Stat

(1) What is statistics?

→ Collecting, analyzing, interpreting and drawing conclusion.

(2) What is data?

(3) Type of stats.

1. Descriptive stats:
It consist of organizing & summarizing data. (graphs, tables, descriptive

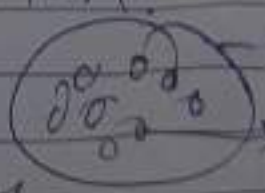
2. Inferential stats:
using Data, we can make conclusion using same techniques

etc.

(4) Sample and Population.

Population - N
→ whole collected D
known as population

sample - n
→ small subsets of data taken from pop



5 sampling techniques

(1) simple Random sampling.

↳ Every member of the population has an equal chance of getting selected in sample(n)

(2) Stratified sampling.

↳ splitting data into non-overlapping groups.
eg: age, gender.

(3) systematic sampling.

↳ Selects every nth member of the population after a random starting point.

(4) Convenience sampling.

↳ Sample are chosen based on ease of access and availability.

Variables

↳ It is a property that can hold/store / take any value.

eg. { 8, 10, 15, 20 ... 25 ... }
marks : { 78, 80, 95 ... }.

⟹ type

1. Qualitative Variable (Categorical Value)

→ based on some characteristics can derive categorical values

Eg	0 – 10	–	Low
	10 – 50	–	Avg
	> 50 - 90	–	Good.

2. Quantitative Variable (Numerical value)

→ measurable numerically.

→ Height - { 162, 159, 165 ... }
weight - { 57, 55, 56 ... }

i) Discrete - value are countable and finite (whole numbers) (eg. no. of students)

ii) Continuous - value can take any value within range. (decimal no.) eg. Height

Variable Measurement scales.

① ordinal scale. - (order)
ex- rank, satisfaction.

② Nominal scale - (categorical value) (no order)
↳ colors, classes, degree.

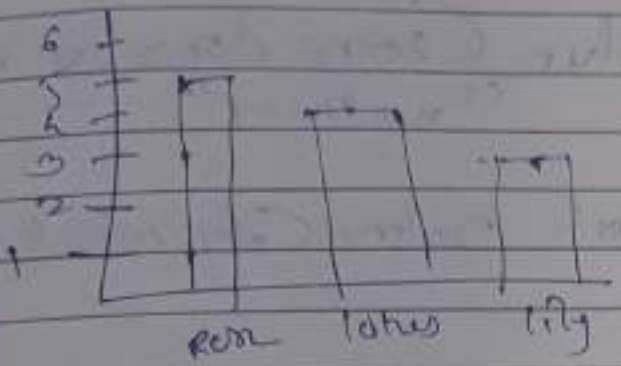
③ Interval scale : (No zero / absolute point)
ex- temperature.

④ Ratio scale. (zero means nothing)
ex- Height, weight, or age
BP, Income.

Frequency

data : flowers
{Rose, lily, lotus, Rose, rose, rose, lotus, lily, lotus, lily, lotus}

Flowers	frequency	CF cumulative
Rose	5	5
lily	3	8
lotus	4	12
	12	



bar graph / chart

Histogram

marks = [12, 15, 13, 15, 21, 21, 24, 35, 34, 36, 39, 41, 45]

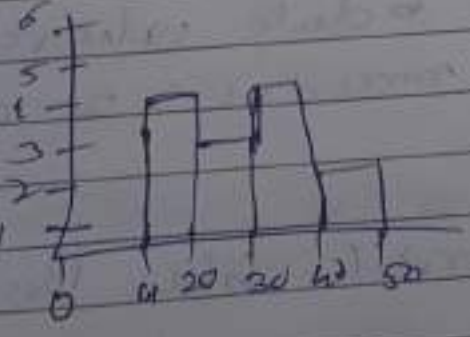
bins

10-20 4

20-30 3

30-40 4

40-50 2



Interval scale -> Diff measure the diff btw value but does not have a zero points

→ no absolute value (zero does not mean "nothing")

→ can not calculate ratio (eg, 20/10 ≠ 2)

Ex: temp

0°C does not mean "no temperature"

20°C is not twice as hot as 10°C

Celsius.

i.e. b/w 2000 & 2020 is 20 years but 2020 is not twice of 10.

Ratio scale.

It measures the absolute where there is zero point meaning of zero refers complete absence.

→ Equal interval b/w values.

→ zero allows ratios

→ can perform all maths

Measure of Central tendency

Avg → Mean.

pop. Sam.

$$\mu = \frac{\sum x_i}{N} \qquad \bar{x} = \frac{\sum x_i}{n}$$

mean: It refers to the measure used to determine the centre of the distribution or the data.

{1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 10} ← outlier

$$\frac{32}{10} = 3.2 \qquad \frac{132}{11} = 12$$

Median → middle value

Even

$$\frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th}}{2}$$

Odd - data = {11, 12, 13, 14, 17, 18}

n > 6

$$\therefore \frac{\left(\frac{6}{2}\right)^{th} + \left(\frac{6}{2}\right)^{th}}{2} \text{ th}$$

$$\therefore \frac{3^{th} + (3+1)^{th}}{2} = 13.5$$

Odd

$$\left(\frac{n+1}{2}\right)^{th}$$

{ 11, 12, 13, 14, 15 }

$$\left(\frac{5+1}{2}\right)^{th} = \frac{6^{th}}{2} = 3 \rightarrow (3)$$

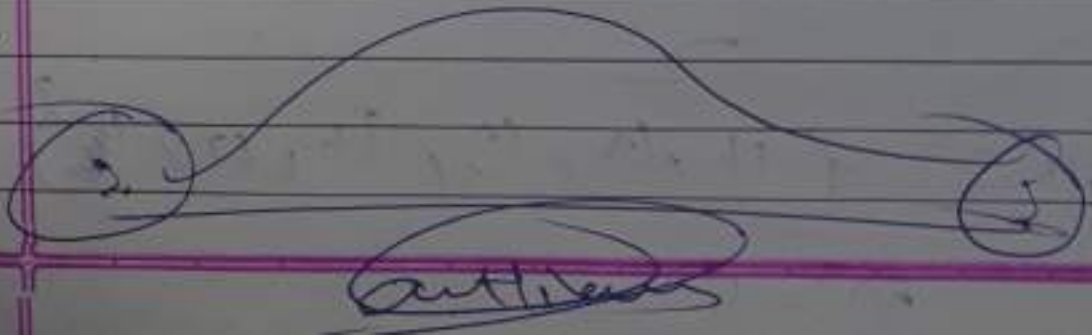
(5) outliers

{ 21, 23, 25, 29, 3 & 100 }

$$\frac{25+29}{2} = \frac{54}{2} = 27$$

outliers

→ a data set who doesn't follow pattern or trend in data set then it is consider as outliers

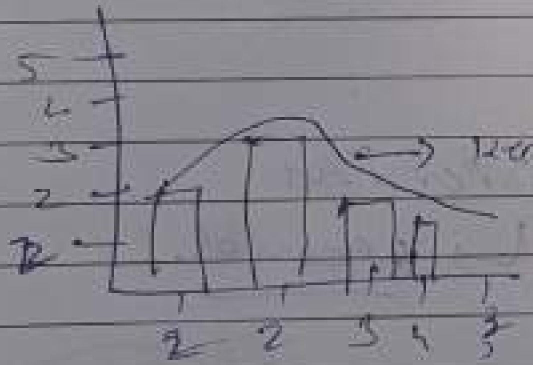


Mode

→ most frequent value (repeated)

{1, 2, 2, 3, 4, 2, 3, 1, 0}

Mode = 2



⇒ for categorical many data

0 - 5 %. mode

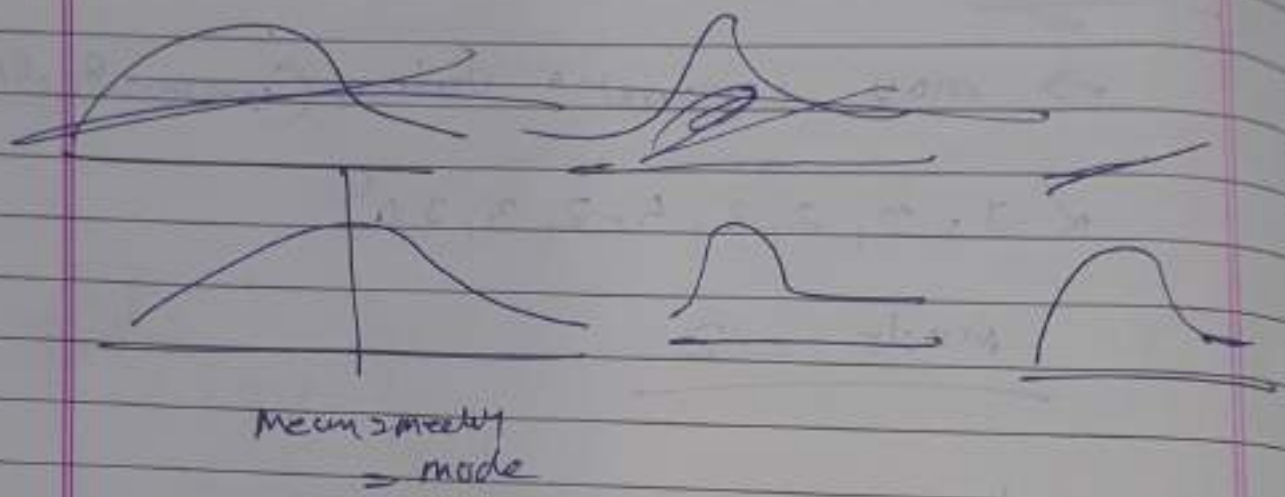
→ new category "missing" "values" "unknown"

or →

⇒ for numerical value

Gaussian distribution

Skewed & median.



Measures of dispersion

by spread.

[5, 1, 1, 1, 2] → $\frac{5+1}{5} = 2$

[2, 2, 2, 2, 2] $\frac{}{2} = 2$

→ Variance (σ)

It measures how far the numbers in a dataset are from the mean.

(how each value differ from a dataset or mean).

High Variance → more spread

Low " → closer to mean

Pop.

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

Sum

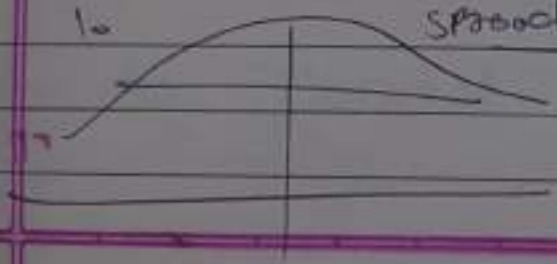
$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

	$\bar{x}$	$x_i - \mu$	$(x_i - \mu)^2$
1		-1.83	3.34
2		-0.83	0.68
3		-0.83	0.68
4		0.17	0.02
5		1.17	1.36
6		2.17	4.70
Σ		0.02	10.78

$$\sigma^2 = \frac{10.78}{6} = 1.79$$

$$\sigma = 1.33$$

1. Spread 1



2.

* Standard Deviation

→ Just a squared mean of Variance

→ It gives your a measure of spread that is in the same unit as the original data, making it easier to interpret than Variance

[Same Unit, easily comparable]

For.

$$\sigma = \sqrt{\frac{\Sigma (x_i - \mu)^2}{N}} \qquad \text{Sample} \quad s = \sqrt{\frac{\Sigma (x_i - \bar{x})^2}{n-1}}$$

* (key point (Variance & SD))

→ Variance gives you the average squared distance from the mean.

→ SD gives you a measure of spread the same unit

Variance Formula

* Pop. $\frac{\Sigma (x-\mu)^2}{N}$ (N = total no. of data pts)

When we have data from the entire population, we use 'N' in the denominator. This gives us an exact measure how the data pts vary around the Pop. mean (μ)

* Sample $\frac{\Sigma (x_i - \bar{x})^2}{n-1}$

When we working with a sample, we only have sample mean $\bar{x}$, which is an estimate of the POP. mean μ.

→ Using Sample mean in calculating tend to make the Variance slightly smaller than than the true POP. Variance.

to Corrected this bias (Underestimating variability)

→ To Correct this bias we divide by 'n-1' instead of n. This makes variance unbiased. accurate.

→ By Subtracting 1 from n, we account for the fact that the $\bar{x}$ is not perfectly representative of μ this correction ensures that the sample variance is an unbiased estimator of the pop. variance.

* Percentage

1, 2, 3, 4, 5

∴ 3/5 of the no.s = 0.6 → × 100 → 60%

* Percentile

→ A Percentile is a value below which a certain p.c. observation lie

dataset = [2, 3, 3, 4, 6, 5, 5, 6, 7, 8, 8, 7, 9, 16, 11, 11, 12]

P.P = No of Value below x / N

= 16/2 = 80

Q. What Value exist at percentile ranking of '25'?

Value = (Partile / 100 × N) + 1

= (25/100 × 20) + 1

= 5 + 1 = 6th

75 percentile value

= 75/100 × 20 + 1

= 16th

Five number Summary

1. minimum
2. first quartile (Q_1)
3. median
4. third quartile (Q_3)
5. maximum

[1, 1, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 7, 8, 8, 9, 10, 23]

lower fence = $Q_1 - 1.5$ (IQR)
Higher " = $Q_3 - 1.5$ (IQR)

Q_1 = 25 Percentile (25%)

Q_3 = 75 " (75%)

IQR = Inter quartile range.

= $Q_3 - Q_1$

$Q_1 = \left(\frac{25}{100} \times 20\right)^{th}$

$Q_1 = 3$

$Q_3 = \left(\frac{75}{100} \times 20\right)^{th}$

$= 15$

$= 8$

$IQR = 8 - 3 = 5$

Lower Fence: $Q_1 - 1.5\,IQR$

$= 3 - 1.5(5)$

$= -4.5$

Upper Fence $= Q_3 + 1.5\,IQR$

$= 8 + 1.5\,IQR$

$= 8 + 7.5$

$= 15.5$

$Q_1 = \left(\frac{25}{100} \times 20\right)$ th

$Q_1 = 3$

$Q_3 = \left(\frac{75}{100} \times 20\right)$ th

$= 15$

$= 8$

$IQR = 8 - 3 = 5$

Lower Fence: $Q_1 - 1.5\,IQR$

$= 3 - 1.5(5)$

$= -4.5$

Upper Fence: $Q_3 + 1.5\,IQR$

$= 8 + 1.5\,IQR$

$= 8 + 7.5$

$= 15.5$

Terminology data

1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 7, 8, 9, 10, 12, 14

S.NO terms

MM = 1
Q1 = 3
Median = 5
Q3 = 8
Mean = 15

Box plot

Outliers

Data Distribution.

- If refers to a way in which values or data points are spread or arranged.
- It shows how often different values occur in data set & describes the overall pattern of the data.

→ Center.

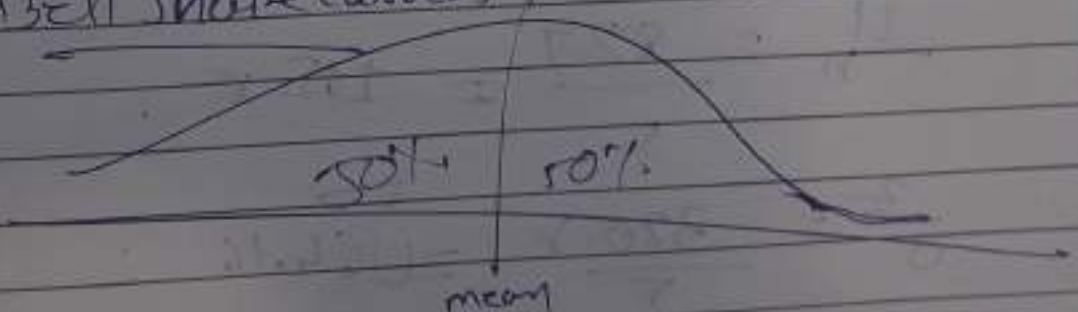
→ spread

→ shape.

→ outliers.

Gaussian / Normal distribution. (Bell shape)

Bell shape curved.



$\mu \pm 1\sigma = 68\%$

$\mu \pm 2\sigma = 95\%$

$\mu \pm 3\sigma = 99.7\%$

Empirical rule

Z-Score

$$Z\text{-Score} = \frac{x_i - \mu}{\sigma}$$

Example

Height	weight	$h_i - \mu_h$	
164	60	3.2	10.24
172	65	5.2	38.44
150	45	-15.8	249.65
168	70	2.2	4.84
170	77	4.2	17.64
829			320.8

$$\mu_h = \frac{829}{5} = 165.8$$

$$\sigma^2 = \frac{320.8}{5} = \sqrt{64.16}$$

$$\sigma = 8.007$$

$z_1 = \frac{169 - 165.8}{8.007} = \frac{3.2}{8.007} = 0.4$

$z_2 = \frac{6.2}{8.007} = 0.7$

$z_3 = \frac{+15.2}{8.007} = +1.9$

$z_4 = \frac{2.2}{8.007} = 0.27$

$z_5 = \frac{4.2}{8.007} = 0.52$

≈ 0

Normalization

$$x = \frac{x - x_{min}}{x_{max} - x_{min}}$$

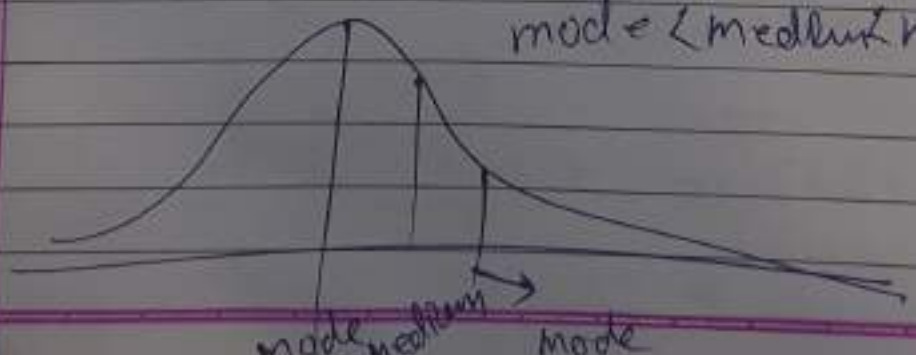
positively skewed distribution

What?

In positively skewed distribution, most values are concentrated on the lower end, with a long tail extending to the right. A few high values pull the average to the right of the median.

Skewness :→ A distortion or asymmetry that deviates from the symmetrical bell curve

→ Also called right-skewed or right tailed distribution.



when?

Useful for data with some but significantly high values such as income levels when few individuals earn much more than the rest.

⇒ the mean is greater than the median due to outliers on the higher end.

Negatively skewed distribution,

~~when?~~

→ why?

Because it makes comparing different data sets easy.

eg: 2022

D1 → AM = 25° VK = 263 σ = 10

D2 : AM → 280 VK = 261 σ = 12

$z_1 = \frac{265 - 255}{10} = 1$

$z_2 = \frac{263 - 260}{12} = 3/12 \quad 0.25$

$z_3 = \frac{263 - 255}{10} = 8/10 = 0.8$

$z_4 = \frac{265 - 250}{12} = 5/12 = 0.41$

2/12

Exponential Distribution. (λ = lambda) (constant rate)
(continuous)

→ It describes the time b/w events in a process where events occur independently at a constant rate λ.

$$\Rightarrow f(x) = \lambda e^{-\lambda x} \quad x \geq 0$$

Bernoulli distribution:
(Discrete)

It models a single experiment with two possible outcomes

⟹ Success ($x = 1$); failure ($x = 0$)

Binomial distribution: (Discrete)

Binomial distribution extends Bernoulli to n independent trials.

$$\Rightarrow P(x = k) \Rightarrow \binom{n}{k} p^k (1-p)^{n-k}$$

x = random variable (no. of success in n trials)

n = total number of trials / P = P(success in 1 trial) (0.1)

k = number of success (0 ≤ k ≤ n)

$= P = P(\text{success, } n, \text{ trial})$

$$\binom{n}{k} = \frac{n!}{(n-k)!\,k!}$$

(g) $n = 5 \quad k = 3$

$$P^{(3)} = \frac{5!}{2! \cdot 3!} \times 0.3^{(3)} \times 0.7^{(2)}$$

Uniform distribution

⟹ (continuous)

$[a, b]$

$$P(x) = \frac{1}{b-a} \qquad a \le x \le b$$

The probability of any value coming in the range $[a, b]$ is same

⇒ (discrete) :

'all outcomes are equally likely.

$$P(x) = \frac{1}{n} \rightarrow \text{total no.}$$

Confidence Intervals (CI)

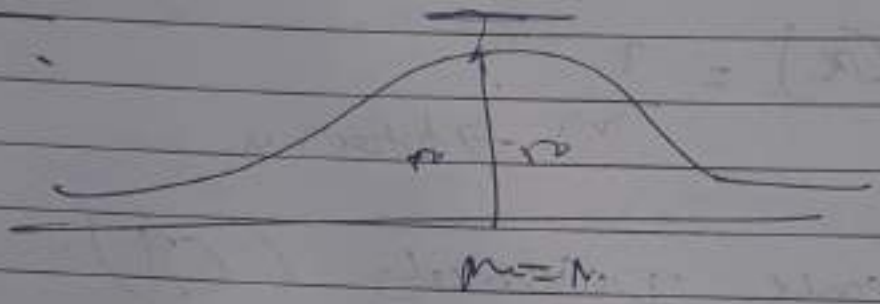
$\bar{x} = 50 \rightarrow$ point estimate - μ

It is a range of values within which we expect a particular population to fall.

Confidence Interval = point estimate ± margin of error

Confidence level

→ Sampling & Statistical parameter



Hypothesis testing

A statistical hypothesis test is a method

→ Null Hypothesis (H_0)

→ Alternate Hypothesis (H_A)

25/12

Rejection Region Method

1. Ho & Ha
2. α ⇒ value → LOSC (Level of significance)
   ↳ Significance level → 95%.
3. Computing.
4. decide test → z-test, t-test
5. Value.
6. test conclusion
7. Reject / Accept
8. State the Reject or Reject.

* Probability → possibility chance.

→ It is a measure of the likelihood of an events

Eg - Roll a dice {1, 2, 3, 4, 5, 6} → sample space

No. of ways an event can occur
total no. of possible outs.

a even no:

P(E) = {2, 4, 6} → Favourable.

{1, 2, 3, 4, 5, 6}

= 3/6 = 1/2 = 0.5

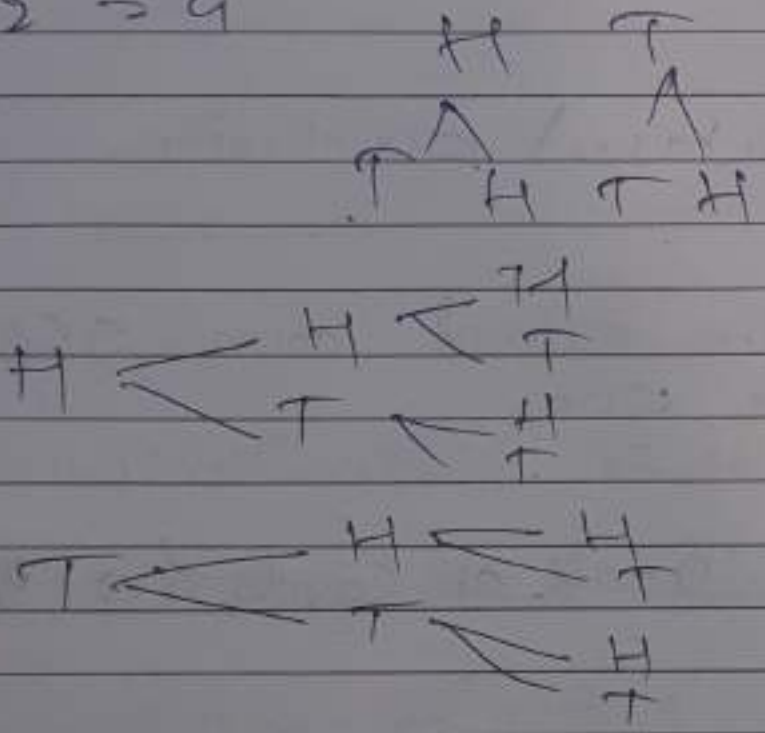
* two event are occuring together

→ tossing one coin

H T → $2^0 = 2$

toss 2 coin

$2^2 = 4$



dice roll → $6^1 \rightarrow 6$

{(1,1) (1,2) }

* Mutual Exclusive event

two events are called mutual exclusive if they can not occur at the same time.

Roll a dice {5,6} → X

roll a coin {H,T} → X

* Non-mutual exclusive.

Multiple event can occur at the same time.

⇒ - Deck of cards {9, ♥}

* Addition rule (or - probability)

Q I toss a coin, what is the probability that the coin comes either on head or tail ?

mutual exclusive

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

$= \frac{1}{2} + \frac{1}{2}$

$P(A \text{ or } B) = 1$

Dice → 1, 3, 6

P(1 or 3 or 6) = 1/6 + 1/6 + 1/6

= 3/6

= 1/2 = 0.50

(3) Multiplication Rule.

* Independent Event

Dice {1, 2, 3, ...}

2 dice → indi

* Dependent event

P(G) = 3/5

P(P) = 2/4

Q. What is the prob of getting a "J" from

Roll a dice {5,6} → X

roll a coin {H,T} → X

Non-mutual exclusive.

multiple event can occur at the same time.

Ex- Draw of cards {9, ♥}

Addition rule for probability

Q I toss a coin, what is the probability that the coin lands either on head or tail ?

mutual exclusive

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

$= \frac{1}{2} + \frac{1}{2}$

$P(A \text{ or } B) = 1$

① Dice → 1,3,6

$P(1 \text{ or } 3 \text{ or } 6) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6}$

$= \frac{3}{6}$

$= \frac{1}{2} = 0.50$

Multiplication Rule.

* Independent Event

① Dice {1, 2, 3, ...}

2 dice → 3 ...

1/6 1/6

* Dependent event

$P(G) = 3/5$

$P(P) = 2/4$

Q What is the prob of pulling a "J" from

Hypothesis testing

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

Null Hypothesis Ho.

The null hypothesis assumes that there is no significant relationship or effect b/w two variable

It serves as a starting point for HT & represents the 'status quo' or the assumption of no effect until proven otherwise.

The purpose of HT is to gather evidence to reject or fail null hypothesis in favour of alternative.

hypothesis, which claims there is significant effect or relation.

Alternate Hypothesis

It is a statement, that contradicts contradicts the NH & claims there is significant effect or relation.

Rejection Region Method

1. H_0 & H_1
2. α → value. ⟷ LOS
   ↳ significance level → 95%
3. ~~assumptions~~ assumptions.
4. decide test → z-test
   t-test
5. Value.
6. test conclusion
7. Reject / Accept / ∴ that proves

Q.1. 50 → Unit per day

$\sigma = 5$

30 days → 53 unit per day

1. $\mu = 50 \rightarrow H_0 \quad H_a = \mu > 50$

2. $\alpha = 0.05$, U.T. on left test

3. data - Normal, σ, random.

$n = 30$

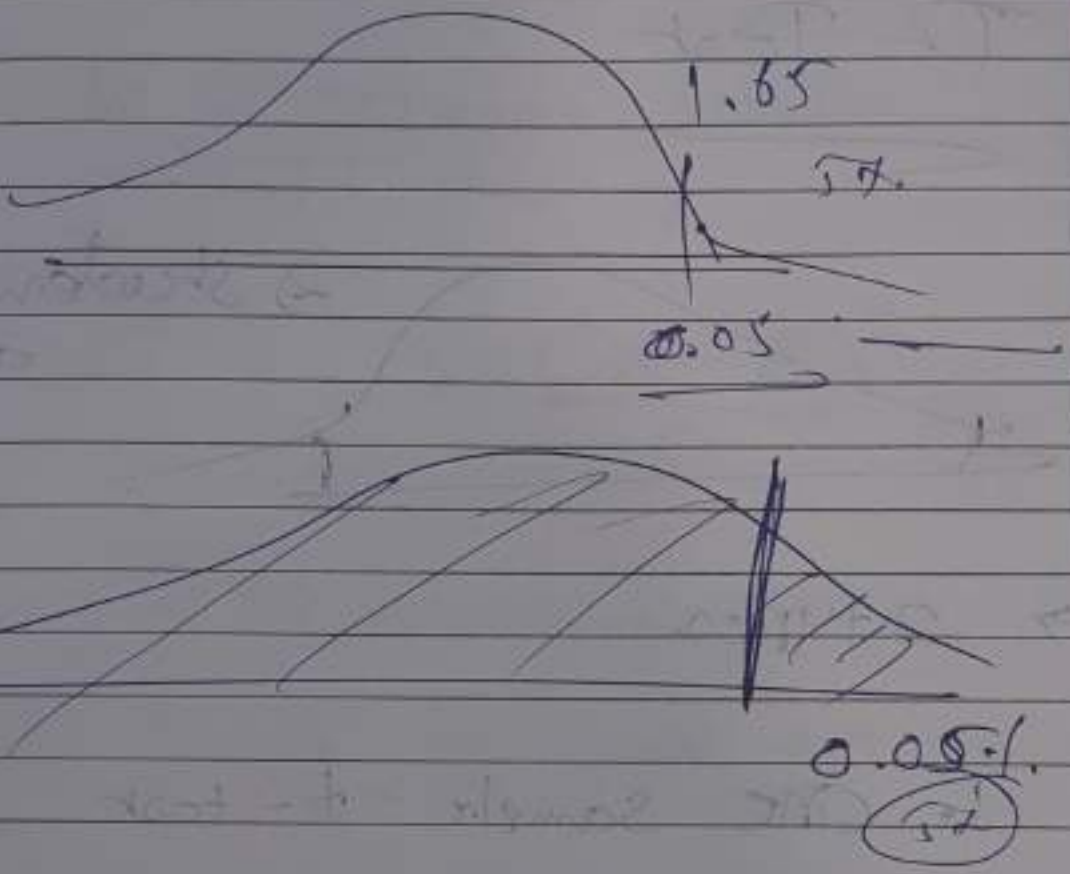
4. Z - test

5. Z - Score $= \frac{x_1 - \mu}{\sigma}$ - P.P.

$\frac{x_1 - \mu}{s/\sqrt{n}} = \frac{53 - 50}{5/\sqrt{30}}$

$Z = 3.28$

σ



∴ Rejection of H_0

& $\mu > 50$

T- Test

→ 3 types

1. One sample t-test

Compares the mean of single sample to a known μ

2. Independent two sample t-statistic

3. paired t-test

Chi - Square test

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

O → Observed freq.

E → Expected freq.

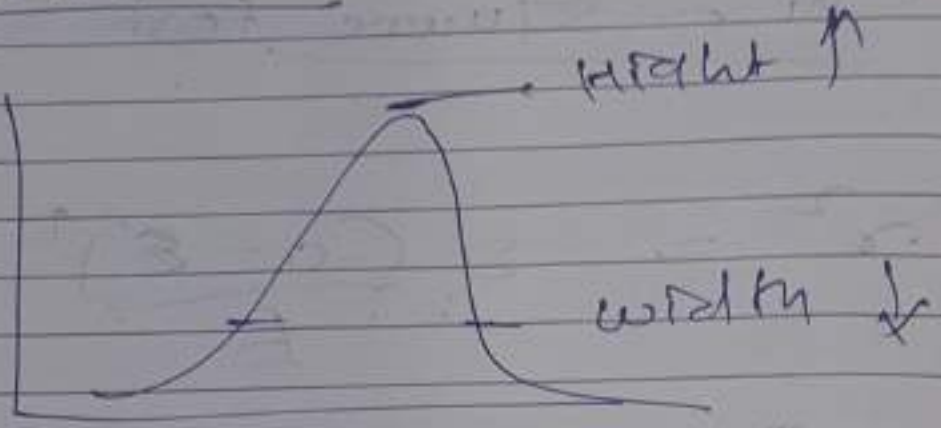
E = $\frac{\text{Row total} \times \text{Coll. total}}{\text{Grand total}}$

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

r = no. of row

c = no. of coll.

Kurtosis



→ kurtosis measures "tailed nell" of a distribution or how extreme the outliers are.

→ 1. Mesokurtic.

→ tails are similar to ND

→ ~~distribution~~ distribution with kurtosis ≈ 3.

→ Eg. standard normal distribution.

→ 2. Leptokurtic.

→ A distribution with kurtosis > 3

→ Heavy tails (with more/extreme outliers)

Platykurtic

→ distribution with kurtosis < 3

→ light tails.

→ Eg - Uniform dist.

Excess kurtosis → K - 3

K = kurtosis

EK > 0 → leptokurtic

EK < 0 → platy "

$$K = \frac{n \cdot \Sigma (x_i - \bar{x})^4}{(\Sigma (x_i - \bar{x})^2)^2} - 1/n$$

n = no. of observations

x_i = each data points

$\bar{x}$ = mean

⟹ High kurtosis (leptokurtic)

- More extreme outliers.
- Higher likelihood of rare, extreme values.
- ~~Higher~~ 1.
  eg:- Financial return during market crisis / ~~crash~~ crash.

- Lower kurtosis (platykurtic)
  - Fewer extreme outliers.
  - Data is evenly spread.
  - kurtosis near 3 (no outliers)
  - Similar to normal distribution