Statistics

* What is Stats?

→ It is a branch of maths which deals with collecting, organizing, interpretating & presenting ~~data~~ Data.

→ Data summarizing / ~~exploration~~, hypothese testing to make some decision.

* What is Data?

→ Facts or pieces of information that can be measured.

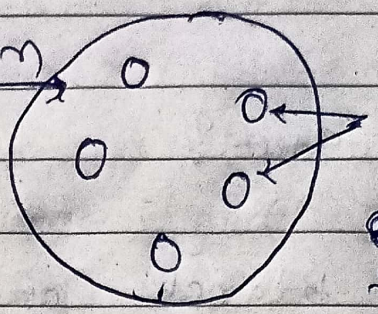
→ DIKW (Data (Raw facts), Information, knowledge, wisdom / insights).

* Types of Stats :-

→ 1. Descriptive → Organization & summarization.
→ 2). Inferential. → Techniques using which we can conclude something from Data or inference.

* Sample & Population :-

Whole Dataset → Population (N)



→ Sample (n)

↓

small subset taken from population.

* Sampling Techniques :-

1. Simple Random Sampling :-

→ Every member of population have an equal chance of getting selected in your sample.

2. Stratified Sampling :-

→ Where we split the population into non-overlapping groups (Strata).

→ Ex :- Gender → Male
→ Female

Age → 0 - 20
→ 21 - 40
→ 41 - 60

3. Systematic Sampling :-

→ Surveying an (N^{th}) individual

4. Convenience Sampling

* Variable :-

→ It is a property that can take any value.

→ Types of variable :-

1. Quantative :- Measurable (numeric)

2. Qualitative :- Categorical value.

→ Types of Quantative variable :-

1). Discrete :- Whole no

2). Continuous :- Decimal value

* Variable Measurement Scales :-

1. Nominal :- Categorical - No order

2. Ordinal :- Ordered (order matters)

3. Interval :- [No zero/absolute Point], Ex:- Calender/ Temperature

4. Ratio :- Ex. :- Height / weight / Marks.

* Frequency :-

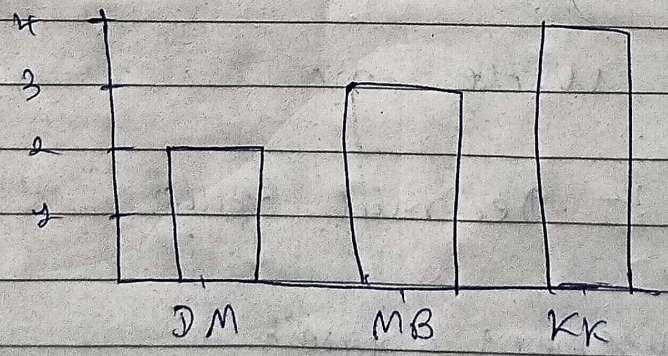
→ It means how many times value ~~refers~~ repeats.

→ Ex :-

1. [DM, MB, KK, KK, KK, MB, DM, KK, MB]

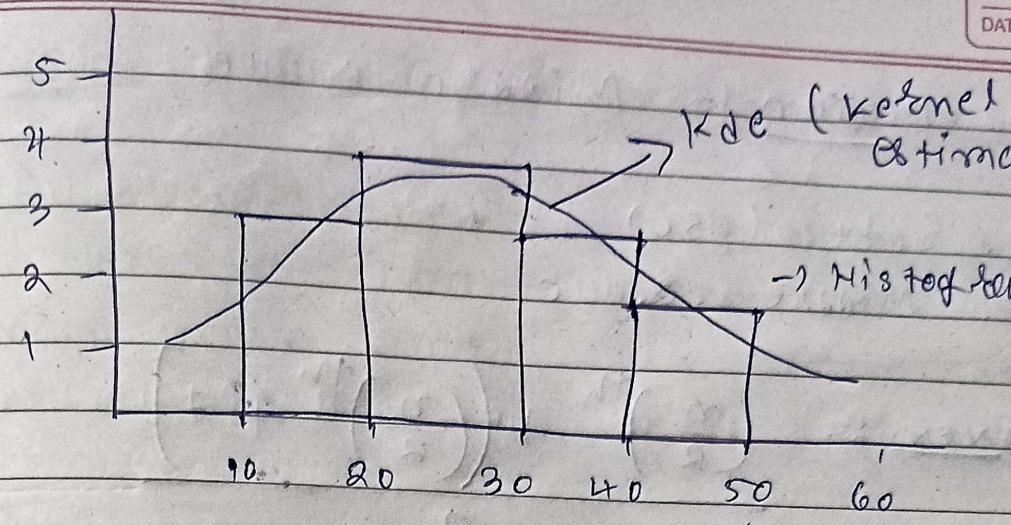
Chocolates	Frequency	Commulative freq.
DM	2	2
MB	3	5
KK	4	9
	9	

Bar graph



2. Age → [11, 12, 15, 25, 26, 29, 24, 33, 37, 39, 42, 47]

10 - 20	20 - 30	30 - 40	40 - 50
3	4	3	2



* Descriptive Stats :-

→ Descriptive stats use methods for summarizing & understanding the key characteristics of data.

1. Measure of central Tendency :-

→ Measure of CT describes the "center" or typical value of a dataset. They provide a summary of the data with a single value that represent the dataset as a whole.

1) Mean :- Avg → refers to the measure used to determine the center of distribution.

μ → Pop avg

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

$\bar{x}$ → sample avg

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

2> Median :- Middle value.

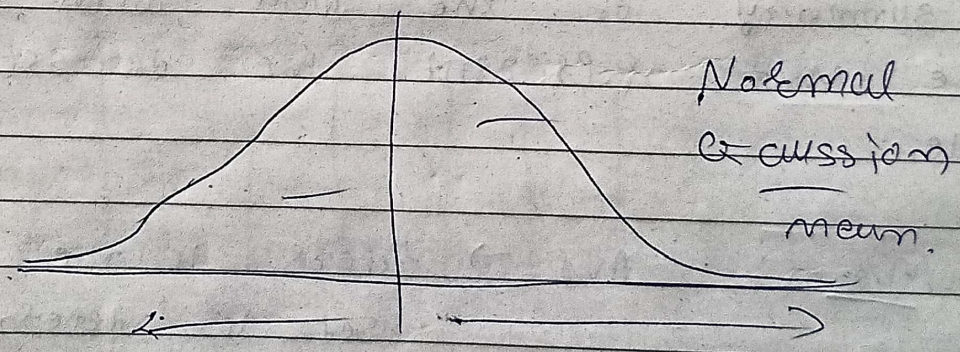
i). Sort (asc)
ii). Middle value

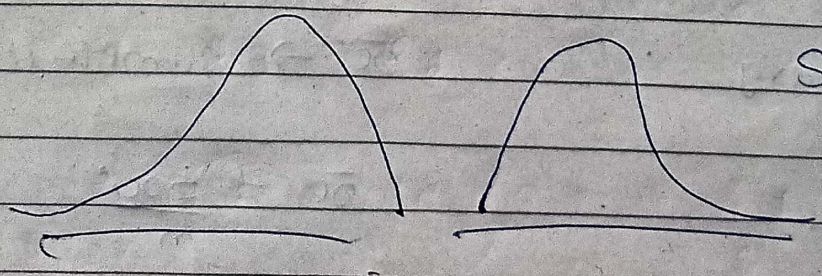
→ Even :- $\frac{\left(\frac{n}{2}\right)^{th} + \left(\left(\frac{n}{2}\right)^{th} + 1\right)}{2}$

→ Odd :- $\left(\frac{n+1}{2}\right)^{th}$

→ Outliers :- a datapoint(s) who doesn't follow the pattern or trend of the data / or are extreme points.

3> Mode :- Most frequency value





2. Measure of Dispersion :- Spread

→ MOD describes the spread of variable of interest. They indicates how much individual data points differ from the central value.

1). Variance :-

→ Variance measure how far each data pt in the dataset is from the mean.

→ High variance → more spread
→ Low " → less "

Pop

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

Sum

$$s^2 = \frac{\Sigma (xi - \bar{x})^2}{n-1}$$

Bessel's correction.

2). Standard Deviation :-

→ Sqrt of variance.

Pop

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

Sum

$$s = \sqrt{\frac{\Sigma (xi - \bar{x})^2}{n-1}}$$

→ It gives you a measure of spread that is in the same units as the original data, making it easier to interpret than variance.

* What is Percentage and percentile?

→ A percentile is a value below which certain percentage of observations lie.

$$\text{Per}_{10} = \frac{\text{no. of values below } x}{n}$$

$$25\%\text{ lie per} = \left(\frac{\text{percentile}}{100} \times n\right) + 1$$

⇒ Five Number Summary :-

1. Minimum
2. Q_1 → 25 percentile
3. Median
4. Q_3 → 75 "
5. Max

IQR = Inter quartile Range

$= Q_3 - Q_1$

Lower fence / min $= Q_1 - 1.5\,IQR$

Higher " / max $= Q_3 + 1.5\,IQR$

PAGE NO.:
DATE: / /

* Distribution :-

→ The possible values a variable can take & how frequently they occur.

Y → the actual outcome of an event.
y → one of the possible outcome.

* Probability :-

Sample space → Total possible outcomes.

1. Mutually exclusive
2. Collectively "

⇒ Event :-

→ Subset of sample space.

→ Axioms :-

1). non - negative. ∴ $P(A) \geq 0$
2). Normalisation :- $P(\Omega) = 1$
(Ω ∪)
3). Additivity :- if $A \cap B = \phi$
$P(A \cup B) = P(A) + P(B)$

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

PAGE NO.:
DATE: / /

⇒ Multiplication Rule :-

→ To check 2 or more event can occur together.

→ Two versions :-

1). Independent :- 2 outcome independent
2). Dependent :- 2 outcome depend

1). Independent event :-

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

2) Dependent event :-

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

→ conditional prob.

Prob B when A has already occured

$$P(B/A) = \frac{\text{Prob of what we want to know}}{\text{Prob of what we already know}}$$

⇒ Probability Distribution function :-

→ It gives the probability that a random variable takes a specific value or lies within a certain range.

1.) Discrete → Probability Mass fn (PMF)

2). Continuous → Probability Density fn (PDF).

*Types of Distribution :-

1. Discrete :- limite no of outcomes

2. Continuous :- infinite no of outcomes

→ $X \sim N(\mu, \sigma^2)$

X → variable; N → type of Dis.; (μ, σ^2) → characteristics.

1). Discrete Dist

i) Bernoulli dist :-

out → 2 → 0, 1 → Equiprobable

ii) Binomial :- Bernoulli → extension

iii). Poisson dist

2). Continuous :-

i) Gaussian / Normal :-

ii) Chi-square dist

1) Uniform dist [discrete]:-

→ U(a, b) → range of values in a setset.

→ variable X follows uniform dist with range a to b.

$$X \sim U(a, b)$$

→ all outcomes → equi probable
P(a) = P(b)

2) Bernoulli dist:-

→ Bern (P) → Prob of success

$$X \sim \text{Bern}(P)$$

→ Event:-

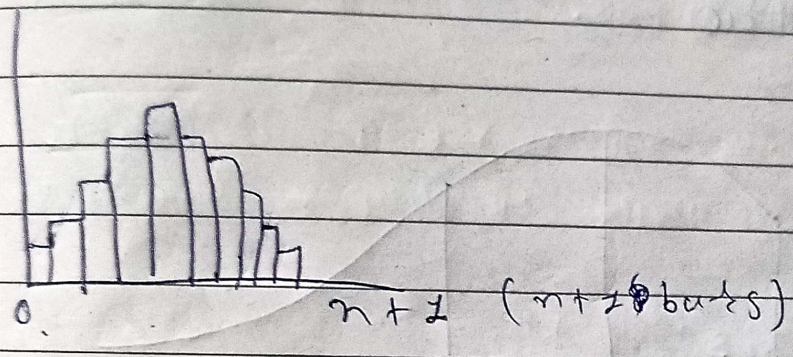
1. 1 trial
2. 2 possible outcomes (T/F)

→ $E(X) = 1 \times P + (0 \times (1-P))P$

3) Binomial Dist:-

→ ~~Satisfie~~ Sequence of Bernoulli event identical

$X \sim B(n, p)$ → Prob of success in each trial



→ $P(y) = {}^{n}C_{y} \cdot p^{y} \cdot (y-p)^{n-y}$

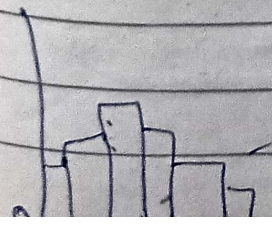
→ $E(X) = n \cdot p$

4) Poisson Dist:-

$P_o(\lambda)$ → frequency

$P(y) = \frac{\lambda^{y} e^{-\lambda}}{y!}$ e = euler's

$E(y) = \lambda$



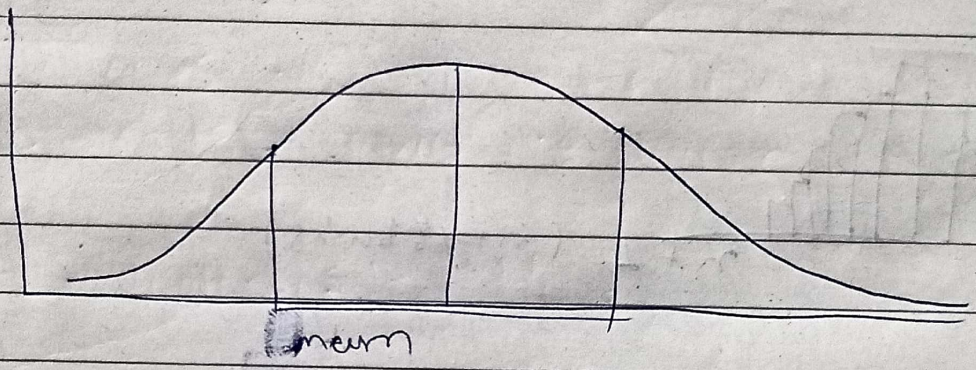
→ no of occurrence.

5) Normal dist (Gaussian) :-

$X \sim N(\mu, \sigma^2)$

ND → Bell shape majority data → mean

Graph → symmetric



$E(X) = \mu$

mean = median = mode

68% → $\mu \pm \sigma$

95% → $\mu \pm 2\sigma$

99.7% → $\mu \pm 3\sigma$

outliers are extremely low/large

→ Standardization :-

$z = \frac{x_1 - \mu}{\sigma}$

score

Standard ND

$\mu = 0$ $\sigma = 1$

ND → SND

$n > 30$

$t \rightarrow$ Dist

$y \sim t(3)$ → Degree of freedom

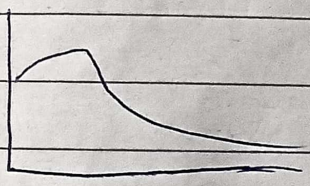
6) Chi - square dist :-

→ asymmetric

→ skewed

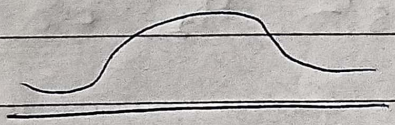
$y \sim t(k)$, $X \sim \chi^2(k)$

$y^2 \sim \chi^2(k)$, $\sqrt{X} \sim t(k)$

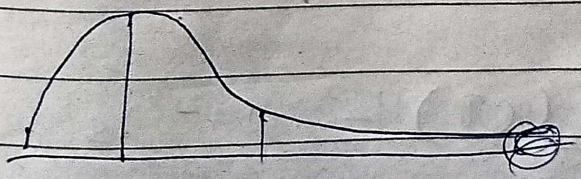


$E(X) = K$

→ skewness

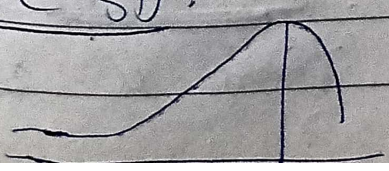


1. + ve SD :-



Mode < median < mean.

2. -ve SD :-



mean < median < mode

* Inferential Stats:-

⇒ Hypothesis testing :-

→ A hypothesis test is a method of statistical inference used to decide whether the data at hand is sufficient to support a particular hypothesis.

→ It allows us to make probabilistic statements about population parameters.

1) NULL Hypothesis :- H_0

→ It assumes that there is no relationship b/w two variables [simply saying nothing new is happening]

2) Alternate Hypothesis :- H_1 or H_a

→ It states that there is significant relationship b/w two variables.

→ Confidence Interval (CI) :- It is a range of values within which we expect a particular population parameter to fall.

→ It's a way to express uncertainty around an estimate obtained from a sample of data.

→ Confidence Level :- It refers to the ~~deg~~ degree of certainty that the statistical estimate will contain the true population parameter.

90 % → Narrower CI → ~~Precise~~ precise

95 %

99 % → wider CI → less precise.

⇒ Loss of Significance :- (α)

→ It is threshold for ~~deci~~ deviding when to reject the H_0. It defines the probability of making a Type I error, which is the error of rejecting H_0 when it is actually true.

* Rejection Region Approach :-

1. H_0, H_a State
2. α → Significance level $\alpha = 0.05$
3. Assumption / value note down.
4. Decide test
5. value
6. Test conduct
7. Reject / Accept Decision
8. Results.

⇒ Types of Error :-

→ Type - I :- Rejecting H_0 when it is actually correct.

→ Type - II :- Accepting H_0 when H_0 is actually incorrect.

* Z - Test :- $Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

1. Random Sampling (Normalized)
2. σ
3. $n \geq 30$

* T - Test :- $df = n - 1$

→ One - sample t - test.

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

→ Independent - 2 Sample t - test.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1}{n_1} - \frac{s_2^2}{n_2}}}$$

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2 / n_1 - 1 + \left(\frac{s_1^2}{n_2}\right)^2 / n_2}$$

→ Pairs t - test.

$$t = \frac{\bar{x}_d}{s_d / \sqrt{n}}, \quad \bar{x}_d \rightarrow \text{mean difference}, \quad s_d \rightarrow \text{SD}$$

PAGE NO. :
DATE : / /

* Chi-Square Test :-

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

O = Observe Freq

E = Expected "

$$E = \frac{\text{Row total} \times \text{col total}}{\text{Grand total}}$$

$$df = (r-1) \times (c-1)$$