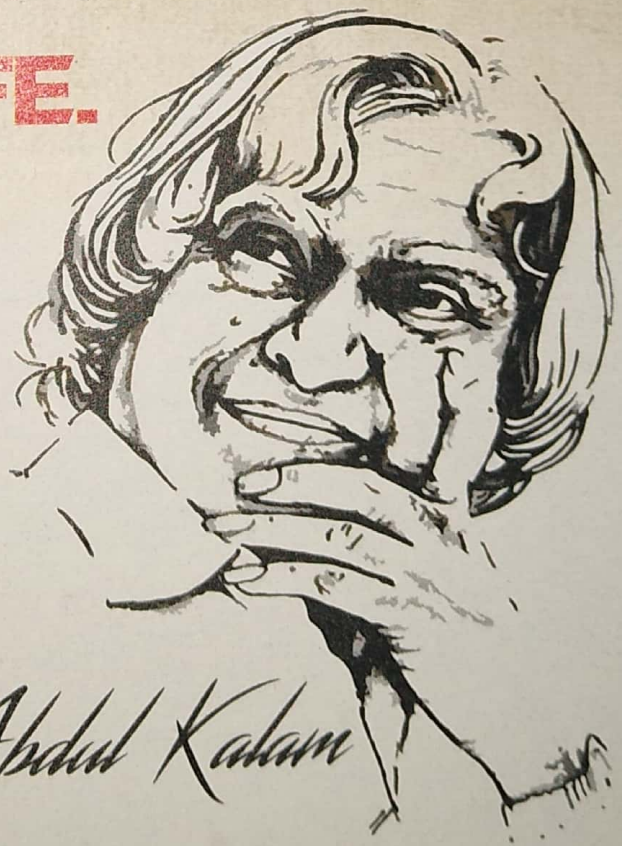
Book 1

THINKING SHOULD BECOME YOUR CAPITAL ASSET. NO MATTER WHATEVER UPS AND DOWNS YOU COME ACROSS IN YOUR LIFE.



Inspirational Quotes by

-A. P. J Abdul Kalam

INDEX

Name SANDEEP. A.C

Standard Section Roll No.

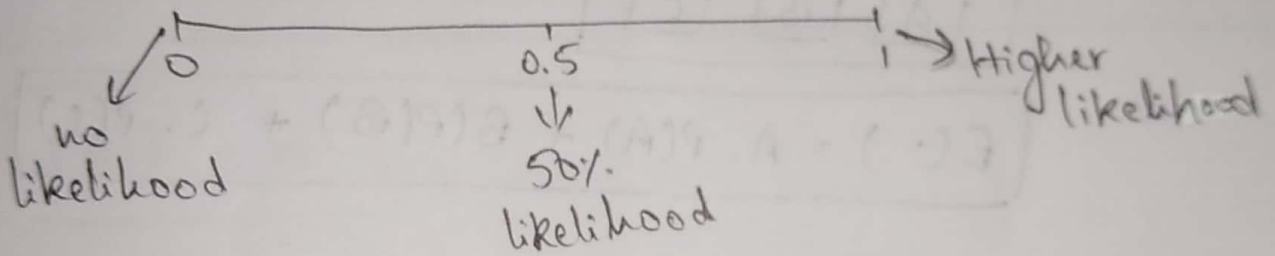
Subject

STATISTICS

1. PROBABILITY

Definition: The likelihood of an event occuring.

Value ranges from 0 to 1.
Higher value indicates higher probability



Eg: A is an event
P(A) is probability of an event to occur.

$$\therefore P(A) = \frac{\text{Favorable outcomes}}{\text{Sample Space}}$$

Note: Let A and B be events

$$P(A \text{ and } B) = P(A) \cdot P(B)$$

* Flipping a Coin
  ↳ There are 2 possible events - Heads or Tails
  ∴ Probability of getting Heads = $\frac{1}{2}$ = 0.5 = 50% chance
* Rolling a die
  ↳ There are 6 possible events - 1, 2, 3, 4, 5, 6
  ∴ Probability of getting 5 = $\frac{1}{6}$ = 0.167

2

Experimental Probabilities

$$P(A) = \frac{\text{Successful Trials}}{\text{All Trials}}$$

Expected Values

→ Categorical Outcomes

$$E(A) = P(A) * n$$

→ Numerical Outcomes

Sample Space

A	B	C

$$E(\cdot) = A \cdot P(A) + B \cdot P(B) + C \cdot P(C)$$

Probability Frequency Distribution

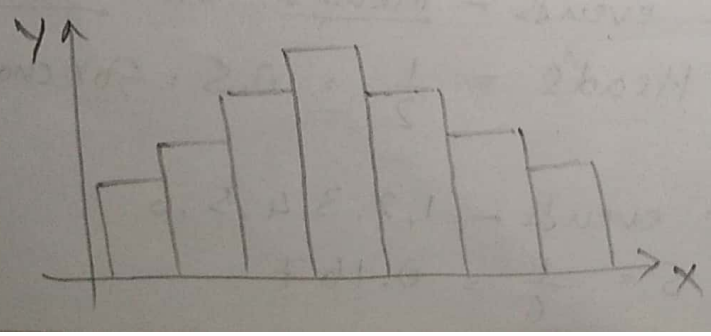
A collection of all the probabilities for the various outcomes is called Probability Frequency Distribution.

Sum	Frequency	Probability

↓

divide the frequency by the size of the sample space.

We can plot Probability Distribution Function with the help of histogram.



Events and their Complements

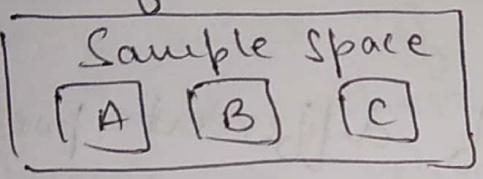
Complement : Everything the event is not.

Event A

complement of Event A = A^c

$A + A^c$ = Sample Space

Sum of All possible outcomes of an event = 1



$P(A) + P(B) + P(C) = 1$

$(A')' = A$

2. COMBINATORICS

Three integral parts
→ Permutation
→ Variation
→ Combination

Permutation

Definition: The number of different possible ways we can arrange a set of elements.

Note: n Factorial (n!)

$$n! = n \times (n-1) \times (n-2) \times \dots \times 1$$

* Negative number dont have factorial
* $0! = 1$

Note: Important Properties

* $n! = (n-1)! \times n$
* $(n+1)! = n! \times (n+1)$
* $(n+k)! = n! \times (n+1) \times (n+2) \times \dots \times (n+k)$
* $(n-k)! = \dfrac{n!}{(n-k+1) \times (n-k+2) \times \dots \times (n-k+k)}$
* $n > k$, $\dfrac{n!}{k!} = (k+1) \times (k+2) \times \dots \times n$

Important Note

* In permutation, order matters.
* Number of permutations (order matters) of n things taken r at a time

$$\boxed{{}^{n}P_{r} = \frac{n!}{(n-r)!}}$$

* Number of different permutations of n objects where there are n_1 repeated items, n_2 repeated items, ... n_k repeated items.

$$\boxed{\frac{n!}{n_1! \times n_2! \times \ldots \times n_k!}}$$

Variations

$$\boxed{{}^{n}V_{r} = n^{r}}$$., → with repetition

n- total number of elements available
r- number of elements we take at a time.

$$\boxed{{}^{n}V_{r} = \frac{n!}{(n-1)!}}$$., → without repetition

Combinations

Important Note:

* In combinations, order does not matter

* $$^{n}C_{r} = \frac{n!}{(n-r)!\, r!}$$

n - total number of elements available

r - total number of elements we take at a time

* $$^{n}C_{r} = {}^{n}C_{n-r}$$

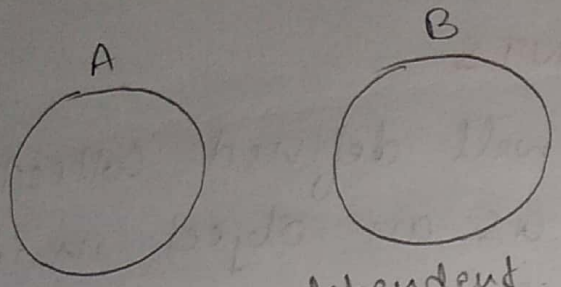
3. SETS & EVENTS

Definition: Set is a well defined collection of objects, considered as an object in its own right.

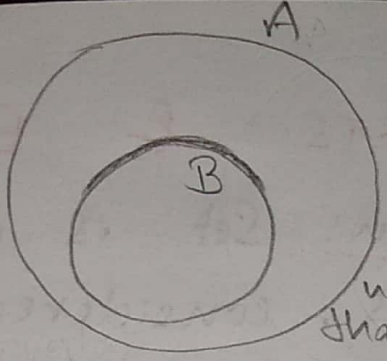
* A Set can be an Empty Set OR Non-Empty
* Empty Set can also be an Null Set. ($\emptyset$)
* Notations
  - → $x \in A$:- x is an element of Set A.
  - → $x \notin A$:- x is not an element of Set A.
  - → $\forall x \in A$:- for all x in A.
  - → $\forall x \in A : x$ is even :- for all x in A, such that, x is even.
  - → $A \subseteq B$:- Every element of A is also an element of B.

* Every Set contains 2 Sub Sets
  - → The set itself
  - → The Null Set.

8

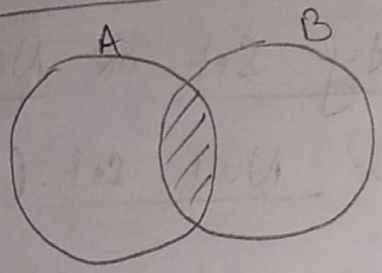


2 sets are independent



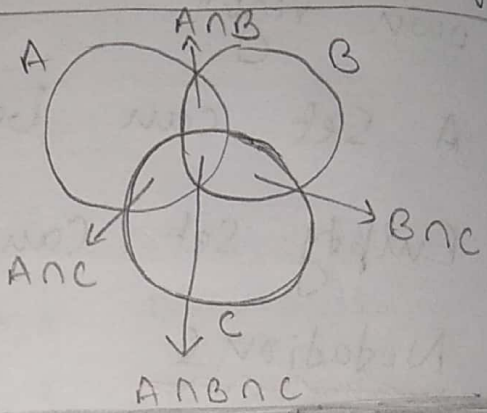
overlap

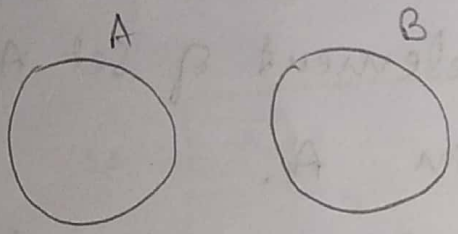
* If B occurs, it definitely satisfies A.
* If A occurs, no guarentee that it satisfies B



A & B happens simultaneous

Intersection

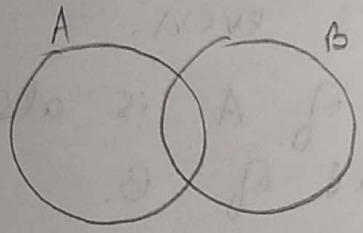




when 2 sets are independent

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

$(A \cap B) = \emptyset$ (Null Set)



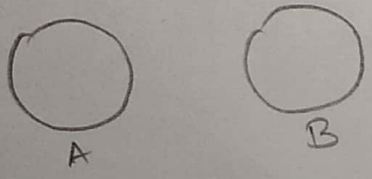
when 2 sets are dependent

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

Mutually Exclusive Sets

Sets which are not allowed to have any overlapping elements.

Their circles never intersect.



Dependent Events

The probability of getting A, if we are given that B has occured → $P(A|B)$

* If $P(A|B) = P(A)$, then A and B are independent events.

* $$P(A|B) = \frac{P(A \cap B)}{P(B)}$$ → ∴ only if $P(B) > 0$ The Conditional Probability Formula

If $P(B) = 0$, then $P(A|B)$ is not interpretable

* $$P(A|B) \neq P(B|A)$$

Law of Total Probability

Let A is union of finitely many events.

$$A = B_1 \cup B_2 \cup \dots \cup B_n$$

$$P(A) = P(A|B_1) \times P(B_1) + P(A|B_2) \times P(B_2) \dots$$

Additive Rule

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

Multiplication Rule

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

10

Bayes Rule/Theorem

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

$$P(A \wedge B) = P(B|A) \times P(A)$$

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

→ This equation is called Bayes Rule.

It is crucial because, it helps us to find a relationship between the different conditional probabilities of 2 events.

4. PROBABILITY - DISTRIBUTIONS

2 types of Distributions
① Discrete distributions
② Continuous distributions

1. Discrete Distribution : has finitely many distinct outcomes

→ * Uniform Distribution. $x \sim U(a,b)$ — a & b are range of distribution

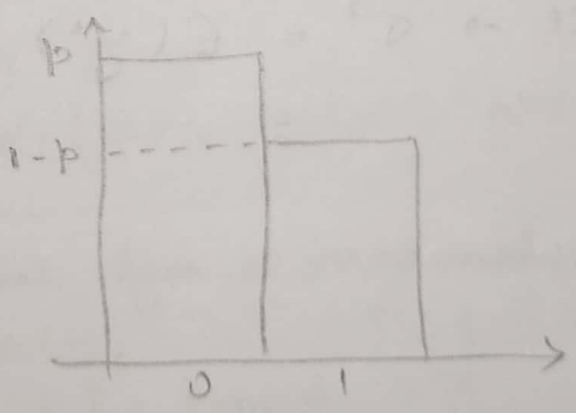
- All outcomes have equal probability
- No predictive power.
- mean and variance are uninterpretable.

→ * Bernoulli Distribution

- $x \sim Bern(p)$ — variable x → prefered outcome; Success p (known value)
- Deals with scenario where there is 1 trial and 2 outcomes.

Variance
$\sigma^2 = p(1-p)$

Standard deviation
$\sigma = \sqrt{p(1-p)}$



- Expected value is either p or $1-p$

→ * Binomial Distribution

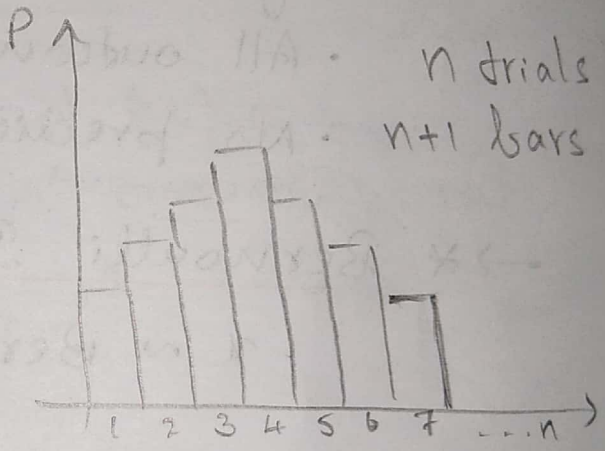
- used to know which of the 2 outcomes are more likely to occur.
- They are sequence of identical Bernoulli events

$x \sim B(n, p)$ — n : no of trials, p : Success

Eg: $x \sim B(10, 0.6)$

Probability Distribution Function

$$p(y) = \binom{n}{y} p^y (1-p)^{n-y}$$



$$E(y) = p \cdot n$$

Variance $\rightarrow \sigma^2 = E(y^2) - E(y)^2$

$= n \cdot p \cdot (1-p)$

$E( )$ - means Estimate Value

→ Poisson Distribution

. Deals with frequency with which an event occurs within an interval.

. $Y \sim Po(\lambda)$

$$P(Y) = \frac{\lambda^y \, e^{-\lambda}}{y!}$$

e - euler Constant ≈ 2.71828

λ - is a known value which tells mean number of occurences in the interval.

y - 0, 1, 2 ....

$$\text{mean}(\mu) = \text{Variance}(\sigma^2) = \text{Std dev}(\sigma)$$

2. Condinous Distribution

→ Normal Distribution

. $x \sim N(\mu, \sigma^2)$

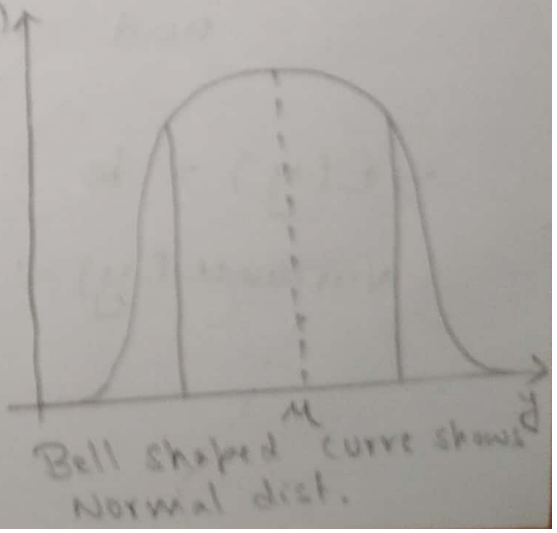
. values apart from mean are less likely to occur.

. Expected value for a normal distribution is its mean(μ)

. Variance can be found after we define the distribution.

otherwise use formula

$$\text{Var}(x) = E(x^2) - E(x)^2$$



Bell shaped curve shows Normal dist.

14

→ * Standard Normal Distribution

Standardization means making mean = 0 and variance = 1

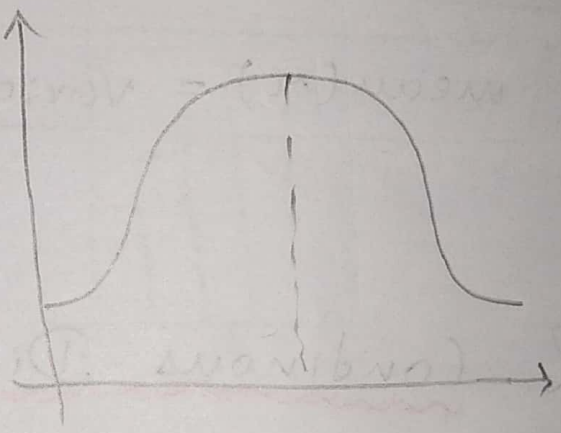
$$Z = \frac{x - \mu}{\sigma}$$

→ * Student's T Distribution

- $y = t(k)$ k - degrees of freedom

If $k > 2$

- $E(y) = \mu$
- $\text{Variance}(y) = \frac{s^2 \cdot k}{k-2}$

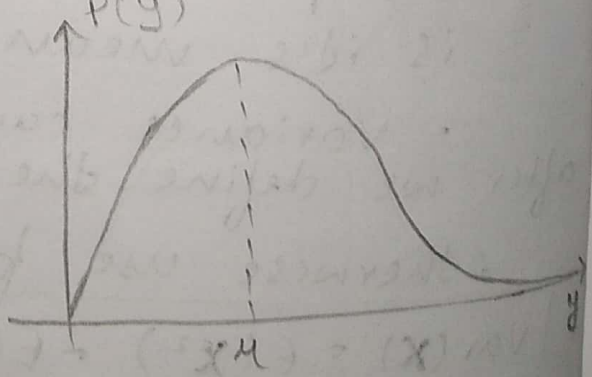


→ * Chi Squared Distribution

$y = \chi^2(k)$ k - degrees of freedom

χ → Greek letter chi

- Graph is asymetric and Right Skewed



- $E(y) = k$
- $\text{Variance}(y) = 2k$

→ * Exponential Distribution

$x \sim Exp(\lambda)$ λ - Scale

Expected value $= \frac{1}{\lambda}$

Variance $(x) = \frac{1}{\lambda^2}$

No table of known variables.

log(Exp dist.) = Normal distribution

→ * Logistic Distribution

$x \sim Logistic(\mu, s)$ μ → mean/ Location

s → scale parameters

Expected value $= \mu$

Variance $= \frac{s^2 \pi^2}{3}$

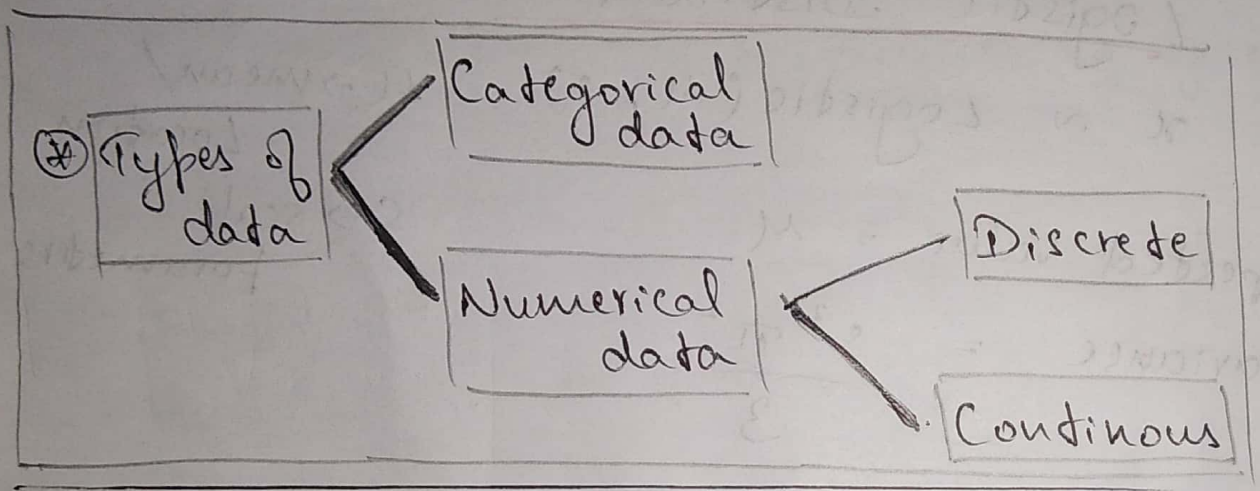
Definition of Distribution

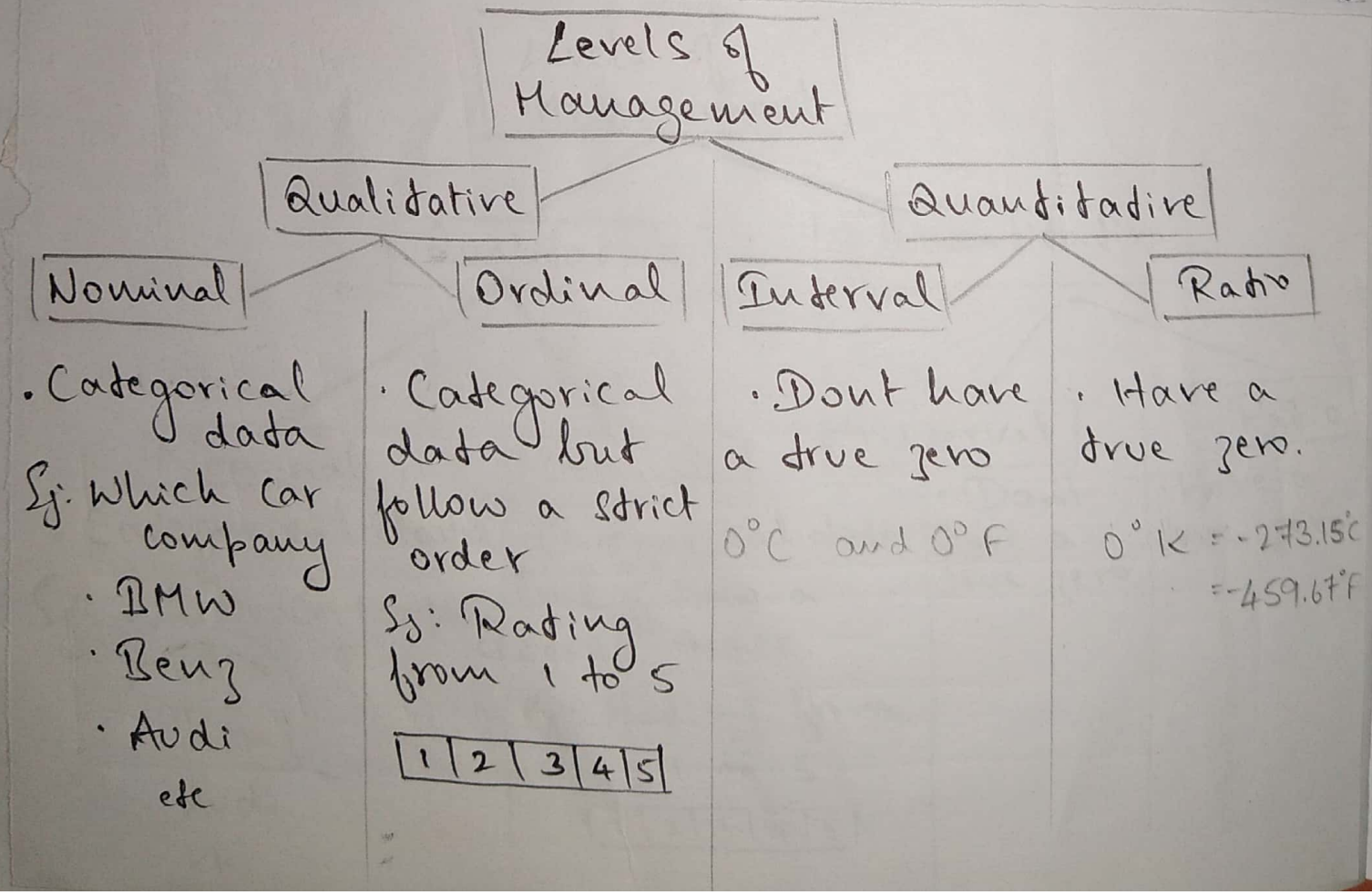
A distribution is a function that shows the possible values for a variable and how often they occur.

5. STATISTICS

Population data : It is the collection of all items of interest to our study. Denoted as N.

Sample data : It is a subset of population. Denoted as n.





Measures of Central Tendency

Central Tendency means tendency for the values of a random variable to cluster round its mean, median and mode

→ * **Mean**

- also known as simple average.
- Denoted as μ for population data. $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$
- Denoted as $\bar{x}$ for sample data. $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

→ * **Median**

- It is the middle number in an ordered dataset.

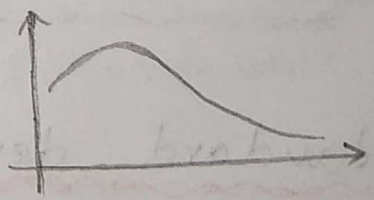
→ * **Mode**

- The number that appears most frequently in a set.

Skewness

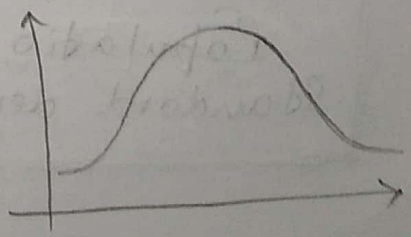
→ * **Right Skewed**

- mean > median
- outliers are towards the right



→ * **No Skewed**

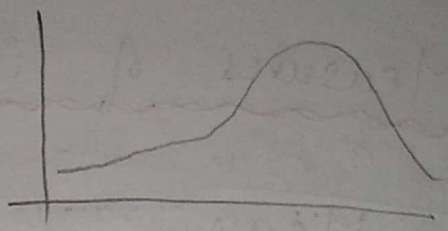
- mean = median = mode
- frequency of occurance is completely symmetrical



18

→ Left skewed

. mean < median



* Skewness tells us about where the data is situated.

Variance

Definition: Variance measures the dispersion of a set of data points around their mean.

Population Variance	$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$
Sample Variance	$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$

Standard deviation

Population Standard deviation	$\sigma = \sqrt{\text{population variance}}$
Sample Standard deviation	$s = \sqrt{\text{sample variance}}$

Coefficient of Variance (CV)

- Another name is relative Standard deviation.

Population formula	$C_v = \frac{\sigma}{\mu} = \frac{\text{Std dev}}{\text{mean}}$
Sample formula	$\hat{C_v} = \frac{s}{\bar{x}} = \frac{\text{Std dev}}{\text{mean}}$

Note:

Mean and Variance Relationship

$$\sigma^2 = E(y^2) - \mu^2$$

CoVariance:

When 2 variables are co-related, the measure of co-relation is called co-variance.

Used to determine the relationship between the movement of two asset prices.

When 2 stocks tend to move together, they are seen as having a positive covariance, when they move inversely, the covariance is negative.

Population formula	$\sigma_{xy} = \sum_{i=1}^{N} \frac{(x_i - \mu_x) * (y_i - \mu_y)}{N}$
Sample formula	$S_{xy} = \sum_{i=1}^{n} \frac{(x_i - \bar{x}) * (y_i - \bar{y})}{n-1}$

Corelation Coefficient

It tells us how much variability of one variable is explained by another.

$$Corr(x,y) = \frac{covariance(x,y)}{Std\ dev(x) * Std\ dev(y)}$$

$$\text{Population formula} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

$$\text{Sample formula} = \frac{S_{xy}}{S_x S_y}$$

If,

Corelation = 1 — positive corelation

Corelation = 0 — no corelation

Corelation = -1 — perfect negative corelation.

Note:

$$Corr(x,y) = Corr(y,x)$$

Corelation Coefficient

It tells us how much variability of one variable is explained by another.

$$Corr(x,y) = \frac{covariance(x,y)}{Std\ dev(x) * Std\ dev(y)}$$

$$\text{Population formula} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

$$\text{Sample formula} = \frac{S_{xy}}{S_x S_y}$$

If,

Corelation = 1 — positive corelation

Corelation = 0 — no corelation

Formula to find the corelation between 2 variables

range = $-1 \leq r \leq +1$

$r = -1$ → Negatively corelated

$r = +1$ → Positively corelated

$r = 0$ → No Corelation

$$r = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left[\sum_{i=1}^{n} x_i\right]\left[\sum_{i=1}^{n} y_i\right]}{n}}{\sqrt{\left[\sum_{i=1}^{n} x_i^2 - \frac{\left[\sum_{i=1}^{n} x_i\right]^2}{n}\right] * \left[\sum_{i=1}^{n} y_i^2 - \frac{\left[\sum_{i=1}^{n} y_i\right]^2}{n}\right]}}$$

Central Limit Theorem

It states that, no matter the distribution of the population, it's sampling distribution of it's mean will approximate a normal distribution.

Not only that, it's mean is same as the population mean.

$$\text{Variance} = \frac{\text{population variance}}{\text{Sample Size}} = \frac{\sigma^2}{n}$$

∴ Bigger the sample size, the variance will be lower and the statistical results will be more accurate.

Why is Central Limit Theorem important?

① Normal distribution has unmatched applicability in calculating confidence intervals and performing tests. Hence even when population data is not Normally distributed, we can perform tests and solve problems using CLT.

Standard Error

Definition: Standard Error is the standard deviation of the distribution formed by the sample means.

In other words, it is the standard deviation of the sampling distribution.

- How to calculate Standard Error

→ From CLT

$$\text{variance} = \frac{\sigma^2}{n}$$

$$\therefore \text{Std dev (of sampling dist.)} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

$$\therefore \boxed{\text{Standard Error} = \frac{\sigma}{\sqrt{n}}}$$

Standard Error decreases with sample size increases.

- It is important because it shows how well you approximated the true mean.

Estimators and Estimates

A specific value is called an Estimate.

There are 2 types of Estimates.
- Point Estimate
- Confidence Interval Estimate.

Point Estimate is located exactly in the middle of the confidence interval.

Confidence Intervals

Definition: A range of values so defined that there is a specified probability that the value of a parameter lies within it.

Confidence level: It is the level of confidence of the interval. It is denoted by $1-\alpha$

α (alpha)

$0 < \alpha < 1$

If we are 95% confident that the parameter is inside the interval, then $\alpha = 5\%$.

Similarly when Confidence level = 99%. then $\alpha = 1\%$.

α	1%	5%	10%
α	0.01	0.05	0.1

Formula for all Confidence intervals

$$\left[\left(\text{Point Estimate} - \text{Reliability Factor} * \text{Standard Error}\right), \left(\text{Point Estimate} + \text{Reliability Factor} * \text{Standard Error}\right)\right]$$

Point Estimate is nothing but the Sample mean

→ * Confidence Interval when Population variance is known

$$\left[\left(\bar{x} - Z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}\right), \left(\bar{x} + Z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}\right)\right]$$

σ - population Std dev

$\bar{x}$ - Sample mean

$\frac{\sigma}{\sqrt{n}}$ - Standard error

α - confidence level

$Z_{\alpha/2}$ is known as Critical Value

To Calculate $Z_{\alpha/2}$:

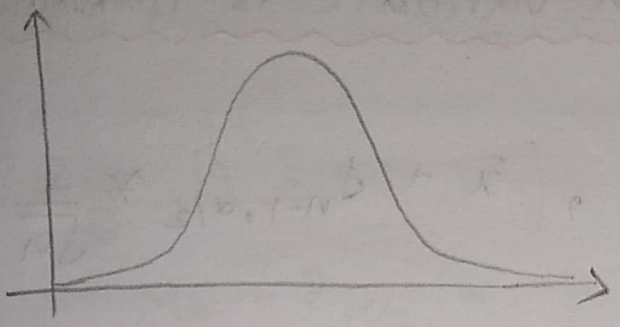
Step 1: Decide the confidence level and find α.

Step 2: Using z table, find the critical value

> Note: z table Summarizes the standard normal distribution critical values and the corresponding $1-\alpha$

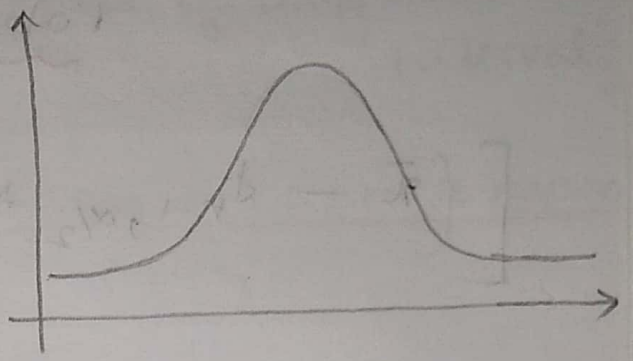
When population variance is known, we use z statistic and population Std dev to Calculate Standard error.

Normal distribution



- Have Thin tails
- Z-Statistic
- Population variance is known
- Big samples
- Sometimes even if variance is unknown.

Student's T distribution



- Have fat tails
- t-statistic
- Formula for t-statistic

$$t_{n-1,\alpha} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$n-1$ - degrees of freedom

α - Significance level

$\bar{x}$ - Sample mean

μ - population mean

$s/\sqrt{n}$ - Standard Error

s - Sample Std dev

- Student's T is just an approximation of the Normal dist.
- Population variance is unknown.
- Small Samples

→ * Confidence Interval when Population variance is unknown.

$$\left[\left(\bar{x} - t_{n-1,\alpha/2} * \frac{s}{\sqrt{n}}\right), \left(\bar{x} + t_{n-1,\alpha/2} * \frac{s}{\sqrt{n}}\right)\right]$$

s - Sample std dev

$\bar{x}$ - Sample mean

$\frac{s}{\sqrt{n}}$ - Standard error

α - confidence level

$n-1$ - degrees of freedom (Sample size - 1)

n - Sample size

Note: t table summarizes the probability for different degrees of freedom

when population variance is unknown, we use t-statistic and Sample Standard dev to calculate the error.

Margin of Error (ME)

Definition: They are the span of confidence interval.

Margin of Error when population variance is known

$$ME = z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}$$

Margin of Error when population variance is unknown

$$ME = t_{n-1,\alpha/2} * \frac{s}{\sqrt{n}}$$

The true population mean fall in the interval

$$\left[(\bar{x} - ME), (\bar{x} + ME) \right]$$

- Smaller ME means confidence interval will be Narrower (Small Confidence interval) [. . .]
- Bigger ME means confidence interval will be wider. [. . . . . .]

① the more observations there are in a sample, higher the chances of getting a good idea about the true mean of the entire population.

Confidence Intervals looking into 2 populations

There are 2 scenarios
- when 2 populations are dependent
- when 2 populations are independent

Dependent	Independent
• Before and After OR • Cause and Effect	• Population variance is known • when population variance is unknown but assumed to be equal • when population variance is unknown but assumed to be different.

→ * Confidence interval for difference of 2 means, dependent samples and population variance unknown

$$\left[\left(\bar{d} - t_{n-1,\alpha/2} * \frac{s_d}{\sqrt{n}}\right), \left(\bar{d} + t_{n-1,\alpha/2} * \frac{s_d}{\sqrt{n}}\right)\right]$$

$\bar{d}$ - Sample mean

s_d - Sample Std dev

→ * Confidence interval for difference of 2 means, independent samples and population variance known

Variance of the difference

$$\sigma_{diff}^2 = \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}$$

σ_x^2 - variance of population Sample 1

n_x - Size of population Sample 1

σ_y^2 - variance of population Sample 2

n_y - Size of population Sample 2

Confidence Interval

$$\left[\left((\bar{x} - \bar{y}) - z_{\alpha/2} * \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}\right), \left((\bar{x} - \bar{y}) + z_{\alpha/2} * \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}\right)\right]$$

$\bar{x}$ and $\bar{y}$ are Sample means

→ * Confidence intervals for differences of 2 means, independent samples and variances unknown but assumed to be equal.

Pooled variance formula

$$s_p^2 = \frac{(n_x - 1)\, s_x^2 + (n_y - 1)\, s_y^2}{n_x + n_y - 2}$$

n_x - Sample size of 1st set

n_y - Sample size of 2nd set

s_x^2 - variance of 1st set

s_y^2 - variance of 2nd set

Confidence interval

$$\left[(\bar{x} - \bar{y}) - t_{n_x + n_y - 2,\, \alpha/2} * \sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}} \right.,$$

$$\left. (\bar{x} - \bar{y}) + t_{n_x + n_y - 2,\, \alpha/2} * \sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}} \right]$$

→ * Confidence intervals for differences of 2 means, independent samples and variances unknown but assumed to be different

$$\left[\left((\bar{x}-\bar{y}) - t_{v,\alpha/2} * \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}\right),\ \left((\bar{x}-\bar{y}) + t_{v,\alpha/2} * \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}\right)\right]$$

v - degree of freedom

$$v = \frac{\left(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}\right)^2}{\frac{\left(\frac{s_x^2}{n_x}\right)^2}{(n_x-1)} + \frac{\left(\frac{s_y^2}{n_y}\right)^2}{(n_y-1)}}$$

6. HYPOTHESIS TESTING

Definition : Hypothesis is an idea that can be tested OR it is an assumption about a population parameter.

Null Hypothesis (H_0) : It is the one to be tested. We always try to reject H_0.

Alternate Hypothesis (H_1): It is the one that directly contradicts the null hypothesis.

The concept of null hypothesis is 'Innocent until proven guilty'
H_0 is true until rejected.

Significance level (Denoted as α (alpha))

It is the probability of rejecting the null hypothesis.

The typical values of α are 0.01, 0.05, 0.1

↓

Most commonly used.

Z - test

Formula

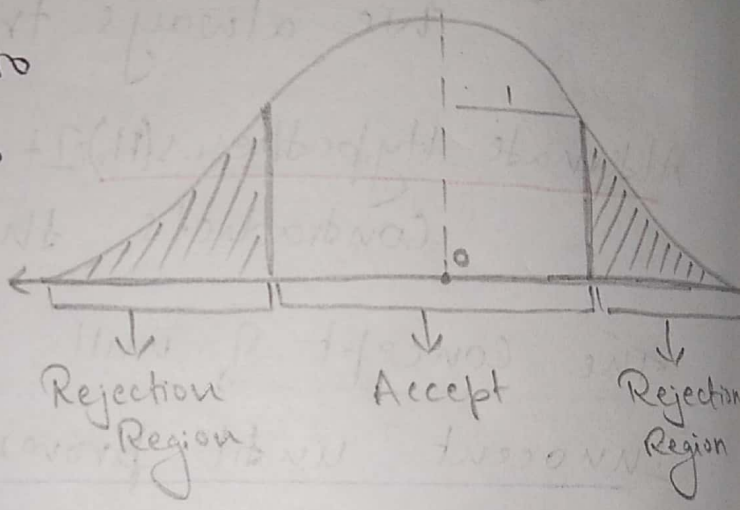
$$Z_T = \frac{\bar{x} - \mu}{S/\sqrt{n}}$$

$\bar{x}$ - Sample mean

μ - Hypothesized mean

$S/\sqrt{n}$ - Standard error

If Z_T value is close to zero or is equal to zero then accept the Null Hypothesis.

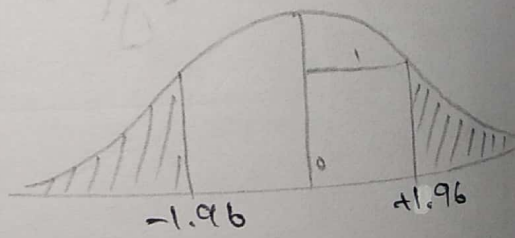


If Z_T value falls in the rejection region, then we have to reject the Null Hypothesis.

Area of the rejection region depends on the significance level.

when level of significance is 0.05, then we have 0.025 on either sides

$\alpha = 0.05 \quad \alpha/2 = 0.025$



check the values from Z-table, when $\alpha = 0.025$ $Z = 1.96$

∴ If the Z_T value is less than -1.96 or if it is greater than +1.96, then we reject the Null Hypothesis.

Few Confidence levels and their Critical Values

Confidence level	α	$\alpha/2$	Critical value
0.65	0.35	0.175	0.935
0.70	0.30	0.15	1.04
0.75	0.25	0.125	1.15
0.80	0.20	0.10	1.282
0.85	0.15	0.075	1.44
0.90	0.10	0.05	1.645
0.91	0.09	0.045	1.70
0.92	0.08	0.04	1.75
0.93	0.07	0.035	1.81
0.94	0.06	0.03	1.88
0.95	0.05	0.025	1.96
0.96	0.04	0.02	2.05
0.97	0.03	0.015	2.17
0.98	0.02	0.01	2.33
0.99	0.01	0.005	2.575

Errors in Hypothesis Testing

There are two types of errors:

1. Type I error
2. Type II error

Type I error

- When we reject a true Null Hypothesis
- Also called False Positive
- Probability of making this error is called Alpha (α)

Type II error

- When we accept a false Null Hypothesis.
- Also called False Negative
- Probability of making this error is Beta(β)
- Probability of rejecting a false Null Hypothesis is $1-\beta$

$\therefore$ $1-\beta$ is called Power of the Test.

The Truth

Ho	Ho is True	Ho is False
Accept	✓	Type II error (False Negative)
Reject	Type I error (False Positive)	✓

Decision Rule

Reject if Z_T is greater than the positive critical value (z).

$$(-)Z_T < -Z \quad <=> \quad (+)Z_T > +Z$$

Z_T - Standardized variable associated with the test called the Z-score.

Z - it is the one from the table and will be refered to as the critical value.

Note: Based on the problem, we have to Calculate the Z-score (Z_T) and T-Score (T_T)

	Population	Sample	Dependent Samples
Z_T	$\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	$\frac{\bar{x} - \mu_0}{s/\sqrt{n}}$	$\frac{\bar{d} - \mu_0}{\text{Std error}}$
T_T	$\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	$\frac{\bar{x} - \mu_0}{s/\sqrt{n}}$	$\frac{\bar{d} - \mu_0}{\text{Std error}}$

126

P - Value

It is the smallest level of significance at which we can still reject the Null Hypothesis.

If P-value is lower than the level of significance, then we reject the Null Hypothesis.

→ Calculate P-value for Z-Statistic

Step 1: Calculate Z_T.

For Eg: $Z_T = 2.31$

Step 2: Look in the Z table

For Eg:

0.01

↓

2.3 → number

number = 0.989

For one Sided Test

P-value = 1 - number = 1 - 0.989 = 0.011

For Two Sided Test

P-value = (1 - number) × 2 = 0.011 × 2 = 0.022

Note:

The closer to 0 your P-value is, the more significant is the result you have obtained

→* Calculate P-Value for T-Statistic

→ Test for mean when population variance unknown

$$Z_T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

μ_0 = Null Hypothesis value

$\bar{x}$ = Sample mean

$\frac{s}{\sqrt{n}}$ = Standard error

s - Sample Std dev

Z (Critical value) → based on level of significance.

$\alpha = 0.1 \;||\; 0.05 \;||\; 0.01$

One Sided OR Two Sided → upto the problem statement.

If 1 sided, then Z_α

If 2 sided, then $Z_{\alpha/2}$

→ Find critical value (z)

* If $(+) Z_T > +Z$, Reject Null Hypothesis.

(OR)

* Find P-Value.

If P-value < level of significance, Reject Null Hypothesis.

→ * Test for mean when population variance is known

$$Z_T = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

μ_0 = Null Hypothesis value

$\bar{x}$ = Sample mean

σ - population Std dev

$\sigma/\sqrt{n}$ = Standard error

Z (Critical value) → based on level of Significance
$\alpha = 0.1 \,||\, 0.05 \,||\, 0.01$

One Sided OR Two Sided → upto the problem statement.

If 1 sided, then Z_α
If 2 sided, then $Z_{\alpha/2}$
→ Find Critical value (z)

(*) If $(+)Z_T > +Z$, Reject Null Hypothesis

OR

(*) Find P value

If P.value < level of Significance, Reject Null Hypothesis

→ * Test for mean for Dependent Samples

Step 1 : Find the Difference

Step 2 : Calculate T score OR Z score based on the problem Statement.

Step 3 : Calculate either Critical value or P-value

Step 4 : Comparing the results, accept OR reject the Null Hypothesis.

→ * Test for mean, for independent samples when population variance is known.

Step 1 :

$$\text{Standard Error} = \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$$

Step 2:

$$Z_T \text{ OR } T_T = \frac{\bar{x} - \mu_0}{\text{Std error}} = \frac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$$

Step 3 : Calculate P-value

If P-value < level of Significance then Reject it.

Example:

	Eugg	Mgmt	Difference
Size	100	70	
Mean	58%	65%	-7.00%
Population Std dev	10%	6%	1.23% → Std error

Hypothesized difference = -4%.

$$Z_T = \frac{\bar{x} - \mu_0}{\text{Std error}} = \frac{(-7\%) - (-4\%)}{1.23\%} = -2.44$$

P-value = 0.015 [2 sided test with significance level 0.05]

0.015 < 0.05

∴ We reject Null Hypothesis.

→ * Test for the mean, for independent samples with unknown variances. but assumed to be equal.

Calculate pooled variance

$$S_p^2 = \frac{(n_x - 1)(S_x)^2 + (n_y - 1)(S_y)^2}{n_x + n_y - 2}$$

Calculate the pooled Std error

$$\text{Std error} = \sqrt{\frac{S_p^2}{n_x} + \frac{S_p^2}{n_y}}$$

Formula to calculate degrees of freedom

$$\text{degree of freedom} = \text{Combined Sample Size} - \text{Number of Variables}$$

Calculate T_T OR Z_T

Calculate P-value

∴ Decide whether to accept OR reject Null Hypothesis.

Note: Rule of Thumb

- Reject the Null Hypothesis when the T-Score (T_T) is bigger than 2.
- Generally, for Z and T-Scores, a value higher than 4 is extremely significant.

MACHINE LEARNING MODELS

1. REGRESSION

Regression is the measure of relationship between two variables.

Linear Regression

Linear Regression is a linear approximation of a casual relationship between two or more variables.

→ * **Simple Linear Regression Model.**

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

y - dependent variable. The value that we want to predict.

x_1 - independent variable.

β_1 - Quantifies the effect of independent variable on dependent variable

β_0 - Constant.

ε - error of estimation. [difference between the actual value and the predicted value.

Simple linear regression Equation

$$\hat{y} = b_0 + b_1 x_1$$

$\hat{y}$ (yhat) : whenever we have a hat symbol, it is an estimated / predicted value.

b_0 : constant OR Intercept

b_1 : quantifies the effect of independent variable on dependent variable.

x_1 : Sample data for independent variable

Slope of the line is the co-efficient of x_1.

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - \left(\frac{1}{n} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}$$

y_i - dependent variable

x_i - independent variable

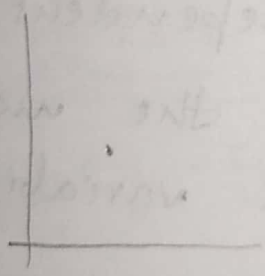
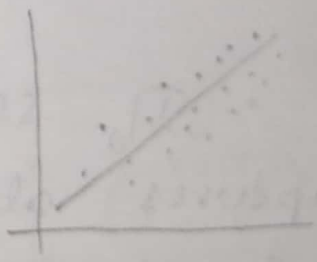
n - size

$$b_0 = \bar{y} - b_1 \bar{x}$$

$\bar{x}$ - mean of independent variable

$\bar{y}$ - mean of dependent variable

Co-relation vs Regression

Co-relation Analysis	Regression Analysis
① Measures degree of relationship between 2 variables.	① Its about how one variable affects other. or what changes it causes to the other.
② Doesn't capture causality. But the degree of inter relation between 2 variables.	② Based on causality. It shows no degree of connection Just cause and effect.
③ Co-relation of x & y is same as y and x $P(x,y) = P(y,x)$	③ One way. x on y and y on x yields different results.
④  Single point	④  Line Best fitting line.

Determinands of Good Regression

① Sum of Squares Total (SST)

$$= \sum_{i=1}^{n} (y_i - \bar{y})^2$$

y_i - observed dependent variable

$\bar{y}$ - mean of that variable

- It is a measure of the total variability of the dataset.
- Another notation is TSS (Total Sum of Squares)

② Sum of Squares Regression (SSR)

$$= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$\hat{y}_i$ - predicted value

$\bar{y}$ - mean of the dependent variable

- If SSR = SST, then the model captures all the observed variability and is perfect.
- Measures the explained variability by your line
- Another notation is ESS (Explained Sum of Squares)

③ Sum of Squares error (SSE)

$$= \sum_{i=1}^{n} e_i^2$$

e - error, which is the difference between the observed value and the predicted value.

- measures the unexplained variability by the regression.
- Smaller the error, better the estimation power of the regression.
- Another notation is RSS (Residual Sum of Squares)

→ * Relation between SST, SSR and SSE

$$SST = SSR + SSE$$

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} e_i^2$$

Total variability = Explained variability + Unexplained variability

- SSE and SSR are inversely proportional to each other. Each time we lower one other goes higher SSE↓ SSR↑

Note: OLS (ordinary Least Squares)

$S(b)$ is the OLS estimator of β for a Simple Linear Regression.

$$S(b) = \sum_{i=1}^{n} (y_i - x_i^T b)^2 = (y - Xb)^T (y - Xb)$$

we can determine the slope and intercept of the line using the above minimization problem.

b_0 is intercept.

b_1 is slope.

OLS is simple and powerful

Other methods to find the regression line

1. Generalized least Squares
2. Maximum likelihood estimation
3. Bayesian regression
4. Kernel regression
5. Gaussian process regression and many more

R-Squared Value (R^2)

$$R^2 = \frac{SSR}{SST} = \frac{\text{variability explained by regression}}{\text{total variability of the dataset}}$$

- Values range from 0 to 1
- If $R^2 = 0$, means your regression line explains none of the variability of the data.

If $R^2 = 1$, means your model explains the entire variability of the data.

-when do I know for sure my regression is good enough.

Ans: There is no definite answer for that.

0.2 ——— 0.9 is most common.

R-Squared measures the goodness of fit.

OR

R-Squared is a measure that measures how well your model fits your data.

... How well your data, fits your model is wrong.

52.

Multiple Linear Regression

Multi Linear Regression Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

Multi Linear Regression Equation

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots b_k x_k$$

$\hat{y}$ - value to be predicted

b_0 - intercept

$b_1, b_2, b_3 \ldots$ - co-efficients

$x_1, x_2, x_3 \ldots$ - independent variables.

$$b_0 = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2$$

- Multi Linear Regression is not about the best fitting line, but it is about the best fitting model.
- Multi Linear Regression has multiple independent variables. and is always better than Simple regressions.

- How to determine optimal number of variables to use.

Ans: Using Adjusted R Squared value.

Adjusted R Squared Value ($\bar{R}^2$)

- Adjusted R Squared value is always smaller than the R Squared value as it penalizes excessive use of variables.
- $\bar{R}^2$ increases if the new predictor (when new independent variable is added, it is called new predictor) enhances the model above what would be obtained by probability.
- Adjusted R Squared value gives the best estimate of the degree of relationship.

Formula for Adjusted R Square

$$\bar{R}^2 = 1 - (1 - R^2) * \frac{n-1}{n-p-1}$$

F Statistic : It is used for testing the overall significance of the model.

- More the value of F statistic, more significant the value is.

Null Hypothesis is (H_0) : $\beta_1 = \beta_2 = \beta_3 = \dots = \beta_k = 0$

Alternate (H_1) : Atleast one $\beta_i \neq 0$

Underfitting and Overfitting

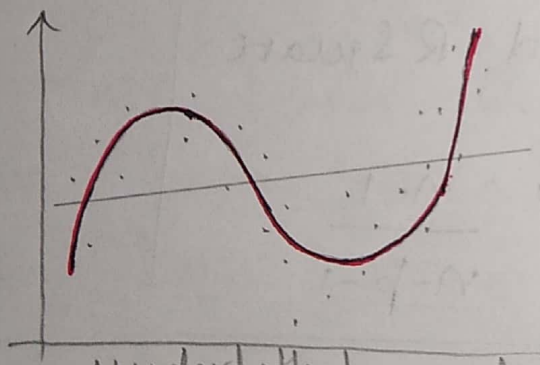
Underfitting

① The model has not captured the underlying logic of the data. It doesnt know what to do and therefore provides answer far from correct.

② No predictive power very low accuracy

③ Low training accuracy.

④



Underfitted model

Overfitting

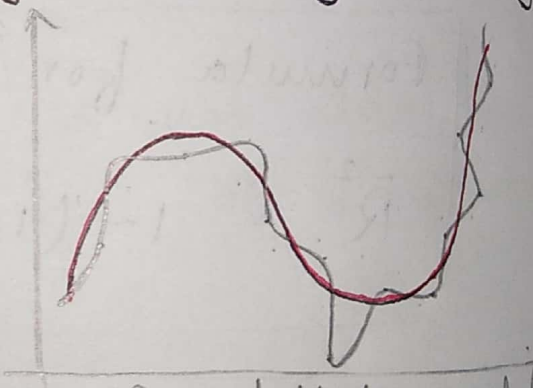
① Our training has focused on the particular training set so much, it has 'missed the point'.

Even a random noise is captured in an overfitted model.

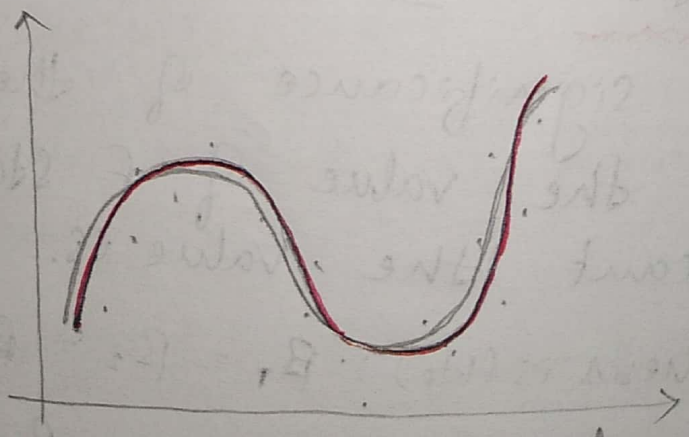
② No predictive power very low accuracy.

③ High training accuracy

④



Overfitted model



A Good Model

Variance Inflation Factor (VIF)

- If VIF = 1, then No Multicolinearity or No Corelation
- If 1 < VIF < 5, then moderately corelated and is accepted.
- If 5 < VIF, then highly corelated and is unacceptable.

VIF Formula

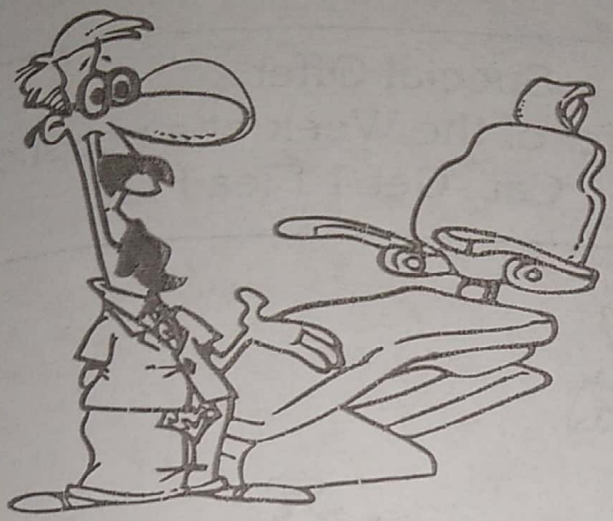
$$VIF = \frac{1}{1 - R^2}$$

R^2 - R-Squared value

Arisdotle once said

- Tell me, I'll forget
- Show me, I'll remember
- Involve me, I'll understand

Invention

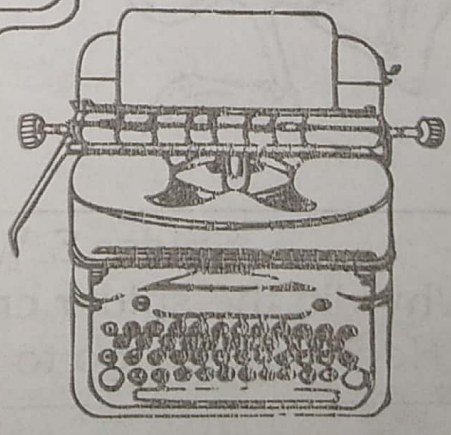


The electric chair was invented by a dentist

Isaac Newton invented the cat door

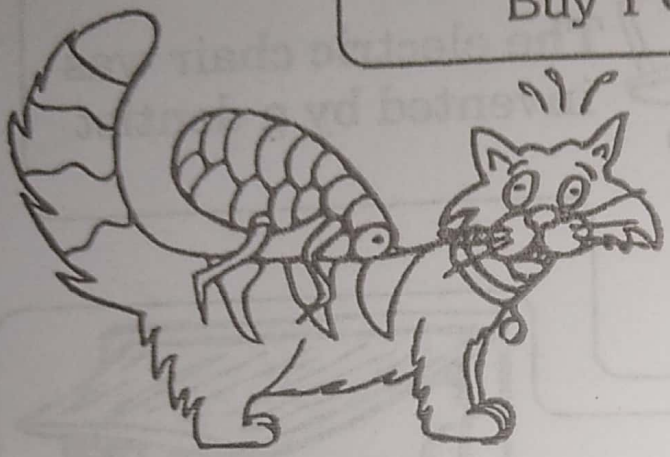
Volleyball was invented in 1895

The typewriter was invented in 1829



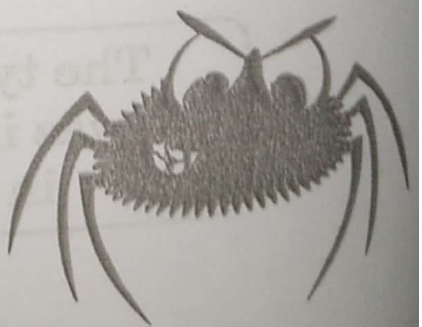
Jokes

Special Offer
Special Offer of the Week at a pet-shop
Buy 1 Cat -Get 1 Flea Free



Dog in Fire
What do you call a dog when it jumps into fire?
A hot dog

Cows & Bells
Why do cows wear the bells
Because their horns don't work!



Spider & Web
Why did the spider cross the road?
He wanted to go to his web-site

FINANCIAL CITY OF INDIA

Chennai

- Chennai is the capital of Tamil Nadu and a primary port city in south India.
- Its GDP of the city is 66 Billion USD.
- The flourishing trades in Chennai are automobiles, software services, medical tourism, hardware manufacturing, and financial services.
- Chennai is the second largest exporter of IT related services.
- Chennai is the biggest electronic exporter in India, accounting for about 50% of the total exports.
- Many automobile giants like Ford, Nissan, BMW etc have their headquarters in Chennai, hence the name "Detroit of India".

The only way

to do great work,

is to Love

What you

DO

Steve Jobs