

# Statistics

# Stats is all about data

## Raw data

- May have errors
- Not validated
- Unformatted
- Uninterpretable
- Not cleansed
- ...

State	District	Jun	Jul	Aug
Andhra Pradesh	Adilabad	213.245	260.107	449.676
Andhra Pradesh	Vizianagaram	265.521	139.41	266.612
Arunachal Pradesh	Changlang	264.225	323.216	371.473
Arunachal Pradesh	Dibang Valley	214.674	316.565	336.434
Assam	Karimganj	164.937	270.082	589.803
Assam	Kokrajhar	580.901	312.208	564.161
Bihar	Supaul	173.777	152.199	200.673
Bihar	Vaishali	126.427	120.119	134.309
Chandigarh	Chandigarh	87.6	236.5	134.6
Chattisgarh	Bastar	318.126	255.674	366.698
Chattisgarh	Rajnandgaon	213.481	378.729	229.806
Chattisgarh	Surguja	227.882	210.418	159.516
Dadra & Nagar Haveli	Dadra & Nagar Haveli	341.727	603.201	234.86
Delhi	New Delhi	80.69	272.234	125.493
Gujarat	Ahmadabad	55.405	335.661	81.557
Gujarat	Amreli	55.892	376.289	103.858
Gujarat	The Dangs	280.156	585.72	242.533
Haryana	Ambala	93.162	237.152	141.453
Haryana	Bhiwani	58.104	219.684	75.383
Himachal Pradesh	Chamba	90.188	145.487	141.654
Himachal Pradesh	Hamirpur	96.383	201.116	147.078

Rainfall in Districts of Indian states during the monsoon season

## Processed data

- No errors
- Validated
- Formatted
- Interpretable
- Cleansed
- ...

State	Average (cm)
Andhra Pradesh	239.383
Arunachal Pradesh	239.4495
Assam	372.919
Bihar	150.102
Chandigarh	87.6
Chattisgarh	265.8035
Dadra & Nagar Haveli	341.727
Delhi	80.69
Gujarat	55.6485
Haryana	75.633
Himachal Pradesh	93.2855

Average rainfall (in cms.) in Indian states

# About data

## Data collection

Data is collected in different ways:

- Census
- Observation
- Convenience sample
- Random samples
- Historical data (data collected over time)
- Any other

## Data forms

Data can be in any of these forms

- Structured (rows and columns)
- Semi-structured (XML / JSON)
- Unstructured (free text)

## Data collected method

- Batch
- Real-time

# Data Types

## Numeric

Numeric data can be of 2 types:

- **Discrete data**

**Eg:**

**Year** → 1972,1998,2005,2018..

**Age** → 12,18,24,39,40..

- **Continuous data**

**Eg:**

**Weight** → 43.1,55.4,76.9 ..

**Temperature** → 31.1,33.4,90.5 ...

## Character

Character data can be of 2 types:

- **Strings and literals**

**Eg:** "A", "computer", "Statistics" ...

- **Factors / Categorical**

## Date

# Factor Data Types

Nominal

Ordinal

Interval

Ratio

# Factor Data Types

**Nominal**

Ordinal

Interval

Ratio

- Used as names / labels without any quantitative measure
- No numerical significance

- Examples:

## Gender

Male  
Female

## Marital Status

Single  
Married  
Divorced

## Religion

Hindu  
Jain  
Buddhism  
Sikh  
Christian

# Factor Data Types

Nominal

Ordinal

Interval

Ratio

- Order is important, rather than the name
- Difference between 2 values is not really known

- Examples:

## Income Level

- 1 = Low
- 2 = Middle
- 3 = High
- 4 = Very high

## Feeling today

- 1 = Very unhappy
- 2 = Unhappy
- 3 = Ok
- 4 = Happy
- 5 = Very happy

## Rating

- 3 = Very Good
- 2 = Good
- 1 = Bad

# Factor Data Types

Nominal

Ordinal

Interval

Ratio

- Numerical scales where order and difference are known
- Do not have a true 0

- Examples:

Temperature

Time

Marks

Temp.

0

5

10

15

20

25

Marks

90-100

80-89

70-79

60-69

50-59

40-49

30-39

20-29

0-19

Freq.

2

3

7

11

15

3

4

5

0



# Data Types

Nominal

Ordinal

Interval

Ratio

- Numerical scales where order and difference are known
- Has a true 0 (means “does not exist”)
- Descriptive and Inferential statistical analysis performed

- Examples:

Height

Weight

Age

Income

Years of education

# Data Types cont...

## Long format

Product	Attribute	Value
P1	Weight	18
P1	Colour	Black
P1	Price	15675
P2	Height	10
P2	Price	1980

## Wide format

Product	Weight	Height	Colour	Price
P1	19	6	Blue	15789
P2	11	2	Black	1900

It is easier to read and interpret a wide format data

# Data Mining

- To discover patterns in a dataset involving statistics, database concepts and machine learning
- Extract this information to transform them into useful interpretations
- Data Mining is an essential part of a process commonly known as KDD (Knowledge discovery in databases)

## **KDD (Knowledge Discovery in Databases) - Some key characteristics**

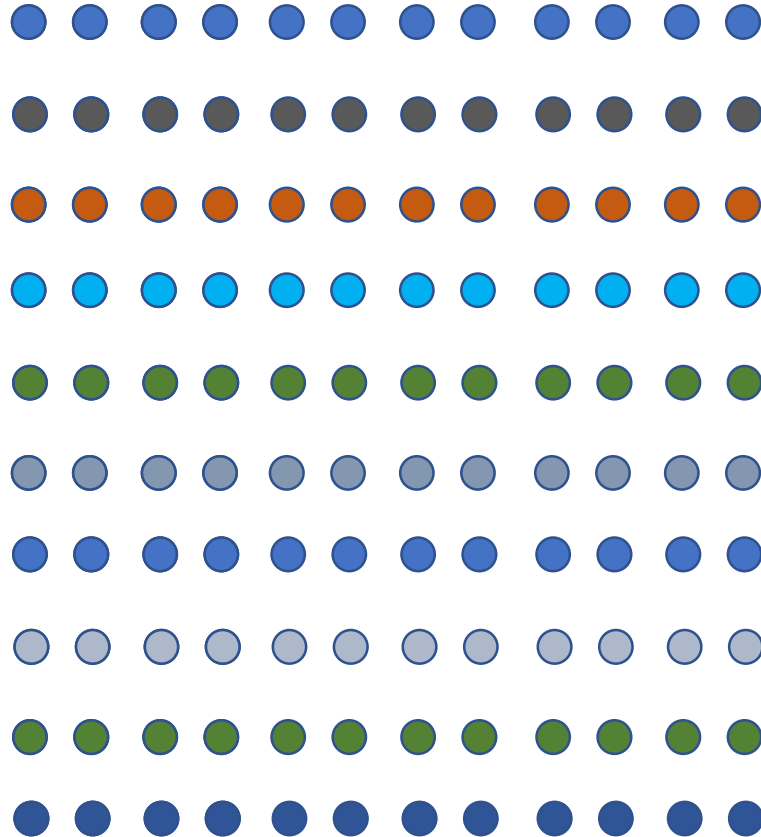
- Data preparation and selection
- Data cleansing
- Incorporating prior knowledge on data sets
- Interpreting accurate solutions from the observed results

## **Data mining – some typical tasks**

- Anomaly detection in data
- Association rule analysis (Apriori)
- Clustering (Unsupervised machine learning)
- Classification and Regression (Supervised machine learning)
- Visualization

# Population vs Sample

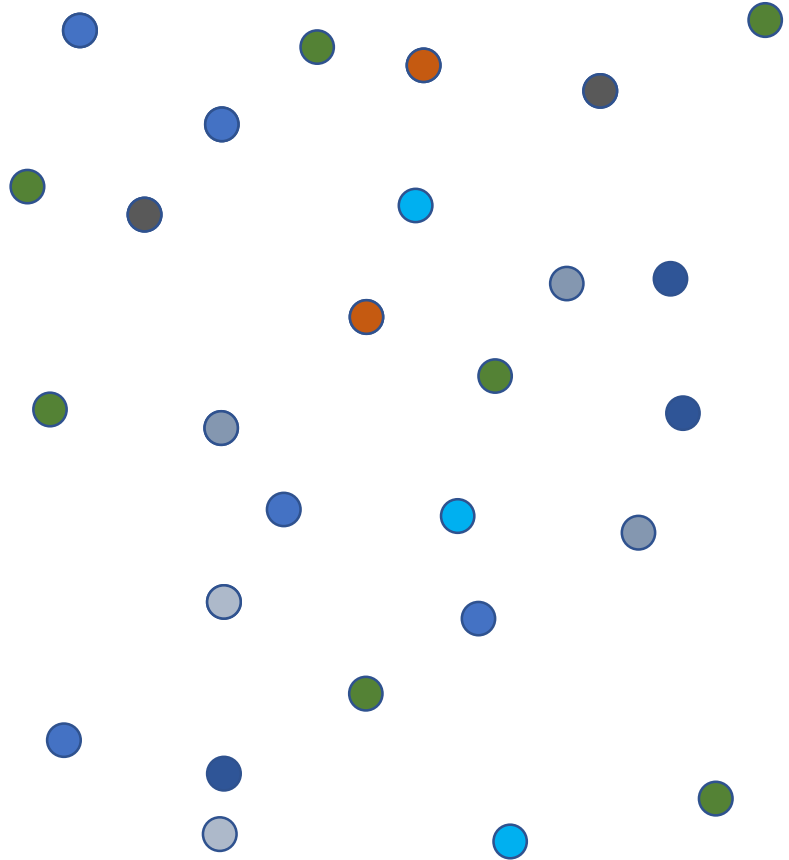
Population



Parameter

$\mu$  ← Mean →  $\bar{X}$   
 $\sigma$  ← Standard Deviation →  $s$

Sample



Statistic

# Types of Statistics

## Descriptive Statistics

*Describes the various aspects of dataset*

Measure of Central  
Tendency

Mean  
Weighted Mean  
Geometric Mean  
Median  
Mode

Measure of Dispersion

Range  
Interfractile Range  
Quartiles  
Interquartile Range  
Standard Deviation  
Variance

Measure of Association

Correlation  
Covariance  
Coefficient of Covariation  
Rank Correlation

## Inferential Statistics

*What conclusion can be drawn from the dataset*

Estimation

Point Estimate  
Range / Interval Estimate

Hypothesis Testing

# **Descriptive Statistics**

# **I. Measure of Central tendency**

Central tendency measures the centre value / middle value / average value of a given dataset

- 1. Mean**
- 2. Weighted Mean**
- 3. Geometric Mean**
- 4. Median**
- 5. Mode**

# 1. Mean

- Arithmetic mean is the “average” of a range of data (numeric)
- Common examples: Test Marks, Temperature, Runs scored in cricket etc.
- Conventional symbols:
  - ✓  $n$  = sample size
  - ✓  $x$  = observation(s)
  - ✓  $\bar{x}$  = sample mean
  - ✓  $\mu$  = population mean
- Arithmetic Mean  $\bar{x} = (\sum x / n)$

## Advantages

- A single number represents a whole **dataset**
- Intuitively clear
- Only 1 mean per dataset – easy for comparison

## Disadvantages

- Affected by **extreme values** – so not a reliable measure
- Every value is taken for calculation (use **grouped data**)
- Cannot compute mean for **open-ended classes**



## 2. Weighted Mean

- Calculate Average by considering the importance of each value to the overall total

### Actual

Student	Homework	Quiz	Assignment	Term	Final
1	85	89	94	87	90
2	78	84	88	91	92
3	94	88	93	86	89
4	82	79	88	84	93
5	95	90	92	82	88

**Overall = 20%\*Homework +  
10%\*quiz +  
10%\*Assignment +  
25%\*Term +  
35%\*Final**

### Weighted mean

Student	Homework	Quiz	Assignment	Term	Final	Overall
1	17	8.9	9.4	21.75	31.5	88.55
2	15.6	8.4	8.8	22.75	32.2	87.75
3	18.8	8.8	9.3	21.5	31.15	89.55
4	16.4	7.9	8.8	21	32.55	86.65
5	19	9	9.2	20.5	30.8	88.5

### 3. Geometric Mean

- Quantities change over a period of time; need to know average rate of change
- A good example is '*Interest rate*'
- Each value of a new year depends on the value of the previous year
- ***Growth factor*** =  $1 + (\text{interest\_rate}/100)$

- $GM = \sqrt[n]{\text{Product of all } X \text{ values}}$

year	Amount (a)	interest (%)	growth_factor (gf)	savings (a*gf)
1	100.00	7	1.07	107.00
2	107.00	8	1.08	115.56
3	115.56	10	1.1	127.12
4	127.12	12	1.12	142.37
5	142.37	18	1.18	168.00

Excel formula

C1							
	A	B	C	D	E	F	G
1	107		130.3247				
2	115.6						
3	127.1						
4	142.4						
5	168						

#### Geometric Mean

$$\sqrt[5]{107 \times 115.56 \times 127.12 \times 142.37 \times 168} \\ = 130.32$$

## 4. Median

- **Position based single value** that measures the central item in a dataset
- Middlemost / Centremost item in a dataset
- About half of the items lie above this point; and the other half below it
- To calculate Median, data needs to be sorted (Ascending / Descending)
- Formula for Median
  - ✓ For non-grouped data
    - $[(n+1) / 2]^{\text{th}}$  item, when  $n$  is odd
    - $[(n/2)^{\text{th}} + ((n/2)+1)^{\text{th}}] / 2$  item, when  $n$  is even
  - ✓ For grouped data
    - $(n/2)^{\text{th}}$  item

### Advantages

- Extreme values do not effect Median strongly
- Easy to calculate

### Disadvantages

- Needs sorting of data before calculation
- Can be time consuming in large datasets

# Median calculation

Item	Time
1	10.2
2	10.3
3	10.7
4	10.8
5	11
6	11.1
7	15

**n = 7**

$[(n+1)/2]^{\text{th}} = 4^{\text{th}}$

4<sup>th</sup> element = 10.8

Median = 10.8

## Exercise: Calculate the Median

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
42	53	90	81	120	41	42	29	87	11	35	69	40	77	97	63

### 1: Sort the dataset

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
11	29	35	40	41	42	42	53	63	69	77	81	87	90	97	120

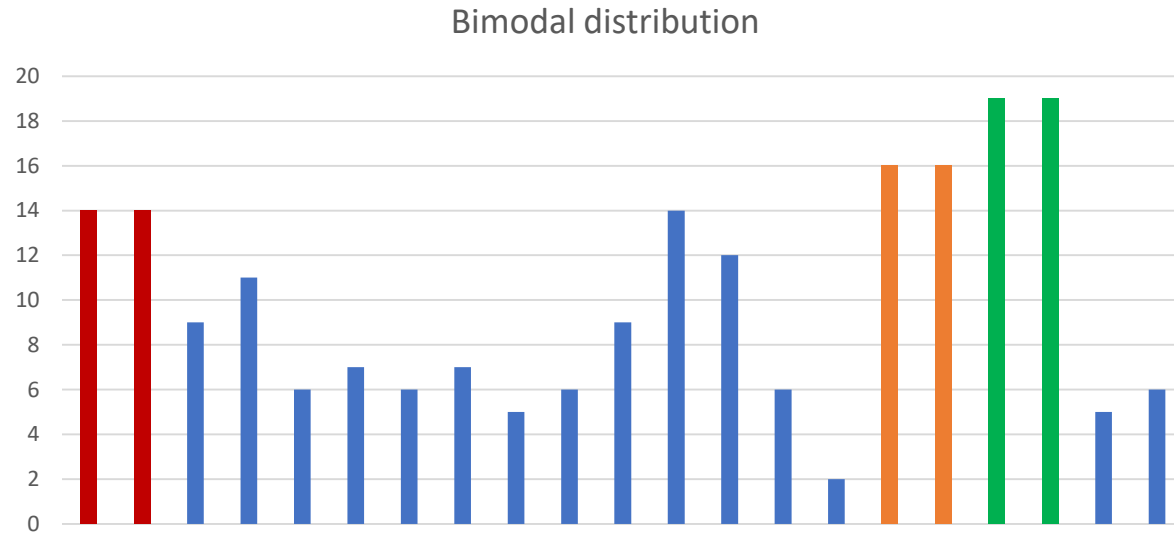
**n = 16**

$[(n/2)^{\text{th}} + ((n/2)+1)^{\text{th}}] / 2$

$[8^{\text{th}} + 9^{\text{th}}] / 2 = (53 + 63) / 2 = 58 \rightarrow \text{Median}$

# 5. Mode

- A single value that is repeated most often
- Used for both qualitative and quantitative data
- Bimodal distribution – different values repeated same number of times



## Advantages

- Not affected by extreme values
- Can be used even for open-ended classes

## Disadvantages

- Datasets may not contain repeated values
- In case of many modes, interpretation may be difficult

Unit produced / hour

1	3	7	10	20
1	3	7	11	20
2	5	9	12	20
2	5	9	16	24

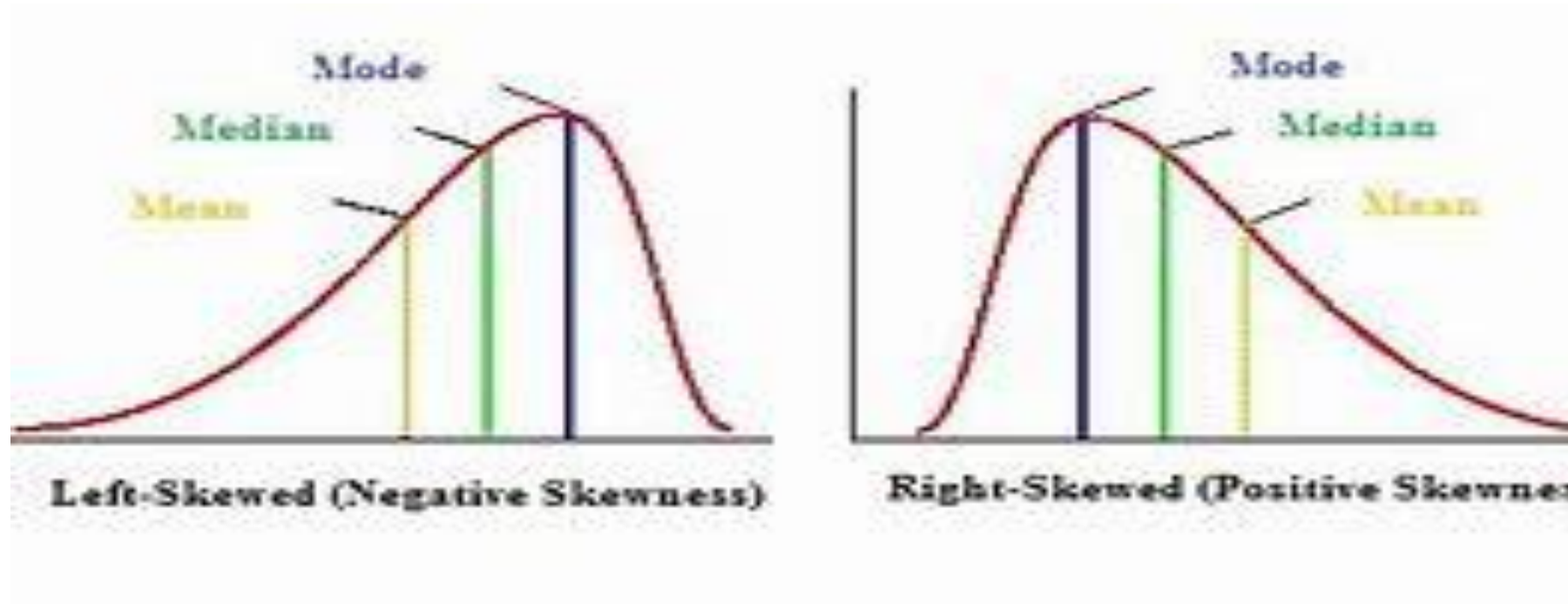
Most popular colours for dresses

Black	White	Green	Blue	Blue
Blue	Green	Yellow	White	Green
Black	Pink	Black	Yellow	Green
Green	Green	Green	Blue	Blue

Frequency Distribution

0-2	4
3-9	9
10-20	7
> 20	1

## Representing the Mean, Median and Mode graphically



### Quiz

For a symmetrical curve, what will be the Mean, Median and Mode ?

It will be the same

# Statistical measures on Data Types

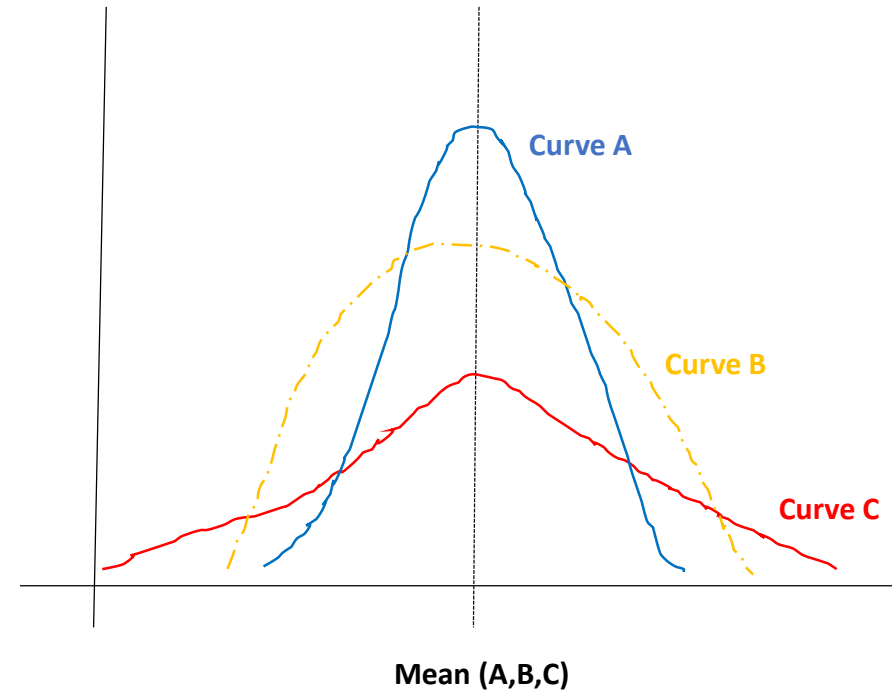
Statistical measures

	Nominal	Ordinal	Interval	Ratio
Mode	<input checked="" type="checkbox"/> Yes	<input checked="" type="checkbox"/> Yes	<input checked="" type="checkbox"/> Yes	<input checked="" type="checkbox"/> Yes
Median	<input checked="" type="checkbox"/> No	<input checked="" type="checkbox"/> Yes	<input checked="" type="checkbox"/> Yes	<input checked="" type="checkbox"/> Yes
Mean	<input checked="" type="checkbox"/> No	<input checked="" type="checkbox"/> No	<input checked="" type="checkbox"/> Yes	<input checked="" type="checkbox"/> Yes
Frequency distribution	<input checked="" type="checkbox"/> Yes	<input checked="" type="checkbox"/> Yes	<input checked="" type="checkbox"/> Yes	<input checked="" type="checkbox"/> Yes
Range	<input checked="" type="checkbox"/> No	<input checked="" type="checkbox"/> Yes	<input checked="" type="checkbox"/> Yes	<input checked="" type="checkbox"/> Yes
Add & Subtract	<input checked="" type="checkbox"/> No	<input checked="" type="checkbox"/> No	<input checked="" type="checkbox"/> Yes	<input checked="" type="checkbox"/> Yes
Multiply & Divide	<input checked="" type="checkbox"/> No	<input checked="" type="checkbox"/> No	<input checked="" type="checkbox"/> No	<input checked="" type="checkbox"/> Yes
Standard deviation	<input checked="" type="checkbox"/> No	<input checked="" type="checkbox"/> No	<input checked="" type="checkbox"/> Yes	<input checked="" type="checkbox"/> Yes

## II. Measure of Dispersion

Dispersion measures the spread or variability of data

1. Range
2. Quartiles
3. Interquartile Range
4. Variance
5. Standard Deviation





# 1. Range

- Difference between the highest and the lowest observed values in a dataset
- Easy to understand and find
- Usefulness as a dispersion measure is limited – only 2 values are considered
- Heavily influenced by extreme values
- Range values may change from one sample to another
- For open-ended class, there is no *range*

## Example

Values	Max	Min	range
22	90	6	84
49			
78			
6			
78			
76			
44			
90			
18			
63			
49			
62			

## 2. Quartiles

- Division of data into 4 segments according to the distribution of values
- The width of the four quartiles need not be the same
- Each part contains 25% data
- Quartiles are the highest values in each of the 4 parts
- Formula to calculate the quartiles:
  - ❑  $Q1 = [(n+1)/4]^{\text{th}}$  value
  - ❑  $Q2 = [(n+1)/2]^{\text{nd}}$  value
  - ❑  $Q3 = [3(n+1)/4]^{\text{th}}$  value

Lowest observation

Q1

Q2

Q3

Q4

Highest observation

S.No	Data
1	10
2	11
3	14
4	16
5	17
6	18
7	19
8	21
9	21
10	23
11	24
12	26
13	29
14	30
15	32
16	33
17	34
18	35
19	36
20	37
21	39
22	40
23	42
24	45

quartile	value	interpretation
1st quartile	18.75	25% values are $\leq 18.75$
2nd quartile	27.5	50% values are $\leq 27.5$
3rd quartile	35.25	75% values are $\leq 35.25$
4th quartile	45	100% values are $\leq 45$

### Excel calculation

quartile	formula
1st quartile	= QUARTILES(<range>,1)
2nd quartile	= QUARTILES(<range>,2)
3rd quartile	= QUARTILES(<range>,3)
4th quartile	= QUARTILES(<range>,4)

### Q1

S.No	Data
1	10
2	11
3	14
4	16
5	17
6	18

### Q2

S.No	Data
1	10
2	11
3	14
4	16
5	17
6	18
7	19
8	21
9	21
10	23
11	24
12	26

### Q3

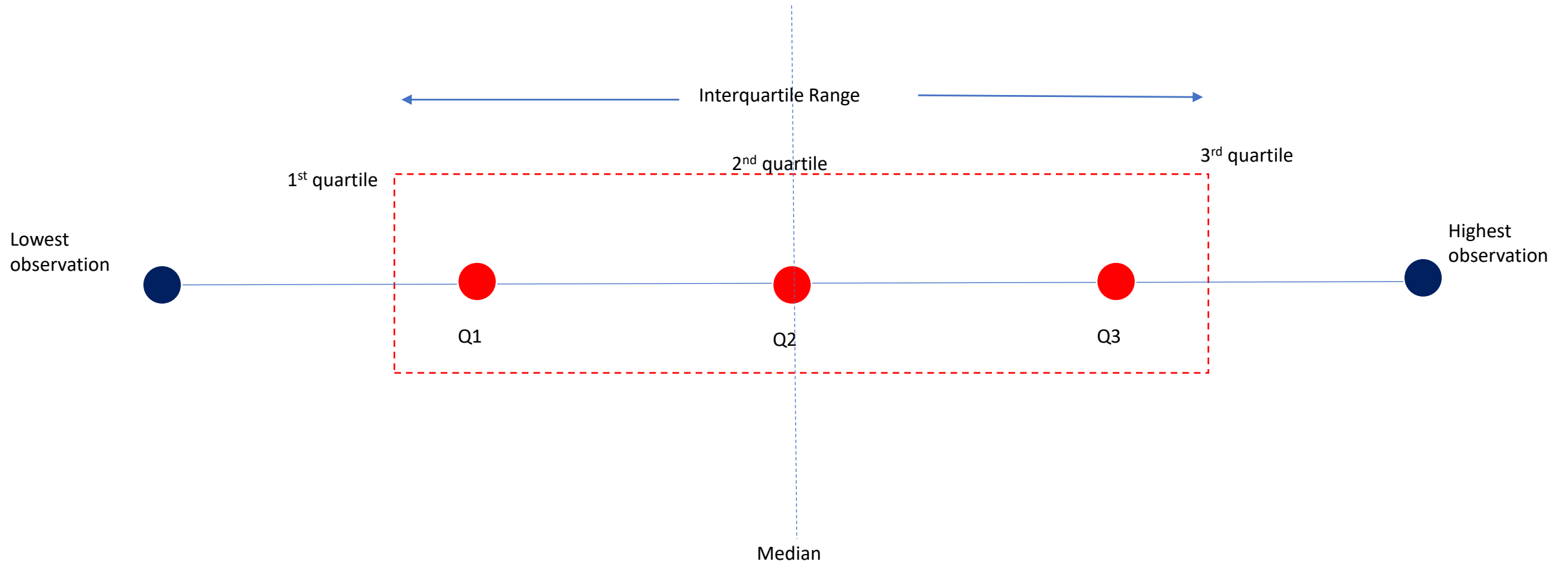
S.No	Data
1	10
2	11
3	14
4	16
5	17
6	18
7	19
8	21
9	21
10	23
11	24
12	26
13	29
14	30
15	32
16	33
17	34
18	35

### Q4

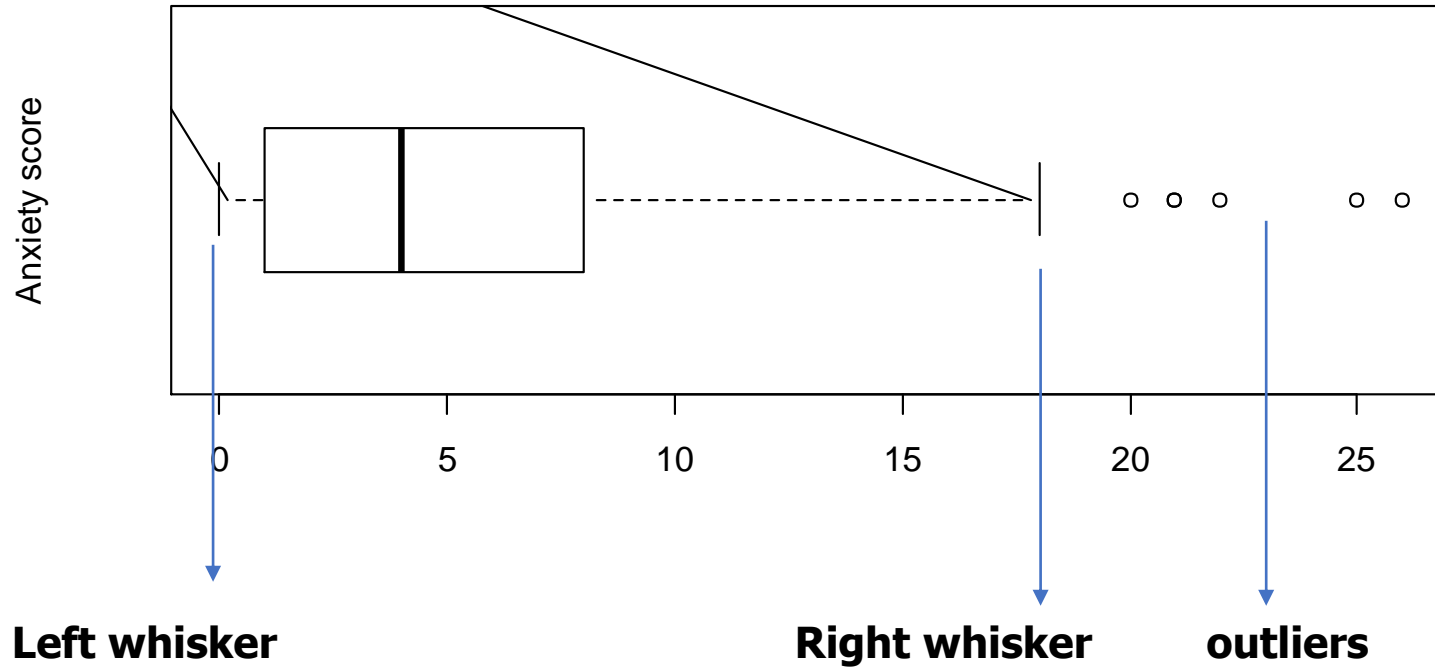
S.No	Data
1	10
2	11
3	14
4	16
5	17
6	18
7	19
8	21
9	21
10	23
11	24
12	26
13	29
14	30
15	32
16	33
17	34
18	35
19	36
20	37
21	39
22	40
23	42
24	45

### 3. Interquartile Range

- Approximately measures how far from the median on either side to include one-half of data
- IQR is the difference between the values of the first and third quartiles



**Boxplot for Anxietyscore**



Min	Q1	Q2	Q3	Max	IQR
0	1	4	8	26	7

**left whisker** = 1st quartile - (3/2 of IQR)

**right whisker** = 3rd quartile + (3/2 of IQR)

## 4. Variance

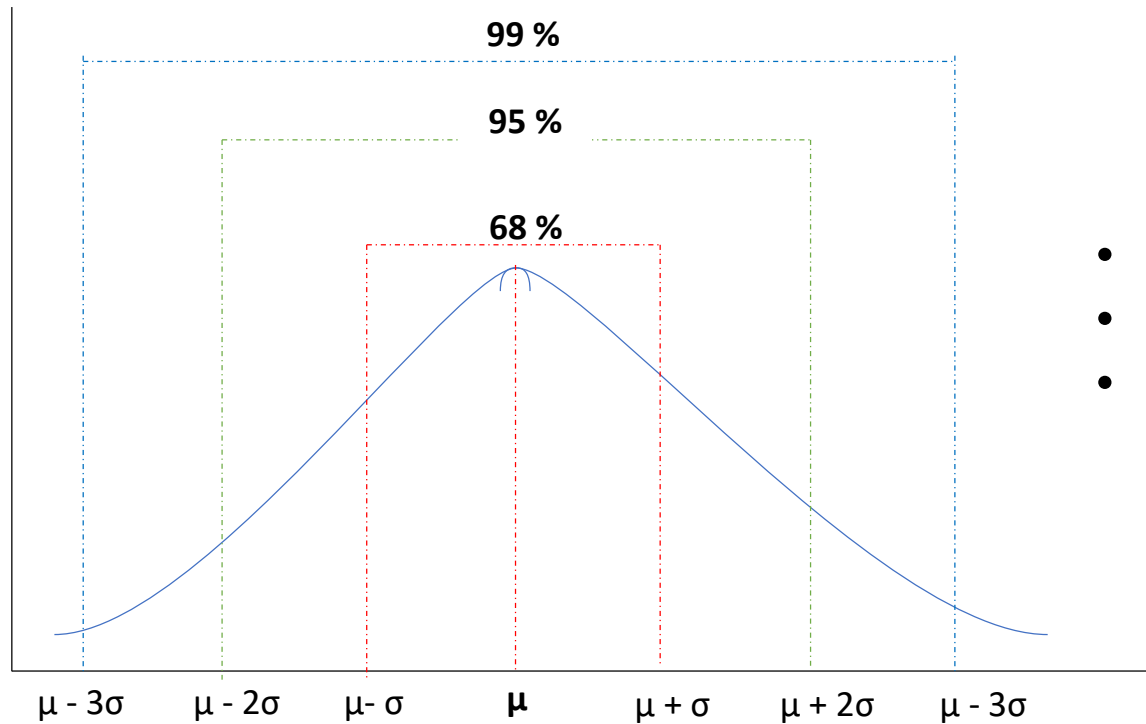
- Average deviation from some measure of central tendency
- Every population / sample has variance
- Represented by the symbol  $\sigma^2$
- Formula to calculate variance
$$\sigma^2 = (\sum(x - \mu)^2) / N$$
  - $\sigma^2$  : population variance
  - $x$  : observed value
  - $\mu$  : population mean
  - $N$  : total number of items in population
- Units of variance are *squares of units* of data – eg: squared miles, squared rupees etc.
- Not intuitively clear or interpreted in the right way

## 5. Standard Deviation

- Square root of the average of the squared distances of observation from the mean
- Represented by the symbol  $\sigma$
- Formula to calculate Standard Deviation:

$$\sigma = \sqrt{\sigma^2} = \sqrt{(\sum (X - \mu)^2) / N}$$

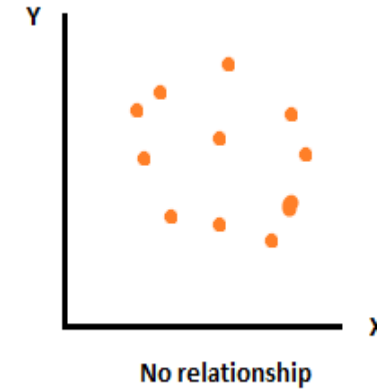
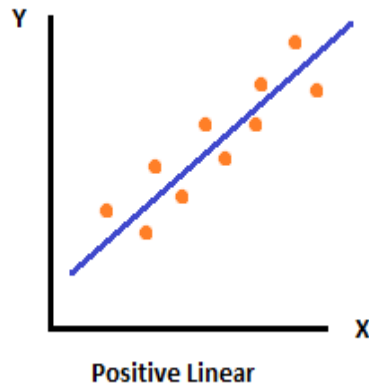
- Units of SD are in the same units as that of the data
- SD enables to determine, with a high accuracy, the values of the frequency distribution in relation to the mean



- About **68%** data lies within  $\pm 1$  SD from the mean
- About **95%** data lies within  $\pm 2$  SD from the mean
- About **99%** data lies within  $\pm 3$  SD from the mean

# III. Measure of Association

Measures the relationship (degree and strength) between two variables that are linearly related



1. Covariance
2. Correlation
3. Coefficient of Variation
4. Rank Correlation



# 1. Covariance (+ -)

- Covariance is the joint variability of two random variables
- Measures the **direction / sign** of relationship only (+ or -)
- How X and Y variables are linearly associated, working in tandem

❑ Eg: **Weight lifter training time** vs **Sprinter training time**

- Weight lifter trains more and lifts more weight (+)
- Trainer trains more and runs in less time (-)

- Covariance measured as ***positive, negative*** or ***zero***

❑ ***Positive:*** indicates direct or increase linear relationship

- X up - Y up
- X down - Y down

❑ ***Negative:*** indicates indirect or decrease in linear relationship

- X up - Y down
- X down - Y up

- Covariance can be any number and not restricted to 0 and 1

- **Formula**

- **Sample  $\text{CoV}_{xy}$**  =  $(\sum xy - n\bar{x}\bar{y}) / n-1$

- **Population  $\text{CoV}_{xy}$**  =  $(\sum xy - n\bar{x}\bar{y}) / n$

**where**

**x** and **y** are the 2 random variables

$\bar{x}$  and  $\bar{y}$  are the means of the 2 random variables

## Exercise


Calculate the Covariance for the following sample dataset

X: 2.1, 2.5, 3.6, 4.0

Y: 8, 10, 12, 14

	X	Y	XY	$\Sigma XY$	$n\bar{x}\bar{y}$	covariance
	2.1	8	16.8			
	2.5	10	25			
	3.6	12	43.2			
	4	14	56			
mean	3.05	11		141	134.2	<b>2.27</b>

Excel calculation



=COVARIANCE.S(B2:B5,C2:C5)						
A	B	C	D	E	F	G
	x	y	covariance			
	2.1	8	2.27			
	2.5	10				
	3.6	12				
	4	14				

**Positive Covariance**

## 2. Correlation (°)

- Measures the **degree** to which one variable is linearly related to the other
- 2 measures are used to describe correlation

### □ Coefficient of Correlation ( $r$ )

- $0 \leq r \leq -1$  : Inverse relationship -> X-increases, Y-decreases
- $0 \leq r \leq 1$  : direct relationship -> X-increases, Y-increases
- Measures the strength and direction
- Formula (Karl Pearson's Coefficient of Correlation / Product moment)  
 $r = \text{covariance of } x \text{ and } y / (\text{SD of } x) * (\text{SD of } y)$

$$r = (\Sigma(xy) - n\bar{x}\bar{y}) / \sqrt{(\Sigma x^2 - n\bar{x}^2)} * \sqrt{(\Sigma y^2 - n\bar{y}^2)}$$

### □ Coefficient of Determination ( $r^2$ )

- $r^2 = r * r$
- Measured in percentage
- Eg:  $r^2 = 0.83$  means 83% of variation in Y (dependent variable) accounted by X (independent variables)
- $r$  does not mean anything,  $r^2$  conveys the actual meaning



### 3. Coefficient of Variation

- Relative Standard Deviation
- Measured in %
- Shows variations with relation to the mean
- Does not have any units
- Smaller CoV is better → represents better quality
- Formula

$$\text{CoV} = \sigma / \mu$$

#### Example:

Last 15 days, trading of 2 stocks are as follows:

#### Stock A

Average price: 135

SD : 15.35

#### Stock B

Average price: 87.5

SD : 1.02

**Which is more risky ?**

$$\text{CoV}_A = 15.35/115 = \mathbf{0.133}$$

$$\text{CoV}_B = 1.02/87.5 = \mathbf{0.011}$$

**Stock A is more risky**

## 4. Rank Correlation

- Measure the degree of similarity between ranks – i.e. to check Correlation among ranks
- Also known as Spearman's Rank Correlation (test)
- Ranks are ordinal data
- Rank ranges from 0 - 1

- **Example**

- ✓ If a student gets a high rank in Subject 1, will he also get a high rank in Subject 2
- ✓ If rank in trials is high, will the rank in the final be also high?

- Formula

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

where

- $r_s$  – Spearman Rank Correlation
- $d$  – difference in ranks
- $n$  – sample size

**Spearman's Rank Correlation (test)**  
**Scenario 1: When the ranks are given**

Student_Id	Trial_Rank	Actual_Rank	Diff (D) (A-T)	D <sup>2</sup>	ΣD <sup>2</sup>	n	n <sup>2</sup> -1	R <sub>s</sub>
100	3	4	1	1	46	6	35	0.781
101	2	6	4	16				
102	1	3	2	4				
103	4	1	-3	9				
104	6	2	-4	16				
105	5	5	0	0				

**0.781 – high correlation**

**Those who performed well in Trials, will also perform well in the Actuals**

## Exercise

**Twelve participants in a contest were ranked by two different judges.  
Is there a correlation between the rankings of the participants ?**

Participant	Rank-Judge 1	Rank-Judge 2
1	2	1
2	4	7
3	6	8
4	8	9
5	10	10
6	12	11
7	1	12
8	3	6
9	5	5
10	7	4
11	9	3
12	11	2



## **Spearman's Rank Correlation (test)**

**Scenario 2: When the data is given and ranks are not given**

<b>Test 1</b>	<b>Test 2</b>	<b>rank1</b>	<b>rank2</b>
56	50	9	10
63	61	7	8
51	69	10	7
90	82	2	4
95	99	1	1
62	72	8	5
73	70	4	6
72	92	5	3
69	58	6	9
74	96	3	2

**Then follow Scenario 1**

# Probability

$$\text{Probability} = (\text{Favourable cases} / \text{Total cases (Sample Space)})$$

## Some basic terminologies in probability theory

### Deterministic model

- Situations where everything relating to situation is known before with certainty
- Not much uncertainty in decision making
- Frequency distribution / descriptive statistics used to arrive at a decision

### Probabilistic model

- Totality of outcome is known, but uncertain which particular outcome will appear
- Lots of uncertainty in decision making
- Probability distributions are used to make decisions

**Event:** one or more possible outcomes of doing something.

**e.g.:** getting a head by tossing a coin, drawing a red face card from a deck of cards, picking a student out of 100 etc.

**Sample space:** A set of all possible outcomes of an experiment

e.g.  $S = \{\text{heads, tails}\}$ ,  $S = \{\text{red king, red queen, red jack}\}$  etc.

**Mutually Exclusive Events:** Only one event can happen from all possible outcomes

e.g. Pass **or** Fail, Rain **or** No Rain etc.

**Odds (in favour / against):** Ratio of an event happening vs not happening

e.g. **Odds of** India winning a match is 3:2, **Odds against** India winning a match is 3:2

**Single Probability:** Only one event can take place

Probability of an event A is expressed as **P(A)**

e.g. Probability of a student getting picked up for a competition out of a class of 50 =  $1/50 = 0.02$

**Addition Rule for Probabilistic Events:** Two events (**A** and **B**), not mutually exclusive, can occur together

$$P(A \text{ or } B) = P(A) + P(B) - P(AB)$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

where

**P(A or B)** = either A or B occur, **P(A)** = only A occurs, **P(B)** = only B occurs, **P(AB)/P(A ∩ B)** = both occur

**Q:** From a pack of cards, what is the probability of getting an Ace and a Heart ?

$$\begin{aligned} P(\text{Ace or Heart}) &= P(\text{Ace}) + P(\text{Heart}) - P(\text{Ace} \cap \text{Heart}) \\ &= 4/52 + 13/52 - 1/52 \\ &= 16/52 \\ &= \mathbf{30.7\%} \end{aligned}$$

**Addition Rule for Mutually Exclusive Events:** Only one event can happen (this **or** that)

$$P(A \text{ or } B) = P(A) + P(B)$$

$$P(A \cup B) = P(A) + P(B)$$

**Q:** From a class of 5, what is the probability of Ram or Sam getting selected, but not both ?

$$\begin{aligned} P(\text{Ram or Sam}) &= P(\text{Ram}) + P(\text{Sam}) \\ &= 1/5 + 1/5 \\ &= 0.4 \\ &= \mathbf{40\%} \end{aligned}$$

**Exclusive and Exhaustive Events:** For an event A, A happens or does not happen

$$P(A) + P(\neq A) = 1$$

$$P(A) = 1 - P(\neq A)$$

No. of children	0	1	2	3	4	5	$\geq 6$
Family Proportion	0.05	0.10	0.30	0.25	0.15	0.10	0.05

**P(4, 5 or 6 children)**

$$\begin{aligned} &P(4) + P(5) + P(6) \\ &= 0.15 + 0.10 + 0.05 \\ &= 0.30 \\ &= \mathbf{30\%} \end{aligned}$$

**P(less than 6 children)**

$$\begin{aligned} &= 1 - P(6) \\ &= 1 - 0.05 \\ &= 0.95 \\ &= \mathbf{95\%} \end{aligned}$$

**Joint probabilities of independent events:** Probability of two or more independent events occurring together or one after the other

$$P(AB) = P(A) \times P(B)$$

where

$P(AB)$  = probability of events A **and** B occurring together or one after the other

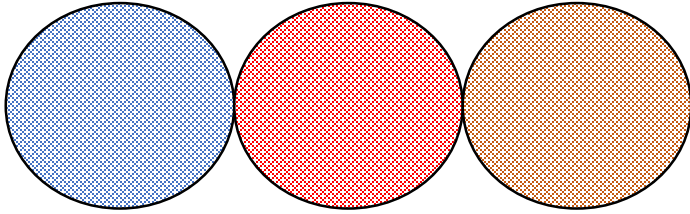
$P(A)$  = marginal probability of event A

$P(B)$  = marginal probability of event B

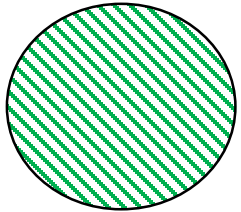
**Q:** What is the probability of getting 3 heads in 3 coin tosses, given the coin is fair ?

$$\begin{aligned} &P(H_1 H_2 H_3) \\ &= P(H_1) \times P(H_2) \times P(H_3) \\ &= \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \\ &= \frac{1}{8} \rightarrow 0.125 \rightarrow \mathbf{12.5\%} \end{aligned}$$

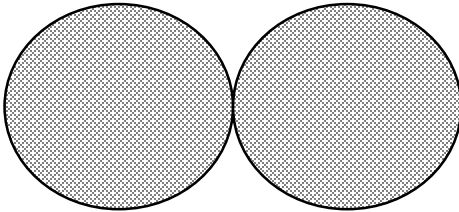
## Conditional Probability Theory for statistically dependent events – example



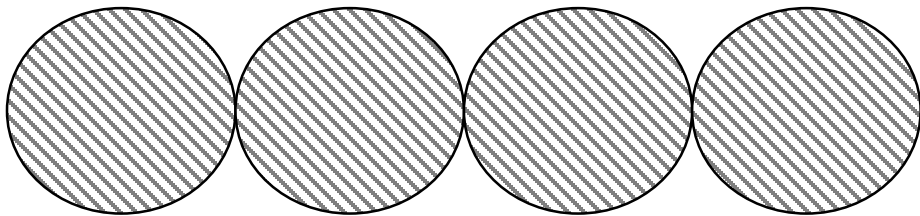
**C and D = 3**



**C and S = 1**



**G and D = 2**



**G and S = 4**

You draw a colour ball, what is the probability is it

a) Dotted ?

b) Striped ?

No. of colour balls = 4

No. of dotted colour balls = 3

$$P(D | C) = 3/4 = 0.75$$

$$P(D | C) = P(DC) / P(C)$$

$$P(D | G) = 2/6$$

$$P(S | G) = 4/6$$

$$P(G | D) = 2/5$$

$$P(C | D) = 3/5$$

$$P(C | S) = 1/5$$

$$P(G | S) = 4/5$$

## Conditional probabilities for Dependent Events / Bayes' Theorem / Posterior probabilities

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

where

**P(A | B)** = Probability of event A occurring, given B has occurred

**P(A)** = Marginal probability of event A

**P(B)** = Marginal probability of event B

-----> **Eq:1**

According to Conditional Probability Theory for statistically dependent events,

$$P(A | B) = P(AB) / P(B)$$

-----> **Eq:2**

Substituting (2) in (1), we get

## Conditional probabilities for Dependent Events / Bayes' Theorem / Posterior probabilities

$$P(B | A) = P(BA) / P(A)$$

where

**P(B | A)** = Probability of event B, given that the event A has occurred

**P(BA)** = Joint probability of events A and B happening together or one after the other

**P(A)** = Marginal probability of event A

## Example

10% of people have a certain disease.

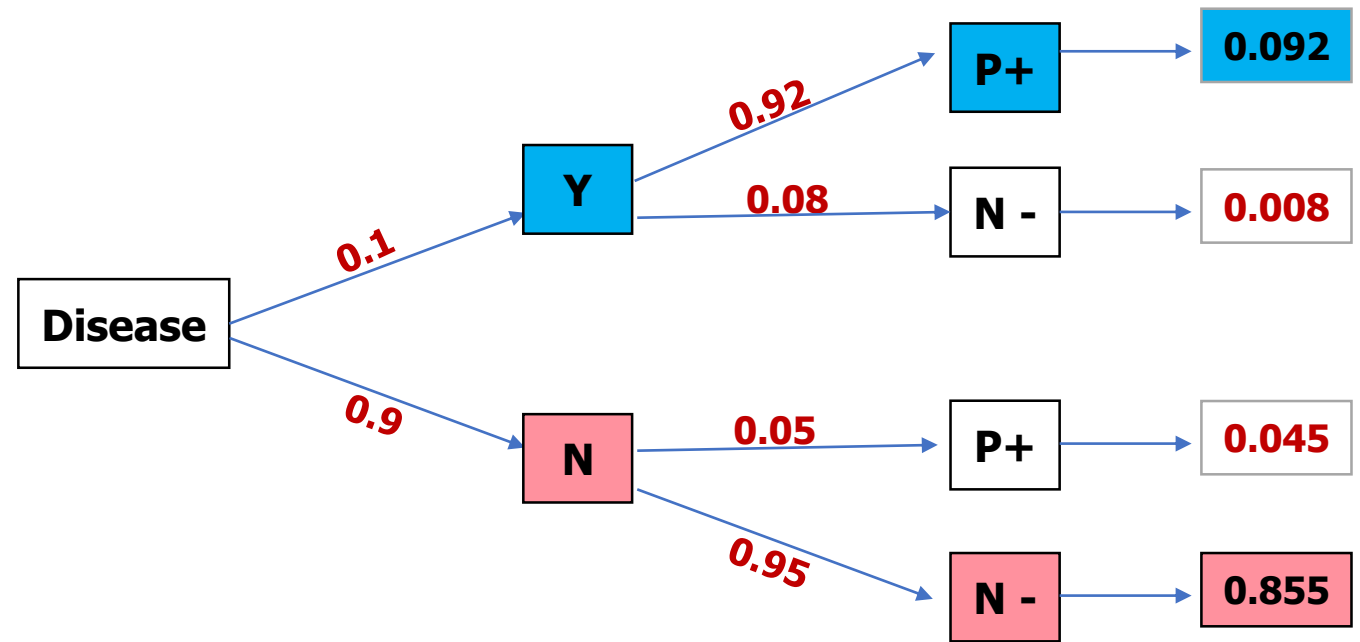
A medical test detects the disease is 92% accurate with a false alarm rate of 5%

Q1) If you test positive, what is the probability you have the disease?

Q2) Your friend tests negative, what is the probability the friend has the disease?



10% have a disease, a test to detect the disease is 92% accurate and a false alarm rate of 5%



**If you test positive, what is the probability you have the disease?**

Test Positive =  $0.045 + 0.092 \rightarrow 0.137$   
Positive and Have disease  $\rightarrow 0.092$

**$P(\text{You have the disease} \mid \text{You test positive})$**   
**=  $0.092/0.137$**   
**=  $0.671$**   
**=  $67.1\%$**

**If your friend tests negative, what is the probability the friend has the disease?**

Test Negative =  $0.008 + 0.855 \rightarrow 0.863$   
Negative and Have disease  $\rightarrow 0.008$

**$P(\text{Friend has the disease} \mid \text{Friend tests negative})$**   
**=  $0.008/0.863$**   
**=  $0.0092$**   
**=  $0.9\%$**

## Example

	Men	Women	<i>Total</i>
Married	10	15	25
Unmarried	30	45	75
<i>Total</i>	40	60	100

**1. One person is selected.  
Probability that the selected person is a married man ?**

$$10/100 = 1 \%$$

**2. One person is selected and is found to be a man.  
Probability that the selected person is a married man ?**

$$10/40 = 25 \%$$

**3. One person is selected and is found to be married.  
Probability that the selected person is a married man ?**

$$10/25 = 40 \%$$

## Marginal probabilities for Dependent Events

$$P(A) = P(AB) + P(AC) + \dots$$

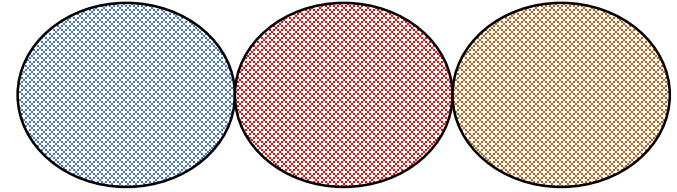
where

**P(A)** = Marginal probability of event A

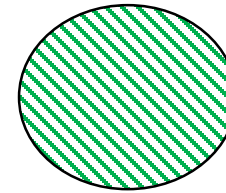
**P(AB), P(AC)** = Joint probabilities that has the event A

$$\begin{aligned} P(G) &= P(GD) + P(GS) \\ &= 2/10 + 4/10 \\ &= 0.2 + 0.4 \\ &= 0.6 \end{aligned}$$

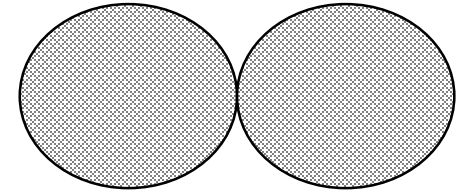
$$\begin{aligned} P(D) &= P(DC) + P(DG) \\ &= 3/10 + 2/10 \\ &= 0.3 + 0.3 \\ &= 0.5 \end{aligned}$$



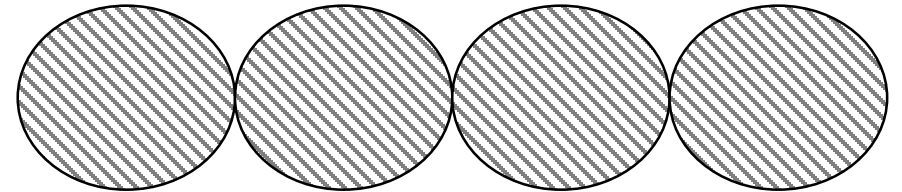
**C and D = 3**



**C and S = 1**



**G and D = 2**



**G and S = 4**

# **Inferential Statistics**

# I. Estimation

- Estimates are characteristics about a sample (mean and proportion) that will represent the population, with a high accuracy
- Calculating the exact mean / proportion is impossible. Estimates give an idea about the expected data
- Helps in decision making
- **Types of estimation**
  - ❑ **Point estimate:** a single number used to estimate the unknown population parameter
  - ❑ **Interval estimate:** a range of values used to estimate the unknown population parameter
- **Estimator**
  - ❑ A Sample statistic used to estimate the population parameter
    - Sample mean ( $\bar{x}$ ) is an estimator to estimate population mean ( $\mu$ )
- **Criteria for a good estimator**
  - ❑ **Unbiasedness** : sample\_mean = population\_mean
  - ❑ **Efficiency** : having smaller standard error (mean/median with smallest SE is a better estimator)
  - ❑ **Consistency** : increase in sample size, value comes close to population value
  - ❑ **Sufficiency** : no other measures are required to represent the population

## Point estimate

**Mean** is the best point estimate for a population because it is unbiased, consistent, efficient

## Sample Variance and Sample Standard Deviation calculation

Packets/Box (x)	Sample mean ( $\bar{x}$ )	std dev ( $x_i - \bar{x}$ )	SD sq ( $(x_i - \bar{x})^2$ )	Sample size (n)	sample variance	sample std dev
109	105	4	16	20	42.58	6.53
97		-8	64			
104		-1	1			
110		5	25			
110		5	25			
102		-3	9			
112		7	49			
111		6	36			
98		-7	49			
106		1	1			
94		-11	121			
109		4	16			
103		-2	4			
110		5	25			
95		-10	100			
106		1	1			
104		-1	1			
113		8	64			
96		-9	81			
94		-11	121			
			809			

If **x** represents the number of packets in each box, we can estimate the **Population Mean** of packets/box by taking a sample of packets and calculating its mean.

Likewise for **Variance** and **Standard Deviation**

## Interval estimate

A range of values within which a population parameter is likely to lie

- From the previous example of packets/box, what is the uncertainty associated with the estimate ?
- This uncertainty is called the ***standard error of the mean***
- Formula to calculate the standard error

$$\sigma_{\bar{x}} = \sigma / \sqrt{n}$$

where

$\sigma_{\bar{x}}$  = Standard error of mean of the population

$\sigma$  = Standard deviation of the population

$n$  = Sample size

- ***Standard error of proportion***
- Formula to calculate the standard error of proportion

$$\sigma_p = \sqrt{(\hat{p}\hat{q}/n)}$$

where

$\sigma_p$  = Standard error of population proportion

$\hat{p}$  = sample proportion in favour

$\hat{q}$  = sample proportion not in favour

$n$  = Sample size

To calculate the Interval estimate, we apply the formula

**Estimate = Population Mean  $\pm$  n (standard error)**

Standard Error
$6.53 / \sqrt{20} = 1.46$

Mean	Standard Error	Est Min Mean	Est Max Mean
105	1 SE (1.46)	$105 - 1.46 = 103.54$	$105 + 1.46 = 106.46$
	2 SE (2.92)	$105 - 2.92 = 102.08$	$105 + 2.92 = 107.92$
	3 SE (4.38)	$105 - 4.38 = 100.62$	$105 + 4.38 = 109.38$

# Hypothesis Testing



## II. Hypothesis Testing

- It is an assumption made about a **population** parameter
- Collect sample data, produce sample statistics and decide how likely our hypothesised parameter is correct
- Calculate the difference between the hypothesized and actual values (goal of HT)
- Judge if the difference is significant
- Hypothesis is either rejected or accepted
  - Objectively and not by intuition

### **Example of hypothesis**

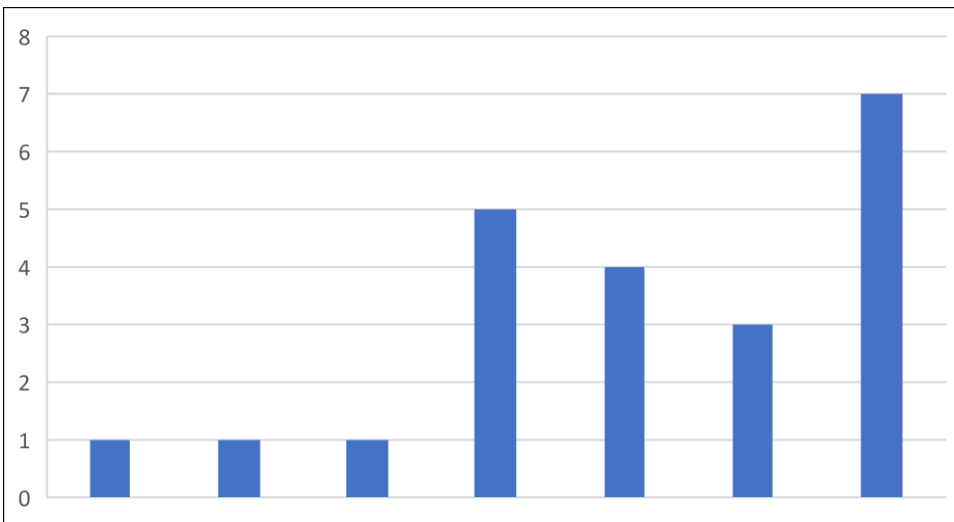
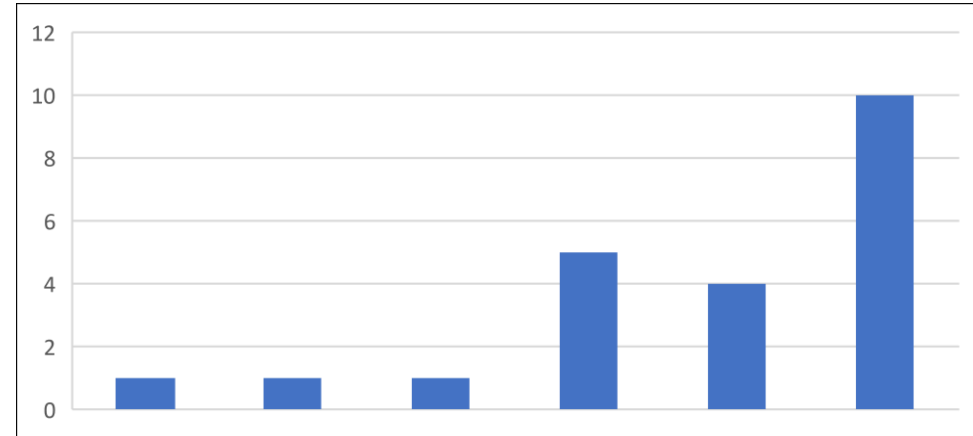
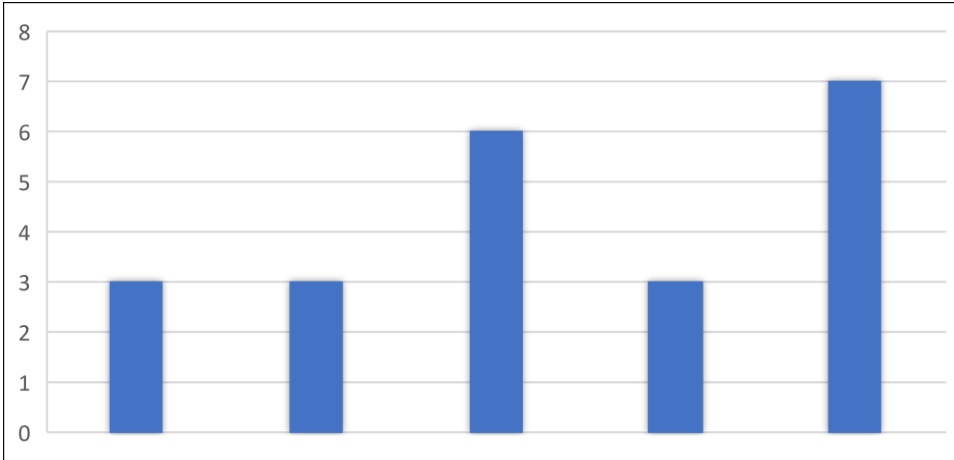
1. Percentage of every 1000 students studying for higher degrees after any graduation is 35%
2. Average salary of a 10-year old experienced person is 10 L/a

### Testing the hypothesis

# **Data Distribution**

# Data Distribution

- A distribution is a listing or a graph of all possible values of a random variable or a population
- Also indicates how frequently they occur



# Kurtosis

- Measure of peak of a frequency distribution
- More peak -> Larger kurtosis (and vice versa)
- Range of Kurtosis of a normal distribution is  $[-3, +3]$

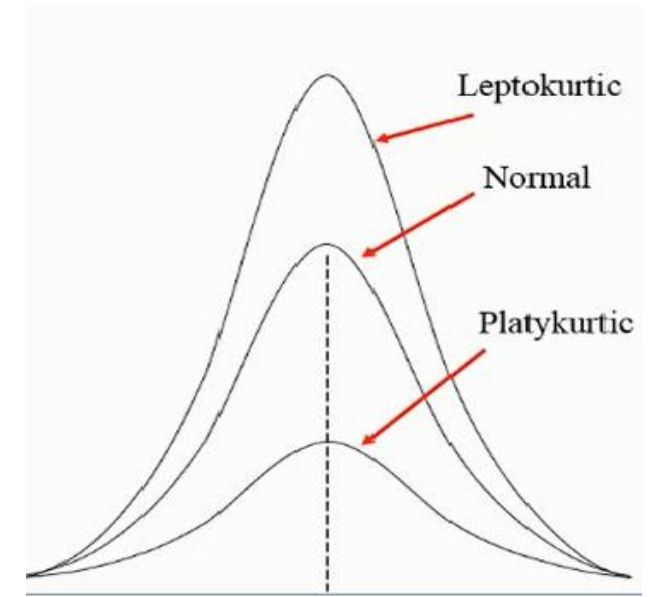
- **Formula (Absolute Kurtosis)**

$$\beta_2 = \Sigma[(x_i - \mu) / \sigma]^4 / N$$

$\beta_2 > 3$ , curve more peaked than normal – **Leptokurtic**

$\beta_2 < 3$ , curve less peaked than normal – **Platykurtic**

$\beta_2 = 3$ , curve with normal peak – **Mesokurtic**



- **Relative Kurtosis**

- Can be negative
- $RK = \text{Absolute Kurtosis} - 3$
- Generally, work with relative kurtosis

- $RK \rightarrow +ve$ , Leptokurtic

- $RK \rightarrow -ve$ , Platykurtic

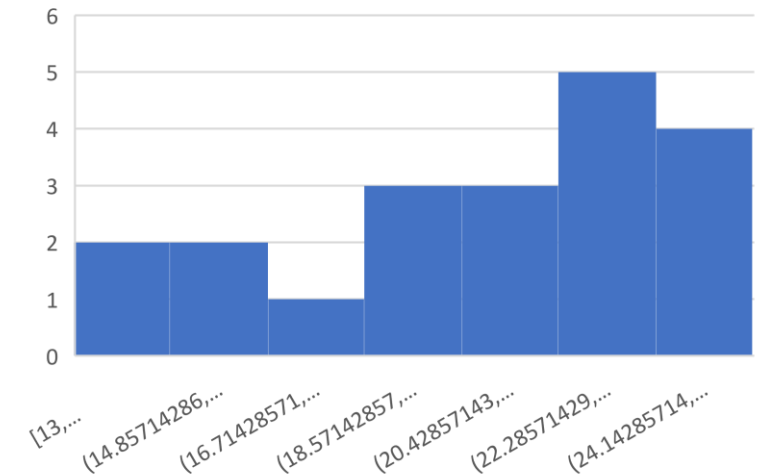
## Example

x
24
22
14
21
26
19
20
13
24
26
24
26
23
20
22
16
24
26
18
15

N	$\mu$	$\sigma$
20	21.15	4.065

$(x-\mu)/\sigma$	$[(x-\mu)/\sigma]^4$
0.701	0.242
0.209	0.002
-1.759	9.568
-0.037	0.000
1.193	2.026
-0.529	0.078
-0.283	0.006
-2.005	16.152
0.701	0.242
1.193	2.026
0.701	0.242
1.193	2.026
0.455	0.043
-0.283	0.006
0.209	0.002
-1.267	2.575
0.701	0.242
1.193	2.026
-0.775	0.360
-1.513	5.237

Sum	$B_2$	Rel Kurt
43.098	2.155	-0.845

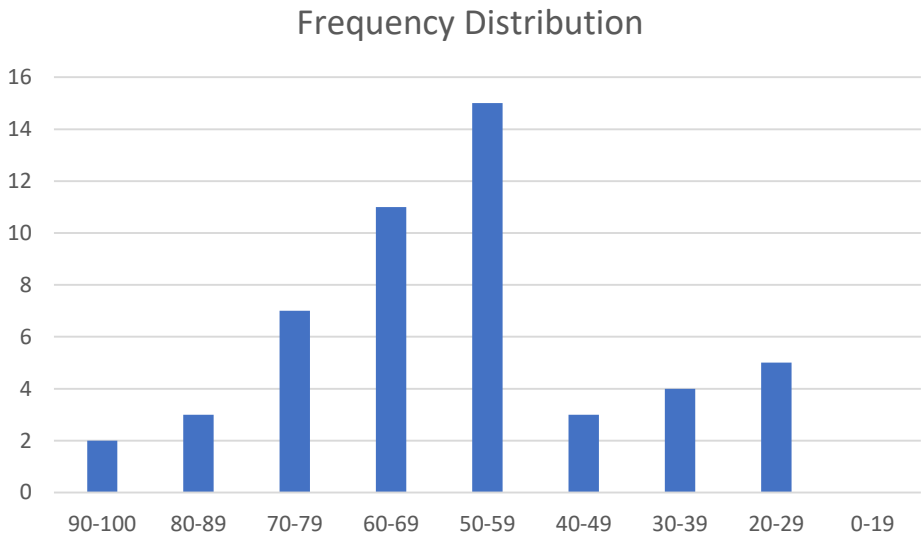


- **Left skewed**
- **Negative Kurtosis**
- **Flatter distribution than normal**
- **Mean < Median < Mode**

# Frequency distribution

Listing of the observed frequencies of all outcomes of an experiment that actually occurred when the experiment was done

Marks	Freq.
90-100	2
80-89	3
70-79	7
60-69	11
50-59	15
40-49	3
30-39	4
20-29	5
0-19	0



Weight	Freq.
10-15	8
16-20	10
21-25	20
26-30	31
31-35	40
36-40	38
41-45	60
46-50	30
> 50	5

## Probability distribution

Listing of all the probabilities of all the possible outcomes that could result if the experiment were done

### Discrete Probability distribution

**Discrete** amount of **outcomes**



Uniform probability distribution

Binomial distribution

Poisson distribution

### Continuous Probability distribution

**Outcome** can be **any value** within a given range

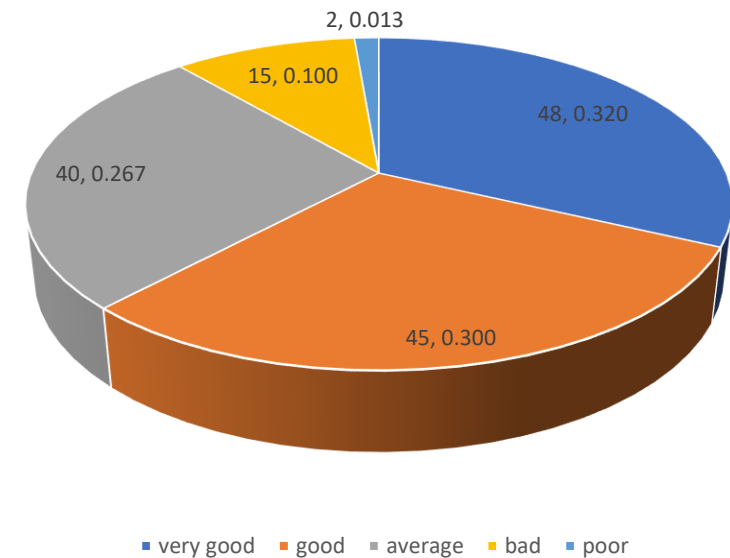
# Discrete Probability Distribution

- Discrete amount of outcomes whose probabilities add up to 1
- Can take limited values (whole numbers)
- **Examples**
  - Probability of someone born in a particular month (12 possible values)
  - Coin flip (2 possible values H / T )
  - Rolling a die to get a number 1-6 (6 possible values)
  - Survey results : Very Good (48), Good (45), Average (40), Bad (15), Poor (2)

## Conditions for Discrete probabilities

- $0 \leq P(x) \leq 1$
- $\sum P(x) = 1$

value	survey	prob
very good	48	0.320
good	45	0.300
average	40	0.267
bad	15	0.100
poor	2	0.013
	150	1.000





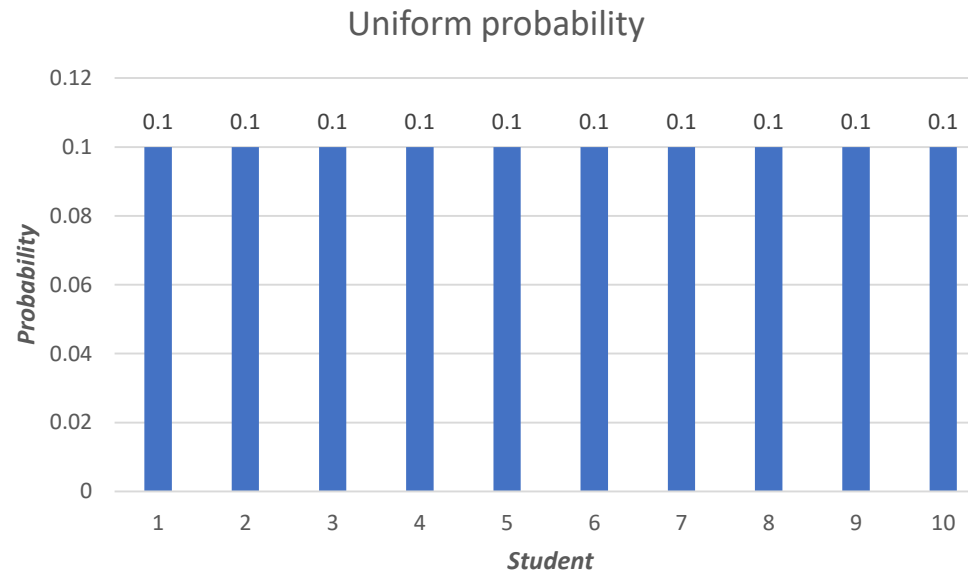
# Uniform Probability Distribution

- Probability of an event happening is equal for every observation
- Each outcome has the same probability
- e.g:
  - Probability of a head or tail in a flip of a coin is 50% each
  - Probability of a number between 1 and 6 in a roll of die is  $1/6 = 0.166$  each

## Exercise

Given a class of 10 students, what is the probability of picking up 1 student at random ?

student	probability
1	0.1
2	0.1
3	0.1
4	0.1
5	0.1
6	0.1
7	0.1
8	0.1
9	0.1
10	0.1



# Binomial Distribution

1. Sequence of 'n' trials / iterations which is fixed
2. Only 2 possible (exclusive) outcomes. Success / Failure
3. Probability value (**p**) does not change from trial to trial. Failure is **1-p**, which is also fixed
4. Trials are independent. Outcome of a trial does not influence future outcomes

## Formula for Binomial Distribution probabilities

$$\frac{n!}{r! (n-r)!} p^r (1-p)^{n-r}$$

where

**n** = number of trials

**r** = number of successes

**p** = probability of success in a trial

**1-p** = probability of failure

## Example

What is the probability of finding 1 defective product out of 5 random samples, given the defect rate of the product is 25%

n (number of trials) = 5  
r (number of successes) = 1  
p (probability of success) = 0.25

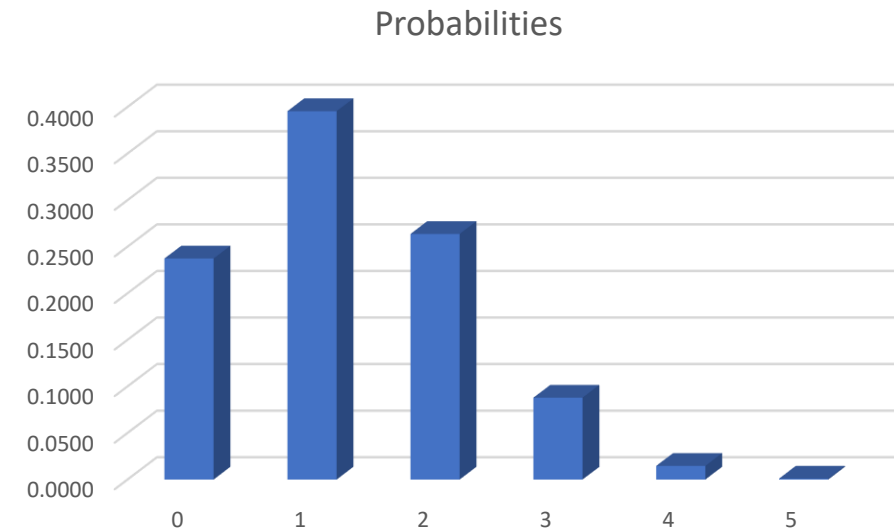
$$\begin{aligned} P(\text{finding one defective product}) &= [5!/(1!)*(5-1)!] (0.25)^1 (1-0.25)^4 \\ &= 5 * 0.25 * 0.316 \\ &= \mathbf{0.39} \end{aligned}$$

## Exercise

Similarly, the probability of finding 0, 2, 3, 4 and 5 defective products out of the 5 random samples can be found out using the formula

n	r	p	n!	r!	(n-r)!	p <sup>r</sup>	1-p <sup>(n-r)</sup>	P
5	0	0.25	120	1	120	1	0.237304688	0.2373
	1			1	24	0.25	0.31640625	0.3955
	2			2	6	0.0625	0.421875	0.2637
	3			6	2	0.015625	0.5625	0.0879
	4			24	1	0.00390625	0.75	0.0146
	5			120	1	0.000976563	1	0.0010

[Excel calculation](#)



# Poisson Distribution

1. Focus on the number of discrete events over a *specified interval of time*
2. Works on a concept of Expected Value  $\rightarrow \mu$  (mean value)
3. Represented by  $\lambda$  (lambda) = (# occurrences / interval)
4. Each event is independent of the other events
5. Expected Value is assumed to be constant in every trial
6. Change of interval should also change the Expected Value

## Formula for Poisson Distribution probabilities

$$\frac{\lambda^x e^{-\lambda}}{x!}$$

### where

$x$  = number of occurrences

$\lambda$  = average

$e$  = exponential value (2.17)

## Example

1. In a given interval of 10 minutes, the Expected value ( $\mu$ ) of vehicles passing through a bridge is 10.
2. Number of vehicles can be 0 or  $\infty$  (both are unlikely events, but theoretically possible)

## Example exercise

What is the probability that exactly 8 cars will cross the bridge in the 10 minute interval ?

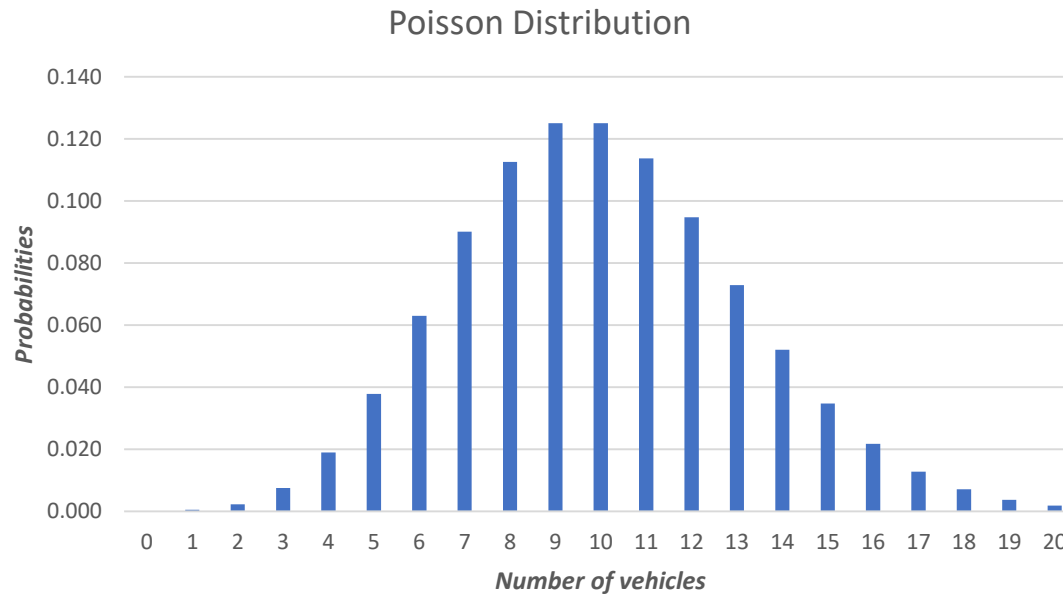
$$x = 8, \lambda = 10, e = 2.17$$

$$\begin{aligned} P(8 \text{ cars}) &= (10^8 * 2.17^{-10}) / 8! \\ &= 0.11 \end{aligned}$$

# Exercise

What is the probability that exactly 'n' cars will cross the bridge in a 10 minute interval ? Given n = 1 to 20

x	poisson
0	0.000
1	0.000
2	0.002
3	0.008
4	0.019
5	0.038
6	0.063
7	0.090
8	0.113
9	0.125
10	0.125
11	0.114
12	0.095
13	0.073
14	0.052
15	0.035
16	0.022
17	0.013
18	0.007
19	0.004
20	0.002



[Excel calculation](#)

## Exercise

What is the probability that more than 10 cars will cross the bridge ?

$$\begin{aligned} P(> 10) \\ &= 1 - 0.58303 \\ &= 0.41697 \end{aligned}$$

=POISSON.DIST(B12,A2,TRUE)				
D	E	F	G	H
	probability of 10			
	0.58303975			

# Geometric Distribution

- Number of trials needed before getting the first success by repeating the independent trials
- When success occurs at trial  $x$ ,  $x-1$  trials must be failures
- Minimum trial is 1
- Maximum trial is  $\infty$

## Formula for Geometric Distribution probabilities

$$P(X=x) = (1-p)^{x-1} * p$$

where

$x$  = number of successes

$p$  = probability of success in a trial

$1-p$  = probability of failure

# Exercise

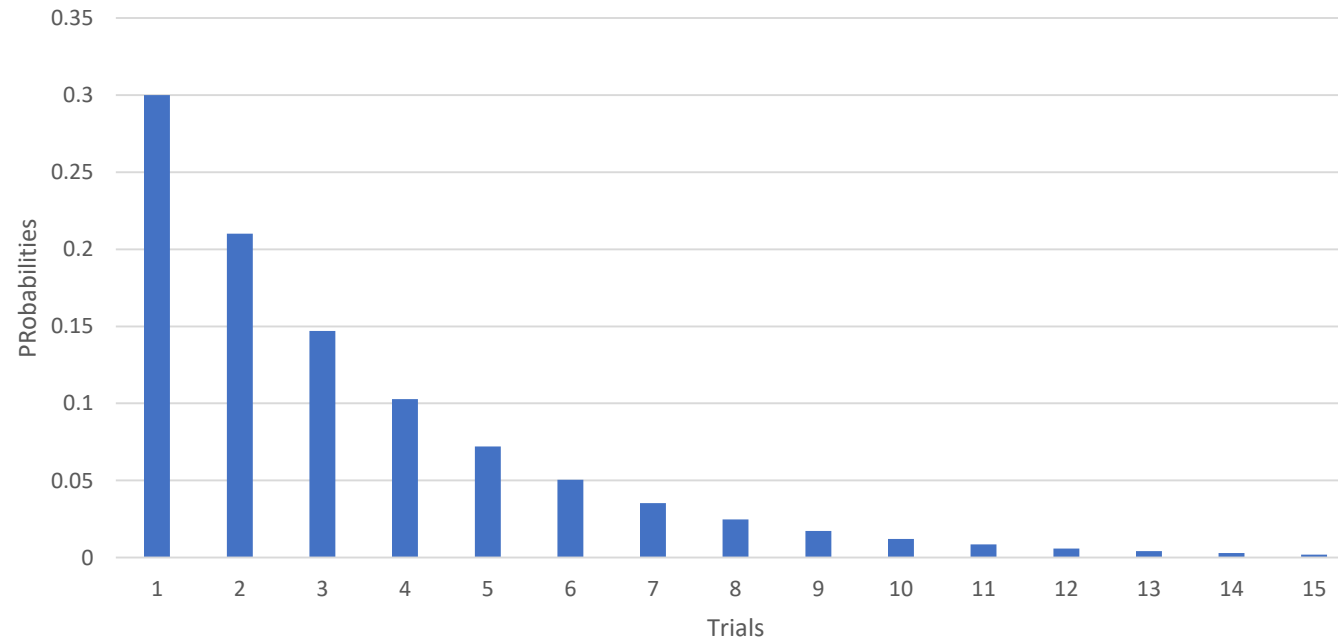
## Exercise

To find the probability that the 7<sup>th</sup> person sampled is the first one to have a certification

Given that 30% of population are certified

p	1-p	x	$P(X=x) = (1-p)^{x-1} * p$
0.3	0.7	1	0.3
		2	0.21
		3	0.147
		4	0.1029
		5	0.07203
		6	0.050421
		7	0.0352947
		8	0.02470629
		9	0.017294403
		10	0.012106082
		11	0.008474257
		12	0.00593198
		13	0.004152386
		14	0.00290667
		15	0.002034669

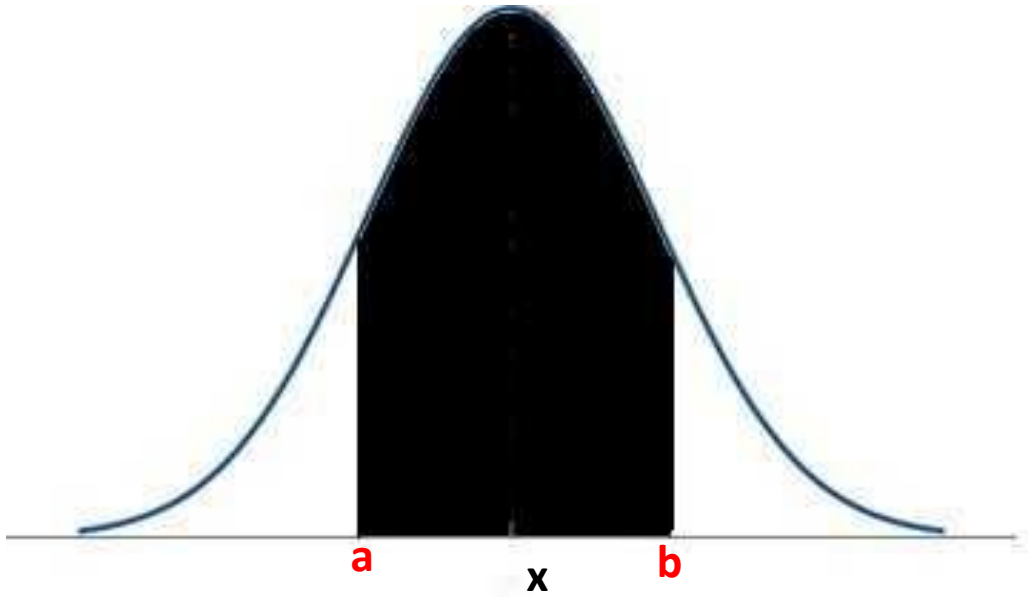
Geometric Distribution



# Continuous Probability Distribution

- Variable can take any value within a given range eg:  $\{3-4\}$ ,  $\{0-\infty\}$
- To model a CPD, we need a different method – Probability Density Function **f(x)**  $\int_3^4 f(x)dx$ 
  - $f(x) = 1 / (n \text{ equal outcomes})$
  - For a coin flip,  $f(x) = 1/2$
  - For a die roll,  $f(x) = 1/6$
- As the number of outcomes increase, probability gradually decreases and tends towards 0
- Probabilities should be  $\geq 0$  and  $\leq 1$
- Probabilities are the areas under the curve
- Some measurements are very **precise** that can have an infinite number of outcomes. Eg:
  - ☐ Temperature
  - ☐ Mass
  - ☐ Distance
- Therefore, probability of a specific outcome (continuous number) is 0
- For continuous data, we can find probability only for a specific interval





For a given random value  $x$ , the probability that it will fall between the points  $a$  and  $b$  is the area under the curve (highlighted In black)

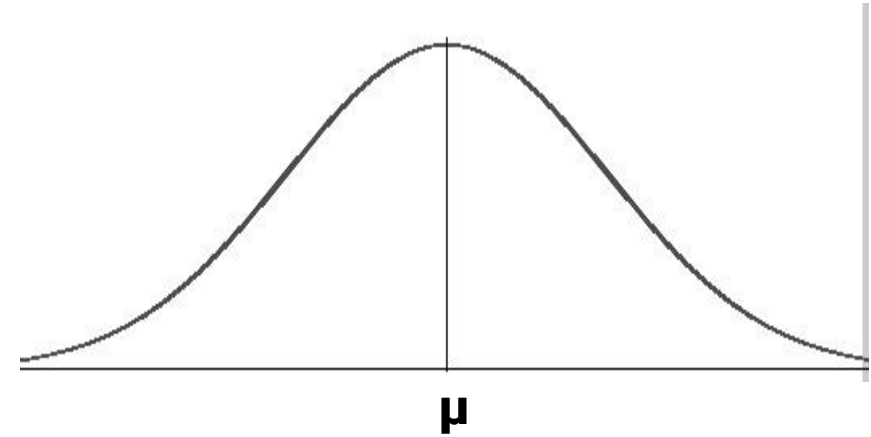
$$P(a < x < b)$$

### Example of a Continuous Distribution

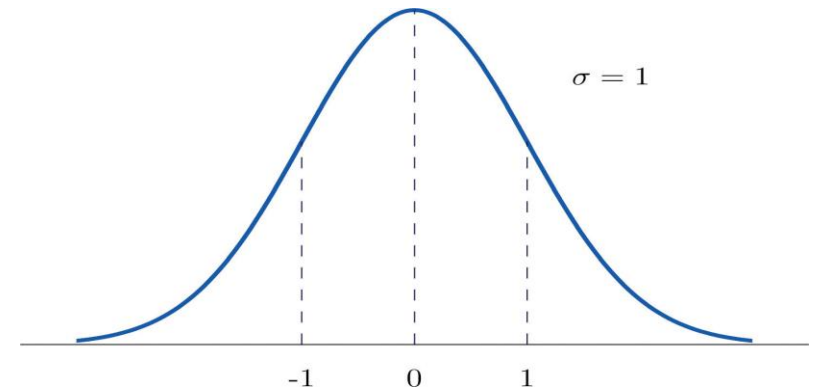
- Normal Distribution Curve (bell shape curve)
- Uniform Distribution
- Exponential Distribution

# Normal Distribution

- Distribution that occurs in situations where data is continuous in nature
- Distribution resembles a bell. Hence, bell-shaped curve
- Represents what percentage of data falls within a given Standard Deviation (SD) **1SD=68%, 2SD=95%, 3SD=99%**
- Mean, Median and Mode are all the same
- Curve is symmetric at the centre (around  $\mu$ )
  - Half values left of the mean and half values to the right of the mean
- Total area under the curve is 1



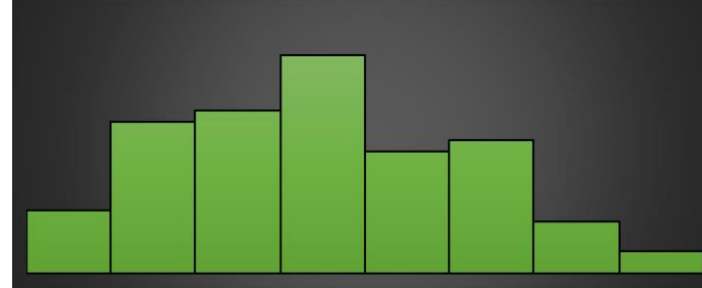
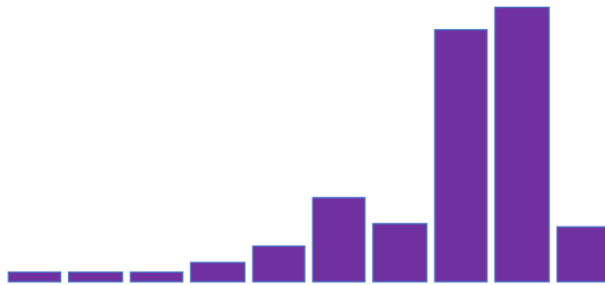
- **Standard Normal Distribution**
  - **Mean = 0**
  - **Standard Deviation = 1**



# Removing Skewness from data

Skewness can be removed either by taking Logs, Square roots or Inverse of the data

measure	range	With actual data	After transformation
Skew	<b>-1 to 1</b>	-1.74	0.06



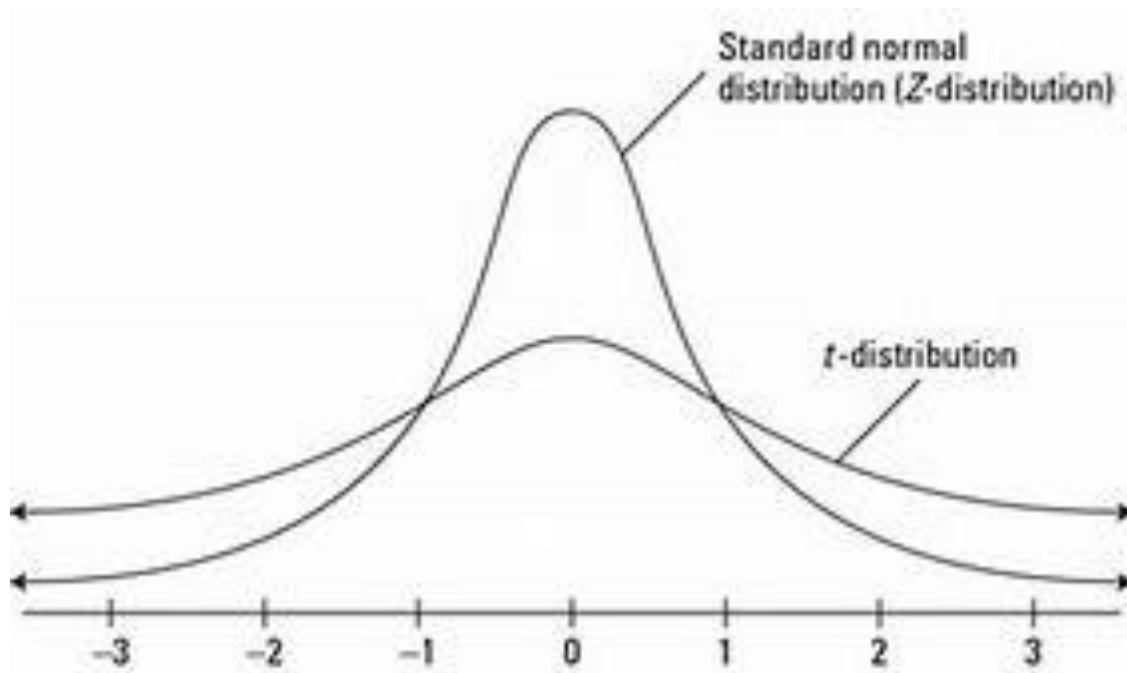
In Excel, histogram can be plotted by using the **Data->Data Analytics->Histogram** option

1. Select the data range
2. Select the Bins

[Actual calculations here](#)

# t-Distribution

- Also called the student's t-distribution
- t-distribution is observed when
  - ✓ Sample size is  $\leq 30$
  - ✓ Population standard deviation is not known
  - ✓ Assumption: Population is normally or approximately normally distributed



## t-distribution vs normal distribution

- Lower at the mean and higher at the tails
- More area in the tails
- Interval widths are wider

# Central Limit Theorem

- The sample mean is approximately normally distributed as the sample size increases, **irrespective of the underlying distribution of data** from which we are sampling
- i.e. if the Population is normally distributed, then  $\bar{x}$  is also normal

# **Sampling Techniques**

## **Sampling**

- A process to determine the characteristics of an entire population
- Sampling can be done for items and people; depending upon the study
- A sample is a portion chosen from a population
- Sampling is done for 2 main reasons
  - Time
  - Cost

## **Types of Sampling**

- Random / probability sampling
- Non-random / judgemental sampling

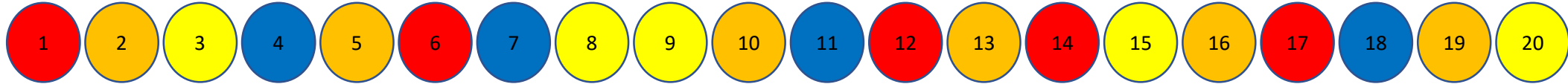
# Sampling techniques

Simple Random

Systematic

Stratified

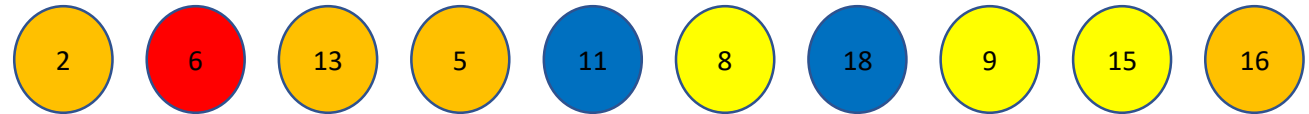
Cluster



**Population**

## Simple Random sampling (SRS)

Number picked at random by some means



**10**

## Systematic sampling

Starting from the first, pick every 3<sup>rd</sup> element

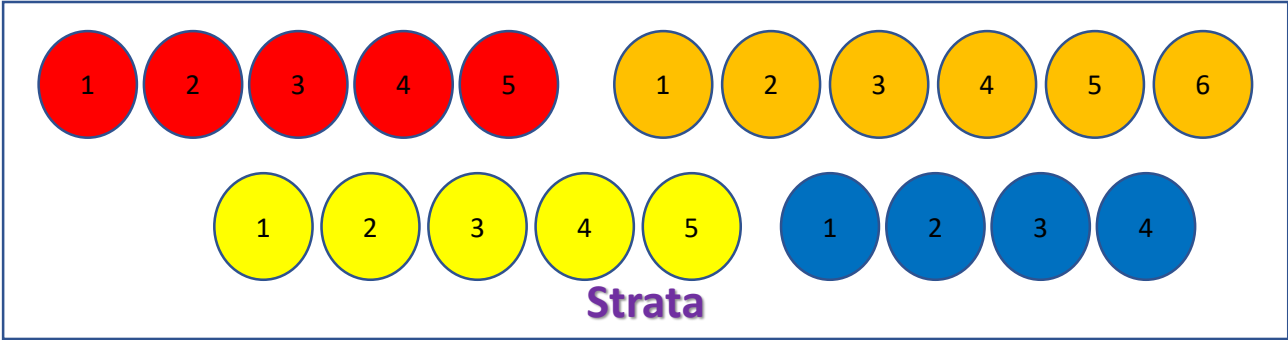


**7**



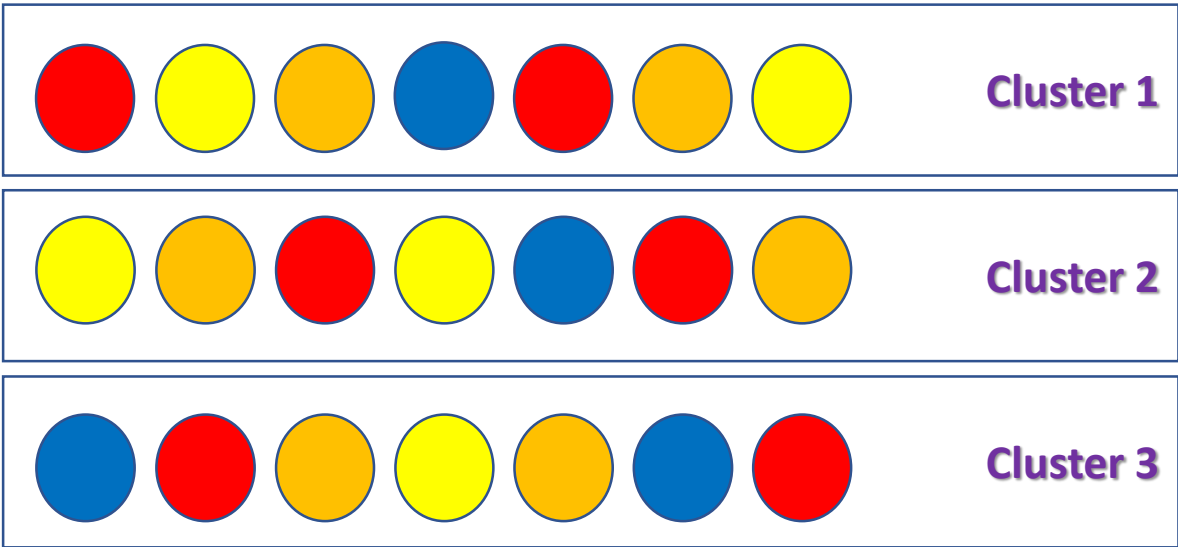
**Strata sampling**

- Group elements by similar characteristics

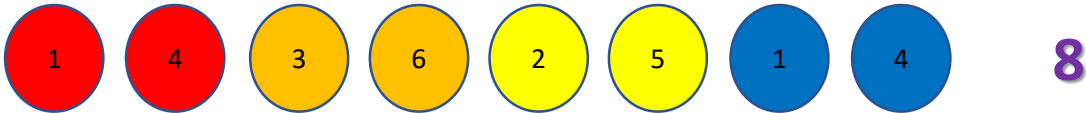


**Cluster sampling**

- Group elements by some characteristics



- Use SRS to get 'n' elements from each strata (e.g. n=2)



- Use SRS to get any 1 cluster from the 3 clusters

