominal	Categorical (>2 Categories)	Ordinal	Quantitative Discrete	Quantitative Non-Normal	Quantitative Normal
Input Variable	Nominal	X^2 or Fisher's	X^2	X^2-trend or Mann-Whitney	Mann-Whitney	Mann-Whitney or log-rank[a]	Student's *t* test
	Categorical (2>categories)	X^2	X^2	Kruskal-Wallis[b]	Kruskal-Wallis[b]	Kruskal-Wallis[b]	Analysis of variance[c]
	Ordinal (Ordered categories)	X^2-trend or Mann-Whitney	[e]	Spearman rank	Spearman rank	Spearman rank	Spearman rank or linear regression[d]
	Quantitative Discrete	Logistic regression	[e]	[e]	Spearman rank	Spearman rank	Spearman rank or linear regression[d]
	Quantitative non-Normal	Logistic regression	[e]	[e]	[e]	Plot data and Pearson or Spearman rank	Plot data and Pearson or Spearman rank and linear regression
	Quantitative Normal	Logistic regression	[e]	[e]	[e]	Linear regression[d]	Pearson and linear regression

Parametric and Non-parametric tests for comparing two or more groups



Parametric test	Non-Parametric equivalent
Paired t-test	Wilcoxon Rank sum Test
Unpaired t-test	Mann-Whitney U test
Pearson correlation	Spearman correlation
One way Analysis of variance	Kruskal Wallis Test

Non-parametric test	Equivalent parametric test	Purpose of statistical test	Example
Wilcoxon rank-sum test	Paired t-test	Compares the mean value of two variables obtained from the same participants	The difference in depression scores before and after treatment
Mann-Whitney U test	Unpaired t-test	Compares the mean value of a variable measured from two independent groups	The difference between depression symptom severity in a placebo and drug therapy group
Spearman correlation	Pearson correlation	Measures the relationship (strength/direction) between two variables	The relationship between fitness test scores and the number of hours spent exercising
Kruskal Wallis test	One-way analysis of variance (ANOVA)	Compares the mean of two or more independent groups (uses a between-subject design, and the independent variable needs to have three or more levels)	The difference in average fitness test scores of individuals who frequently exercise, moderately, or do not exercise
Friedman's ANOVA	One-way repeated measures ANOVA	Compares the mean of two or more dependent groups (uses a within-subject design, and the independent variable needs to have three or more levels)	The difference in average fitness test scores during the morning, afternoon, and evening

Paired vs Unpaired T test

Paired T test	Unpaired T test
A paired t-test is designed to compare the means of the same group or item under two separate scenarios.	An unpaired t-test compares the means of two independent or unrelated groups.
In a paired t-test, the variance is not assumed to be equal	In an unpaired t-test, the variance between groups is assumed to be equal.

Sign Test

- The sign test is a rank test in which the test statistic is calculated by forming differences in paired samples of dependent groups.
- The differences in paired samples are formed by allocating every value from the first group to the respective value from the second group.
- The two groups must have the same sample size.

Yates correction

- The effect of Yates' correction is to prevent the overestimation of statistical significance for small data when 'zero cells' are present in a 2 × 2 contingency table.
- Such zero cells are reported to overestimate the OR measure and the corresponding standard deviation (SD).

Where to apply

- Yates' Correction is used with chi-squared analysis under certain conditions
- Yates' Correction is typically used in X2 analysis with 1 degree of freedom where expected frequencies of less than 10 are found

Simpson’s paradox

- Simpson's Paradox is a statistical phenomenon where an association between two variables in a population emerges, disappears or reverses when the population is divided into subpopulations
- Eg Gender Bias