STATISTICS

It is the science of collecting, organizing & analyzing data.

- Data: Facts or pieces of information.
  Eg: Height of student, no. of sales in term of revenue

→ Types of Statistics

① Descriptive Statistics	② Inferential Statistics
- It consists of organizing, summarizing & visualizing data.	- It consists of using data you have measured to form conclusion.
1) Measure of Central Tendancy [Mean, Median, Mode]	1) z-test 2) t-test 3) chi-square test 4) Anova test
2) Measure of Dispersion [Variance, Z-score, Standard deviation]	(hypothesis testing, P-values, Significance value)
3) Different types of Distribution of data [Histogram, pdf, pmf, Gaussian Dist., Binomial, Bernoulli, Exponential, Poisson, Log Normal Dist]	

EDA & Feature Engineering

EDA: Exploratory data analysis

DESCRIPTIVE STATISTICS

Date
Page

① Population & Sample Data

Population (N) : Group or superset of data that you are interested in studying.

Sample (n) : It is a subset of population data.

→ Types of data

Data

Quantitative | Qualitative

Discrete	Continuous	Nominal	Ordinal
⇓	⇓	⇓	⇓
+ve whole number	any value	No Ranks	Ranks
Eg: no. of children in family - number of people working in a company	Eg: Average rainfall - length of river	Eg: Gender (M,F - Blood grp - colors - location	Eg: Customer feedback

* Scale of Measurement
1) Nominal scale data
2) Ordinal scale data
3) Interval scale data
4) Ratio Scale Data.

1) Nominal Scale Data
- Qualitative / Categorical data
- Order or rank does not Matter
- Eg: Gender, color, labels

Eg: Colors
red → 5 → 50%
yellow → 3 → 30%
orange → 2 → 20%
10

2) Ordinal scale data
- rank is important
- order Matters
- Difference cannot be measured

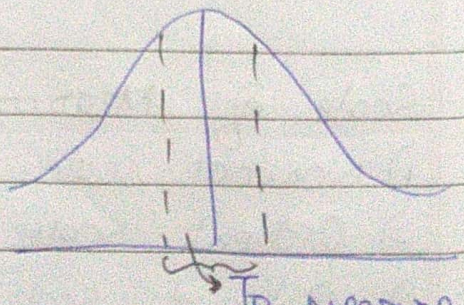
3) Interval scale data
- order Matters
- difference can be Measured
- ratio cannot be Measured
- No 0 (zero) starting point

Eg: Temperature variable
30f
60f → 60:30 = 2:1
90f
0f ✗✗
120

4) Ratio scale data.
- order Matter
- difference are Measurable (ratio)
- contain "0" starting point

Eg: Students Marks in class
3:1
0, 90, 60, 30, 75, 40, 50
3:2

★ Measure of central Tendency
It is very imp in EDA [Exploratory Data Analysis]
1) Mean or average
2) Median
3) Mode



To Measure this central region we need central tendency

① Mean : Avg of all the elements present in distribution

Population (N) Sample (n)

random variable : $X = \{1, 1, 2, 2, 3, 3, 4, 5, 5, 6\}$

Population Mean $(\mu) = \sum_{i=1}^{N} \frac{x_i}{N} = \frac{[1+1+2+2+3+3+4+5+5+6]}{10}$

$= \frac{32}{10} = 3.2$

② Median : whenever there is outlier we use median.

$X = \{4, 5, 2, 3, 2, 2\}$

Steps :

1) Sort the random variable X

$X = \{1, 2, 2, 3, 4, 5\}$

2) No. of elements → if n = even ⇒ Middle 2 element ⇒ Avg
→ if n = odd ⇒ central element

$= \frac{2+3}{2} = 2.5$

③ Mode : frequency of maximum occuring element

$X = \{2, 1, 1, 1, 4, 5, 7, 8, 9, 9, 10\}$

Mode = 1

- In most of scenarios mode is used in categorical value

★ Measure of Dispersion

1) Variance → The more the spread the more variance

Population variance	Sample variance
$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$	$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$
x_i → Data points	x_i → Data points
μ → population Mean	$\bar{x}$ → sample Mean
N → population size	n → sample size

- The sample variable is divide by $(n-1)$ so that we can create an unbiased estimator of population variance.

2) Standard Deviation

Population std	Sample Std
$\sigma = \sqrt{\text{variance}}$	$S = \sqrt{\text{sample variance}}$

- How many std x_i is away from mean

Eg: $\{1, 2, 3, 4, 5\}$, $\mu = 3$

$$\text{Population variance} = \frac{[(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2]}{5}$$

$$= \frac{4+1+0+1+4}{5} = \frac{10}{5} = 2$$

$\sigma = \sqrt{2} = 1.41$

0.18 1.59 3 4.41 5.82 7.23

★ Random variables

It is a process of mapping the output of a random process or experiment to a number

Eg: Tossing a coin

$X = \begin{cases} 0 & \text{head} \\ 1 & \text{Tail} \end{cases}$

Rolling a dice

$Y = \{2, 1, 4, 2, 3, 5\}$

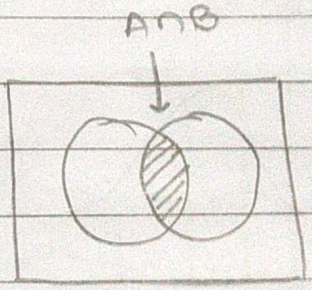
★ Sets

$A = \{1, 2, 3, 4, 5, 6, 7, 8\}$

$B = \{3, 4, 5, 6, 7\}$

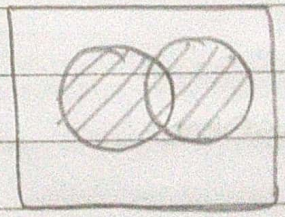
1) Intersection

$A \cap B = \{3, 4, 5, 6, 7\}$



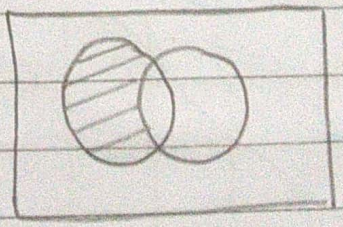
2) Union

$A \cup B = \{1, 2, 3, 4, 5, 6, 7, 8\}$



3) Difference

$A - B = \{1, 2, 8\}$



4) Subset

$A \rightarrow B$: False

$B \rightarrow A$: True

5) Superset

A → B : True → A has all the elements w.r.t B

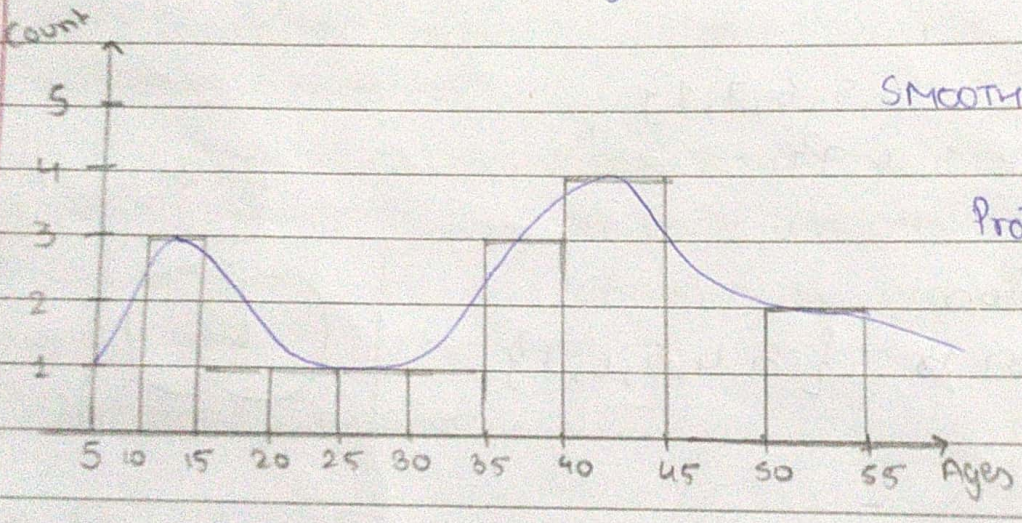
B → A : False

★ Histograms And Skewness

[Frequency] ⇒ Distribution

Ages = {10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51}

Bins : How many groups we are going to make & what will be the group size.



SMOOTHEN THE HISTOGRAM

⇓

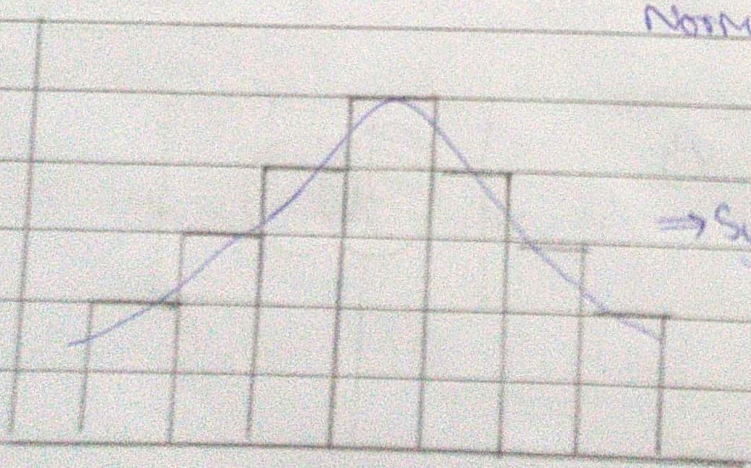
Probability distribution fn

⇓

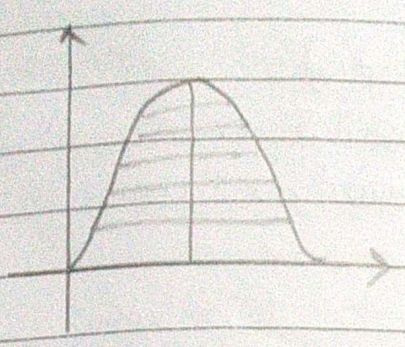
Probability density fn

★ Skewness

Normal / Gaussian Distribution



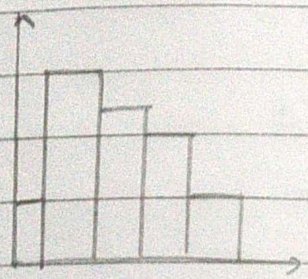
⇒ Symmetrical Distribution



→ No skewness

→ Mean = Median = Mode

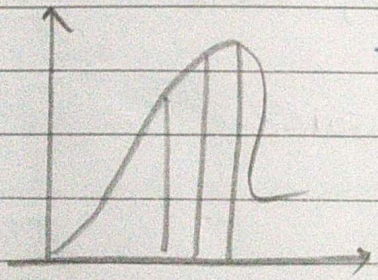
2) Right skewed



→ Positive skewed

→ Mean ≥ Median ≥ Mode

3) Left skewed



→ Negative skewed

→ Mean ≤ Median ≤ Mode

★ Sampling Techniques

1) Random sampling : Simple random sampling gives each member of the population an equal chance of being chosen for the sample.

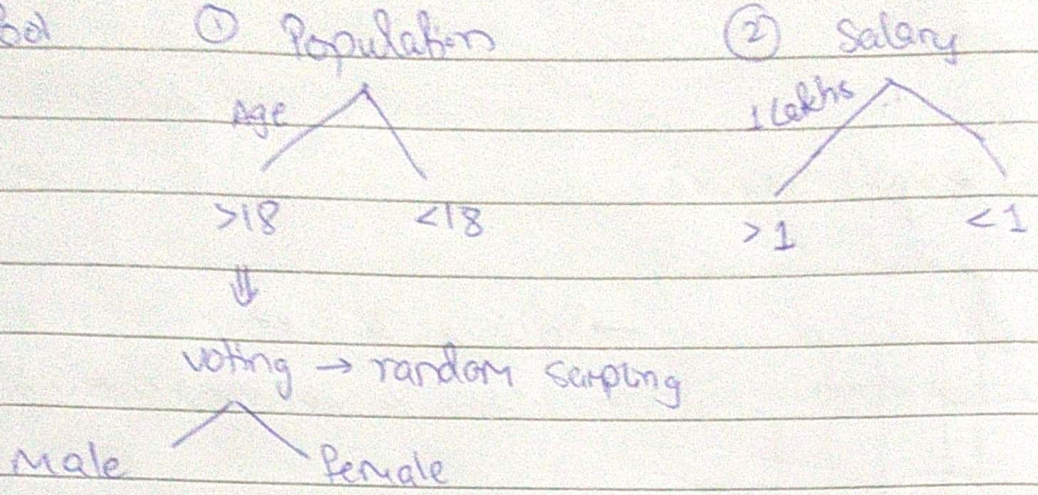
Eg: Vaccination test → 100 → randomly select couple
Accidental Test → 1000 → A → 2 vehicle

Avg IQ of school → select 10 people

2) Stratified sampling [Stratified → layers]

It involves dividing the population into sub-population that may differ in important ways

Eg: Exit Poll



3) Systematic Sampling

It is a statistical method involving the selection of elements from an ordered sampling frame.

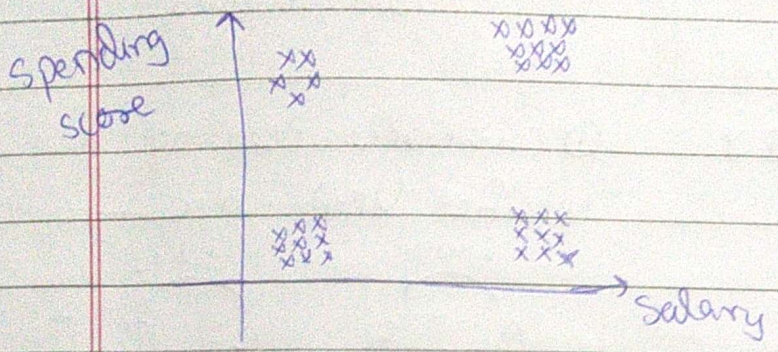
Eg: Mall → fill up a form for a cause.

4) Convenience Sampling

5) Purposive Sampling

Judgemental Sampling is a Method where researchers decides which Members of the target population will be sampled.

6) Cluster Sampling



* Covariance & Correlation

x	y
2	3
4	5
6	7

$x \uparrow \quad y \uparrow$

$x \downarrow \quad y \downarrow$

Relation between x & y

↓

Covariance & Correlation

Covariance

$$Cov(x,y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$Var(x) = \sum_{i=1}^{N} \frac{(x_i - \bar{x})^2}{n-1}$$

$x_i \rightarrow$ Data point of x

$\bar{x} \rightarrow$ sample mean of x

$y_i \rightarrow$ Data points of y

$\bar{y} \rightarrow$ sample mean of y

$$= \sum_{i=1}^{n} \frac{(x_i - \bar{x})(x_i - \bar{x})}{n-1}$$

$Var(x) = Cov(x,x) \Rightarrow$ spread

$Cov(x,y)$ | $x\uparrow\ y\uparrow$ / $x\downarrow\ y\downarrow$ | +ve covariance

$x\uparrow\ y\downarrow$ / $x\downarrow\ y\uparrow$	-ve covariance

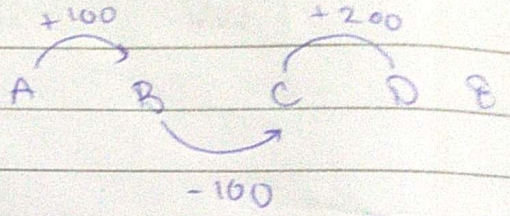
Advantages

① Relationship between x & y +ve or -ve values

A ↔ B +infinity, -infinity

Disadvantages

① Covariance does not have a specific limit value



② Pearson Correlation Coefficient

In order to prevent the disadvantage we use it.

$$\rho_{x,y} = \frac{Cov(x,y)}{\sigma_x \cdot \sigma_y} \rightarrow -1 \text{ to } +1$$

- The more the value towards +1, the more +ve correlated x, y is.
- The more value towards -1, the more -ve correlated x, y is

x	y
8	6
9	5
10	4
12	2
11	1

$\bar{x} = 10$, $\bar{y} = 3.6$ $Cov(x,y) = -3$

$\sigma = \sqrt{2.5}$

$\sigma_x = 1.58$, $\sigma_y = 2.073$ $\text{variance} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$

$\sigma = \sqrt{\text{variance}}$

$$\rho(x,y) = \frac{-3}{1.58 \times 2.073} = \frac{-3}{3.27} = -0.917$$

$x \uparrow \; y \downarrow \Rightarrow 91.7\%$ negative correlation $\rightarrow 91.7\%$

1 unit $x\uparrow$ $y\downarrow$ 0.917

- Pearson Correlation Coefficient does not work well with non-linear data

③ Spearman Rank Correlation [-1 to 1]

$$r_s = \frac{cov(R(x), R(y))}{\sigma(R(x)) * \sigma(R(y))}$$

Ascending order

x	y	R(x)	R(y)
1	2	2	2
3	4	3	3
5	6	4	4
7	8	5	6
0	7	1	5
8	1	6	1

Spearman Rank Correlation

⇓

Non linear relationship