BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
WORK INTEGRATED LEARNING PROGRAMMES
M .Tech (Data Science & Engineering)
MID SEM PREPARATION
I Semester, 2019-20
Course Handout

Course Title	**Introduction to Statistical Methods**
Course No(s)	

Course Description

This course will cover the statistical techniques which are very important in Data Science. It covers the models related to descriptive statistics, inferential statistics, predictive analytics and applied multivariate analytics.

Course Objectives

CO1	Understanding the data representation and analysis which is very important in Data Science
CO2	Understanding the predictive & inferential statistical models used in Data Science

Text Books

No	Author(s), Title, Edition, Publishing House
T1	Probability and Statistics for Engineering and Sciences,8th Edition, Jay L Devore, Cengage Learning
T2	Applied Logistic Regression, Hosmer and Lemeshow,3rd Edition, Wiley
T3	Introduction to Time Series and Forecasting, Second Edition, Peter J Brockwell, Richard A Davis, Springer.

Reference Books

No	Author(s), Title, Edition, Publishing House
R1	Miller and Freund's Probability and statistics for Engineers, 8th Edition, PHI
R2	Statistics for Business and Economics by Anderson, Sweeney and Wiliams, CENAGE learning

Modular Content Structure

1. Descriptive Statistics
    1.1. Data Visualisation
    1.2. Measures of Central Tendency
    1.3. Measures of Variability
2. Probability
    2.1 Probability – Introduction and Basics
    2.2 Conditional probability
    2.3 Bayes' theorem
3. Probability Distributions
    3.1. Random variables – Discrete & Continuous
    3.2. Probability Distributions
        3.2.1. Binomial Distribution
        3.2.2. Poisson Distribution
        3.2.3. Normal Distribution
4. Testing of Hypothesis
    4.1. Sampling & Estimation
    4.2. Type I, Type II errors
    4.3. Testing of Hypothesis – Mean – one and two mean
    4.4. Testing of hypothesis – Proportions – one and several Proportions
    4.5. ANOVA

Learning Outcomes:

No	Learning Outcomes
LO1	Clear understanding of the various statistical models to model the data
LO2	Drawing conclusions from the models selected to understand the data

Part B: Course Handout

Academic Term	I semester, 2019 – 20
Course Title	**Introduction to Statistical Methods**
Course No	

Course Contents

Contact Session 1: Module 1(Descriptive Statistics)

Contact Session	List of Topic Title	Reference
CS - 1	Descriptive Statistics: Data Visualisation, Measures of Central Tendency, Measures of Variability	T1:Chapter 1
HW	Problems on Descriptive Statistics	T1:Chapter 1
Lab		

Contact Session 2: Module 2 - Probability

Contact Session	List of Topic Title	Reference
CS - 2	Probability - Introduction and Basics, Conditional probability, Bayes' theorem	T1:Chapter 2
HW	Problems on probability	T1:Chapter 2
Lab		

Contact Session 3: Module 3 – Probability Distributions

Contact Session	List of Topic Title	Reference
CS - 3	Random Variables – Discrete & Continuous	T1:Chapter 3 & 4
HW	Problems on Random Variables	T1:Chapter 3 & 4
Lab		

Contact Session 4: Module 3 – Probability Distributions

Contact Session	List of Topic Title	Reference
CS - 4	Probability Distributions – Binomial, Poisson and Normal Distributions	T1:Chapter 3 & 4
HW	Problems on probability distributions	T1:Chapter 3 & 4
Lab		

Contact Session 5: Module 4 – Testing of Hypothesis

Contact Session	List of Topic Title	Reference
CS - 5	Sampling & Estimation	R1
HW	Problems on Interval Estimation	R1
Lab		

Contact Session 6: Module 4 – Testing of Hypothesis

Contact Session	List of Topic Title	Reference
CS - 6	Testing of Hypothesis - Type I & II errors, Mean and Proportions models (one mean, Two mean, One proportions and Several proportions with small and big samples wherever applicable)	T1:Chapter 7 ,8,9 & 10
HW	Problems on Testing of Hypothesis	T1:Chapters 7 to 10
Lab		

Contact Session 7: Module 4 – Testing of Hypothesis

Contact Session	List of Topic Title	Reference
CS - 7	Testing of Hypothesis - Problems discussion	T1:Chapter 7 ,8,9 & 10
HW	Problems on Testing of Hypothesis	T1:Chapter 7 ,8,9 & 10
Lab		

Set Theory

In this session we will learn about basic set theory, you may wonder why do we need set theory ?

For understanding probability and in fact to understand probability at a deeper level we need a solid foundation in set theory and that's why we review some concepts in set theory.

We wont go deep into set theory but we will review it.

I assume that you've learned set theory before

$P(B) = \frac{3}{4}$, $P(A \cap B \cap \bar{C}) = \frac{1}{3}$, $P(A \cap B \cap \bar{C}) = \frac{1}{3}$

find the value of $P(B \cap C)$.

2Q. A={1, 2, 3, 4}, B={4,5, 6, 7}, find P(AUB)=?

3Q. A={1, 3, 5}, B={3, 6}, find P(A∩B)=?

Vocabulary

- A *set* is any well defined collection of "objects."
- The *elements* of a set are the objects in a set.
- *Subsets* consists of elements from the given set.
- *Empty set/Null set* is the set that contains no elements.
- *Universal set* is the set of all possible elements.

Sets

- *Definition*. A *Set* is any well defined collection of "objects."
- *Definition*. The *elements* of a set are the objects in a set.
- *Notation*. Usually we denote sets with upper-case letters, elements with lower-case letters. The following notation is used to show set membership
- $x \in A$ means that *x* is a member of the set A
- $x \notin A$ means that *x* is not a member of the set A.

Ways of Describing Sets

- List the elements

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

Give a verbal description

- "A is the set of all integers from 1 to 6, inclusive"

Give a mathematical inclusion rule

$$A=\{\text{Integers } x \mid 1 \le x \le 6\}$$

Some Special Sets

- **The Null Set or Empty Set:** This is a set with no elements, often symbolized by

$$\varnothing \text{ or } \{\}$$

- **The Universal Set**: This is the set of all elements currently under consideration, and is often symbolized by

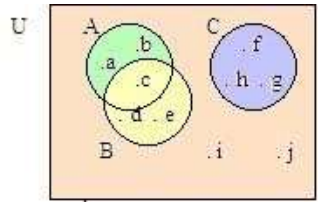
$$U$$

innovate achieve lead

Universal Sets

- The universal set is the set of all things pertinent to a given discussion and is designated by the symbol *U*

The **Universal set** is the set **of all elements under consideration in a given** U = {a, b, c, d, e, f, g, h, I, j}



Properties of the Union

Let A, B, C be three sets, then the following statements are true:

- $A \cup A = A$ (idempotent law)
- $A \cup B = B \cup A$ (commutative property)
- $A \cup (B \cup C) = (A \cup B) \cup C$ (associative property)
- $A \cup \varnothing = A$

- $A \cup A = \{a, b, c, d\} \cup \{a, b, c, d\} = \{a, b, c, d\}$ So, $A \cup A = A$.
- $A \cup \varnothing = \{a, b, c, d\} \cup \{\ \} = \{a, b, c, d\}$ So, $A \cup \varnothing = A$.
- $A \cup B = \{a, b, c, d\} \cup \{c, d, e, f\} = \{a, b, c, d, e, f\}$ $B \cup A = \{c, d, e, f\} \cup \{a, b, c, d\} = \{a, b, c, d, e, f\}$ So, $A \cup B = B \cup A$.
- $B \cup C = \{c, d, e, f\} \cup \{a, d, e, g\} = \{a, c, d, e, f, g\}$ $A \cup (B \cup C) = \{a, b, c, d\} \cup \{a, c, d, e, f, g\} = \{a, b, c, d, e, f, g\}$

  $A \cup B = \{a, b, c, d\} \cup \{c, d, e, f\} = \{a, b, c, d, e, f\}$

  $(A \cup B) \cup C = \{a, b, c, d, e, f\} \cup \{a, d, e, g\} = \{a, b, c, d, e, f, g\}$

  . We can easily see that $A \cup (B \cup C) = (A \cup B) \cup C = \{a, b, c, d, e, f, g\}$.

9

9

innovate achieve lead

Find the Subsets

- What are all the subsets of {3, 4, 5}

{} or Ø

{3}, {4}, {5}

{3,4}, {3,5}, {4,5}

{3,4,5}

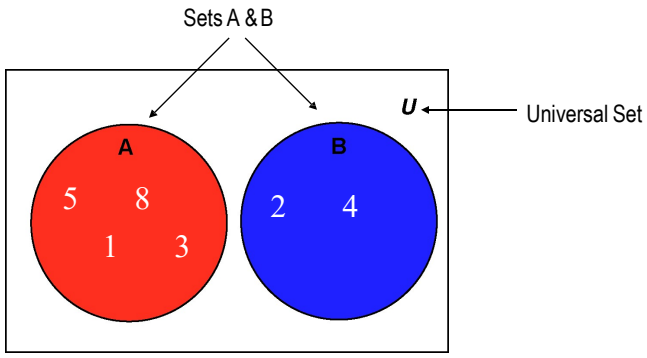
5

10

10

Venn Diagrams

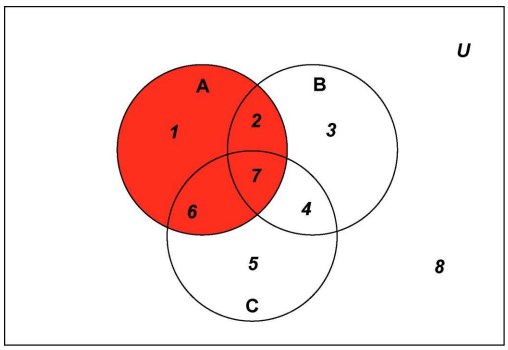
- Venn diagrams show relationships between sets and their elements



Venn Diagram Example

Set Definition

U = {1, 2, 3, 4, 5, 6, 7, 8}



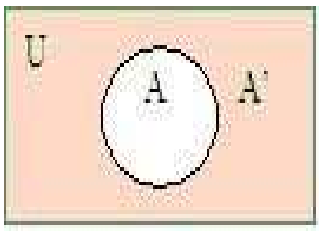
Set Complement

innovate achieve lead

~A or A′

- "*A* complement," or "not *A*" is the set of all elements not in *A*.

What the others have that you don't



13

Properties

Let U be a universal set & $A \subset U$.

- $A \cup A' = U$
- $A \cap A' = \varnothing$
- $U' = \varnothing$
- $(A')' = A$

$\forall\ \varnothing' = U$

14

More Practice:

- U = {1, 2, 3, 4, 5} is the universal set and
  A = {2, 3}. What is A'?

U = {a, b} is the universal set and
T = {a}. What is T'?

U = {+, -, x, ÷, =} is the universal set and
A = {÷, =}. What is A'?

Combining Sets – Set Union

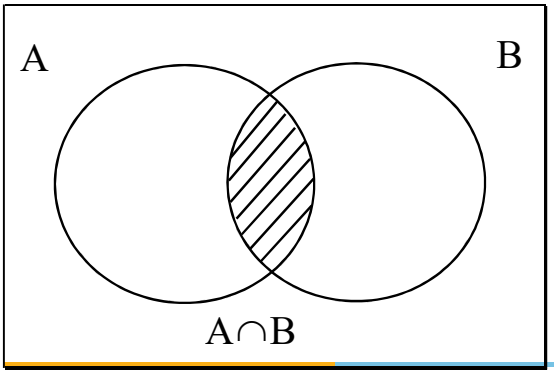
- $A \cup B$
- "*A* union *B*" is the set of all elements that are in *A*, or *B*, or both.
- This is similar to the logical "or" operator.

Combining Sets – Set Intersection

- $A \cap B$
- "*A* intersect *B*" is the set of all elements that are in *both A* and *B*.
- This is similar to the logical "and"

Venn Diagrams

- **Venn Diagrams** use topological areas to stand for sets. I've done this one for you.



Examples

$A = \{1,2,3\} \quad B = \{3,4,5,6\}$

- $A \cap B =$
- $A \cup B =$

Try it on your own!

- Let P = {b, d, f, g, h}, M = {a, b, c, d, e, f, g, h, i, j}, N = {c, k}

P ∪ M

P ∩ M

P ∪ N

N ∩ M

P ∩ N

Properties of the Intersection

Let A, B, & C be three sets, then the following statements are true:

- $A \cap A = A$
  (idempotent law)
- $A \cap B = B \cap A$
  (commutative property)
- $A \cap (B \cap C) = (A \cap B) \cap C$
  (associative property)
- $A \cap \varnothing = \varnothing$
  (identity law)
- $A \cap U = A$

21

Distributive Laws

For any three sets A, B, & C, the following statements are true:

∀ ∩ is distributive over ∪ from the left & the right. $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$

∀ ∪ is distributive over ∩ from the left & the right.

$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
$(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$

22

Some Special Sets

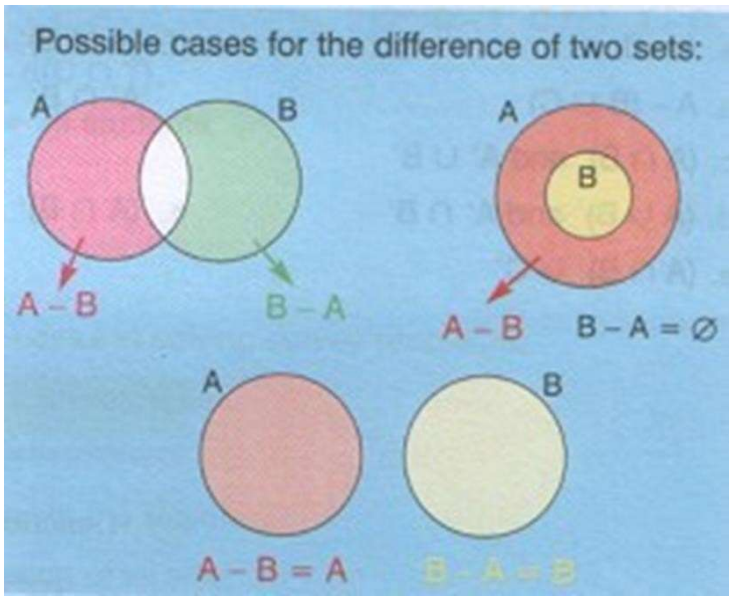
- The Null Set or Empty Set. This is a set with no elements, often symbolized by $\varnothing$
- The Universal Set. This is the set of all elements currently under consideration, and is often symbolized by Ω

Set Difference

$$A - B$$

- The set difference "A minus B" is the set of elements that are in A, with those that are in B subtracted out. Another way of putting it is, it is the set of elements that are in A, *and* not in B, so

$$A - B = A \cap \bar{B}$$



25

25

Properties

- $A - B \neq B - A$
- $A - \emptyset = A$
- $(A - B) \cap (B - A) = \emptyset$
- $A - A = \emptyset$
- $A - B = A \cap B'$
- If $A \subset B$ then $A - B = \emptyset$ De Morgan's

Laws

If A and B are two sets, then

$(A \cup B)' = A' \cap B'$ and $(A \cap B)' = A' \cup B'$.

26

26

Examples

$$U = \{1,2,3,4,5,6\}$$

$$A = \{1,2,3\} \qquad B = \{3,4,5,6\}$$

$$A \cap B = \{3\} \qquad A \cup B = \{1,2,3,4,5,6\}$$

$$B - A = \{4,5,6\} \qquad \overline{B} = \{1,2\}$$

Mutually Exclusive & Exhaustive Sets

- *Definition*. We say that a group of sets is *exhaustive* of another set if their union is equal to that set. For example, if we say that A and B are exhaustive with respect to C.

$$A \cup B = C$$

- *Definition*. We say that two sets A and B are *mutually exclusive* if $A \cap B = \varnothing$, that is, the sets have no elements in common.

SUBSET

A is a subset of B if every element in A is also in B.

eg $A = \{x, y, z\}$, $B = \{w, x, y, z\}$

$A \subseteq B$

Equal Sets: A and B are equal if

$A \subseteq B$, and $B \subseteq A$

$A = \{2, 4, 6\}$, $B = \{2, 4, 6\}$

Random Experiment:
A random experiment is an experiment whose outcome or result is not unique and therefore cannot be predicted with certainty. Ex: In tossing a coin, one is not sure whether a head or tail will occur.

Trail:
Each performance in a random experiment is called a trial. Ex: Tossing a coin first time is first trail, second time is second trial.

Outcome

The result of a trial in a random experiment is called an outcome. In coin tossing experiment getting head and tail are outcomes.

Sample Space:

The set of all possible outcomes of an experiment is called a sample space. Ex: In throwing a die, {1,2,3,4,5,6}will form the sample space.

Discrete Sample Space:- A sample space is said to be a discrete sample space if it has finitely many or a countable infinity of elements.

Ex:- a) A sample space consists of finite no of elements.
-2000 students.
b) The sample space consists of countable infinity of elements -- the whole set of natural nos

Continuous Sample Space:- If the elements of a sample space constitute a continuum – for example, all the points on a line, all the points on a line segment or all the points in plane – the sample space is said to be continuous .

Event:- Every non empty subset of a sample space of a random experiment is called an event. Ex: In throwing a die{1,2,3,4,5,6} -- Sample Space

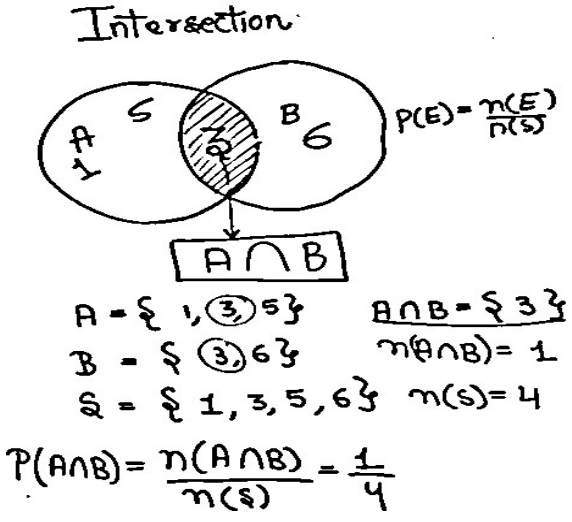
Mutually exclusive events:

- Events are said to be mutually exclusive if the happening of anyone of them prevents the happening of all the others i.e. if no two or more of them can happen simultaneously in the same trail.
- Ex: In tossing a coin the events Head turning up and tail turning up are mutually exclusive.
- The classical probability concept: If there are n equally likely possibilities, of which one must occur and S are regard as favorable, or as a "Success", then the probability of a "Success" is given by S/n.

Ex:-What is the probability of drawing an ace from a well shuffled deck of 52 Playing cards? Solution: There are S=4 aces among the n=52 cards, so we get s/n= 4/52 = 1/13.

31

31

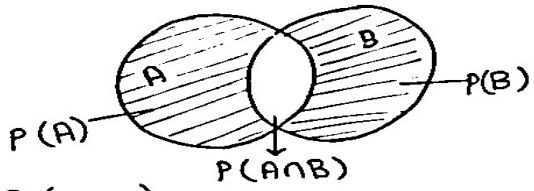


32

32

innovate achieve lead

Union $A \cup B$



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

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

$n(A) = 4$, $n(B) = 4$, $n(S) = 7$, $n(A \cap B) = 1$

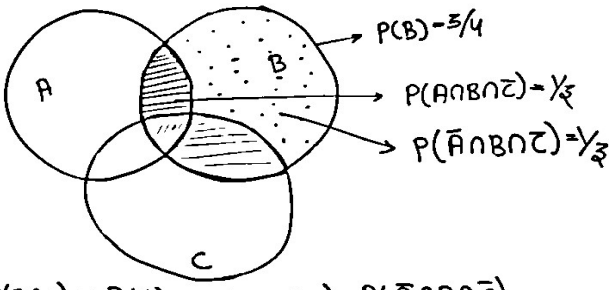
$S = \{1, 2, 3, 4, 5, 6, 7\}$, $A \cap B = \{4\}$

$P(A \cup B) = \frac{n(A)}{n(S)} + \frac{n(B)}{n(S)} - \frac{n(A \cap B)}{n(S)} = \frac{4}{7} + \frac{4}{7} - \frac{1}{7}$

innovate achieve lead

$P(B) = \frac{3}{4}$, $P(A \cap B \cap \bar{C}) = 1/3$, $P(\bar{A} \cap B \cap \bar{C}) = 1/3$

find the value of $P(B \cap C)$.



$P(B \cap C) = P(B) - P(A \cap B \cap \bar{C}) - P(\bar{A} \cap B \cap \bar{C})$

$= \frac{3}{4} - \frac{1}{3} - \frac{1}{3}$

Permutation:
The number of permutations of r objects taken from a set of "n" distinct object is
nPr = n! / (n - r)!

Problem
It repetitions are not allowed, how many 4 digit numbers can be formed from digits 1,2,3,4,5,6,7?
= 7! = 7! = 7x6x5x4 = 840
(7-4)! 3!

Combination
The number of ways in which "r" objects can be selected from a set of n distinct objects is
nCr = n!
r! (n-r)!

Ex: In how many ways 3 students can be selected from 15 students.

Ex: A bag contain 5 red balls, 8 blue balls and 11 white balls. Three balls are drawn together from the box. Find the probability that
One is red, one is blue and one is white.
Two whites and one red.
Three white.

Outcome	HH	HT	TH	TT
Value of X	2	1	1	0

35

35

Solution:
There are 24C3 = 2024 equally likely ways of choosing 3 of 24 balls, so n=2024.
The number of possible cases
5C1. 8C1.11C1 = 440.
Required probability = s/n = 440/2024 =
55/253
No. of possible cases = 11C2.5C1 = 275. Required probability = 275/2024 = 25/184
No. of possible cases = 11C3 = 165 Required probability = 165/2024 = 15/184

This assignment is known as the random variable.
By a random variable we mean a real number X associated with the outcome of a random experiment.

18

Eg: - Suppose two coins are tossed simultaneously then the sample space is
S = HH,HT,TH,TT. We will consider the
random variable ,which is the number of heads (0, 1, 2)

Outcome	HH	HT	TH	TT
Value of X	2	1	1	0

36

36

Thus to each outcome `S' , there corresponds a real number X(s). Note:- The real number is denoted by X(s) and it is defined for each s S. Random variable is also known as stochastic variable or variable.

The random variables are denoted by upper case letters such as X , Y , Z and the values assumed by them are denoted by lower case letters with subscripts as x1 , x2 , x3 ……..

A random variable X on a sample space " S " is a function from S to the set of real numbers R , which assign a real number X (s) to each outcome " s " of S .

Definition of Random variables
The function is given as X : S -> R
Example 1) If a coin is tossed , then sample space is S = { H , T }
Here we consider the random variable

$$X(s) = \begin{cases} 1 & \text{if } s = H \\ 0 & \text{if } s = T \end{cases}$$

2) If a random experiment consists of rolling a die and reading the number of points on the up turned face Sample space is { 1,2,3,4,5,6 }. Events which are considering is whether number of points is even or odd

37

Types of random variables

There are two types of random variables

1) Discrete random variable

2) Continuous random variable

Discrete random Variable (Def)
A random variable X is said to be discrete random variable if its set of all possible outcomes(sample space) is countable . (Finite or an un-ending sequence with as many elements as there in whole numbers)

Continuous Random Variables (Def)
A random variable X is said to be continuous random variable if the sample space contains infinite numbers equal to the number of points on a line segment

OR

A random variable X is said to be continuous if it can assume all possible values between certain limits.

Eg:- Weight, height

38

Definition of Antiderivative: A function F is called an antiderivative of the function f if for every x in the domain of f

F'(x) = f(x) so, dy = f(x) dx

Integration is denoted by an integral sign $\int$.

$$y = \int f(x)dx = F(x) + C$$

Integrand

Variable of Integration

Constant of Integration

F'(x) also = f(x) (first derivative)

39

39

Basic Integration Formulas

$$\int 0dx = C$$

$$\int kdx = kx + C$$

$$\int x^n dx = \frac{x^{n+1}}{n+1} + C$$

$$\int \cos x dx = \sin x + C$$

$$\int \sin x dx = -\cos x + C$$

$$\int \sec^2 x dx = \tan x + C$$

$$\int \sec x \tan x dx = \sec x + C$$

$$\int \csc^2 x dx = -\cot x + C$$

$$\int \csc x \cot x dx = -\csc x + C$$

40

40

Integrate

$$\int 3x\,dx \quad = \frac{3x^2}{2} + C$$

$$\int \frac{1}{x^3}dx = \int x^{-3}dx = \frac{x^{-2}}{-2} + C = -\frac{1}{2x^2} + C$$

$$\int \sqrt{x}\,dx = \int x^{1/2}dx = \frac{x^{3/2}}{3/2} + C = \frac{2x^{3/2}}{3} + C$$

$$\int 2\sin x\,dx = 2\int \sin x\,dx = 2(-\cos x) + C = -2\cos x + C$$

$$\int 1\,dx = x + C$$

$$\int (x+2)dx = \frac{x^2}{2} + 2x + C$$

$$\int \left(3x^4 - 5x^2 + x\right)dx = \frac{3x^5}{5} - \frac{5x^3}{3} + \frac{x^2}{2} + C$$

41

Find the general solution of the equation F'(x) = $\frac{1}{x^2}$ and find the particular solution given the point F(1) = 0.

$$F(x) = \int \frac{1}{x^2}\,dx = \int x^{-2}dx$$

$$= \frac{x^{-1}}{-1} + C = -\frac{1}{x} + C$$

$$\therefore\ y = -\frac{1}{x} + C$$

Now plug in (1,0) and solve for C.

0 = -1 + C

C = 1

Final answer. $y = -\frac{1}{x} + 1$

42

$$\int 12e^{4x}dx = 12\frac{e^{4x}}{4} + C$$

$$= 3e^{4x} + C$$

43

$$z = \int \left(6x^2 + \frac{3}{x} \right) dx$$

$$= \int 6x^2 dx + \int \frac{3}{x} dx$$

$$= \frac{6x^3}{3} + 3\ln x + C$$

$$= 2x^3 + 3\ln x + C$$

44

$$I = \int_0^1 8xe^{-2x}\,dx$$

$$= 8\frac{e^{-2x}}{(2)^2}\left[-2x-1\right]_0^1$$

$$= 2e^{-2}\left[-2(1)-1\right] - 2e^{-0}\left[0-1\right]$$

$$= -6e^{-2} + 2 = 1.188$$

45

45

When deciding what to choose for u, remember L I P E T.

L - logarithmic function

I - inverse trig function

P - polynomial function

E - exponential function

T - trigonometry function

This is usually the preference order in which you would want to choose u.

46

46

Generalized integral

$$\int uv\,dx = uv_1 - u'v_2 + u''v_3 - u'''v_4 \cdots$$

$\int uv\,dx = u \int v\,dx - \frac{du}{dx} \int v\,dx\,dx + \cdots$

$$\int_0^\infty x^2 e^{-2x}\,dx$$

$$\left[x^2\left(\frac{e^{-2x}}{-2}\right) - 2x\left(\frac{e^{-2x}}{4}\right) + 2\left(\frac{e^{-2x}}{-8}\right)\right]_0^\infty$$

47

47

Problem (14) : A continuous random variable has a probability density function.

$$f(x) = \begin{cases} 2ke^{-\lambda x}, & \text{for } x \geq 0, \lambda > 0 \\ 0, & \text{otherwise} \end{cases}$$

Determine (i) k (ii) Mean

(iii) Variance (iv) Standard Deviation.

48

48

Solution:

(i) $\int_0^\infty f(x)\,dx = 1$

$\Rightarrow \int_0^\infty k x e^{-\lambda x} = 1 \quad \Rightarrow k \int_0^\infty x e^{-\lambda x}\,dx = 1$

$\Rightarrow k \left[\left(x \left(\frac{e^{-\lambda x}}{-\lambda^2} \right)_0^\infty - 1 \left(\frac{e^{-\lambda x}}{\lambda^2} \right)_0^\infty \right] = 1$

$\Rightarrow k \left(-\frac{1}{\lambda^2} \left(e^{-\infty} - e^0 \right) \right) = 1$

$\Rightarrow k = \lambda^2$

(ii) Mean $= \int_0^\infty x \cdot k e^{-\lambda x}\,dx = k \int_0^\infty x^2 e^{-\lambda x}\,dx$

$= k \left[x^2 \left(\frac{e^{-\lambda x}}{-\lambda} \right)_0^\infty - 2x \left(\frac{e^{-\lambda x}}{\lambda^2} \right) + 2 \left(\frac{e^{-\lambda x}}{-\lambda^3} \right) \right]$

49

49

$= k \left(-\frac{2}{\lambda^3} \left(e^{-\infty} - e^{-0} \right) \right)$

$= \frac{+2\lambda^2}{\lambda^3} = +\frac{2}{\lambda}$

(iii) Variance $= \int_0^\infty x^3 k e^{-\lambda x}\,dx - (\text{mean})^2$

$= k \left[x^3 \left(\frac{e^{-\lambda x}}{-\lambda} \right)_0^\infty - 3x^2 \left(\frac{e^{-\lambda x}}{-\lambda^2} \right)_0^\infty + 6x \left(\frac{e^{-\lambda x}}{-\lambda^3} \right) - 6 \left(\frac{e^{-\lambda x}}{\lambda^4} \right)_0^\infty \right] - (\text{mean})^2$

$= k \left[-\frac{6}{\lambda^4} \left(e^{-\infty} - e^0 \right) \right] - \frac{4}{\lambda^2}$

$= \frac{6\lambda^2}{\lambda^4} - \frac{4}{\lambda^2} = \frac{6-4}{\lambda^2} = \frac{2}{\lambda^2}$

(iv) Standard deviation $= \sqrt{V(x)}$

$= \sqrt{2/\lambda^2} = \frac{\sqrt{2}}{\lambda}$

50

50

Learning objectives

innovate achieve lead

- Problems on Basic probability
- Conditional Probability
- Baye's theorem

Slide 51 of 23

51

Problem - 1

innovate achieve lead

In a group there are 3 men and 2 women. Three persons are selected at random from this group. Find the probability that one men two women or two men and one women are selected.

Slide 52 of 23

52

Solution -1

innovate achieve lead

Exhaustive number of cases are $^5c_3 = 10$ ways

Let A be the event of selecting one men and two women

Favourable cases for A are = $^3c_1 \,.^2c_2 = 3$ ways

$$P(A) = \frac{\text{Favourable cases for A}}{\text{Exhaustive number of cases}} = {}^3/_{10}$$

Solution -1

innovate achieve lead

Let B be the event of selecting two men and one women

Favourable cases for B are = $^3c_2 \,.^2c_1 = 6$ ways

$$P(B) = \frac{\text{Favourable cases for B}}{\text{Exhaustive number of cases}} = {}^6/_{10}$$

Required probability = **P (A U B)** = P(A) + P(B)

[A and B are disjoint events]

$= {}^3/_{10} + {}^6/_{10}$

$= {}^9/_{10}$

Problem - 2

If two dice are thrown, What is the probability that the sum is

(i) greater than 8

(ii) neither 7 nor 11.

Solution -2

When two dice are thrown the sample space contains 36 elements

S = { (1,1) , (1,2), (1,3), (1,4), (1,5), (1,6)

…..

…..

(6,1) , (6,2), (6,3), (6,4), (6,5), (6,6) }

Solution -2

innovate achieve lead

(i) Let A be the event that the sum on the two dice

P(A > 8) = P(A = 9) + P(A = 10) + P(A = 11) + P(A = 12)

sum is 9 = { (3,6), (6,3), (4,5), (5,4) }

$$P(A = 9) = \frac{\text{Favourable cases for A}}{\text{Exhaustive number of cases}} = {}^{4}/_{36}$$

sum is 10 = { (4,6), (6,4), (5,5) }

$$P(A = 10) = \frac{\text{Favourable cases for A}}{\text{Exhaustive number of cases}} = {}^{3}/_{36}$$

sum is 11 = { (5,6), (6,5) }

$P(A = 11) = {}^{2}/_{36}$

sum is 12 = { (6,6) }

$P(A) = {}^{1}/_{36}$

Solution -2

innovate achieve lead

P(A > 8) = P(A = 9) + P(A = 10) + P(A = 11) + P(A = 12)

$= {}^{4}/_{36} + {}^{3}/_{36} + {}^{2}/_{36} + {}^{1}/_{36}$

$= {}^{10}/_{36}$

(ii) **Let B denote the event of getting the sum of 7**

sum is 7 = { (1,6), (6,1), (2,5), (5,2), (3,4), (4,3) }

P(B) = $= {}^{6}/_{36} = {}^{1}/_{6}$

Let C denote the event of getting the sum of 11

sum is 11 = { (5,6), (6,5) }

P(C) = $= {}^{2}/_{36}$

Solution -2

Required probability = $P(\bar{B} \cap \bar{C}) = P[(B \cup C)^c]$

$= 1 - P(B \cup C)$

$= 1 - [P(B) + P(C)]$ [A and B are disjoint events]

$= 1 - {}^1/_6 - {}^1/_{18}$

$= {}^7/_9$

Slide 59 of 23

59

Problem - 3

The odds that person X speaks the truth are 3:2 and the odds that person Y Speaks the truth are 5:3. In what percentage of cases are they likely to contradict each on an identical point?

Slide 60 of 23

60

Solution -3

innovate achieve lead

Let A = X speaks the truth

$\bar{A}$ = X tell a lie

Let B = Y speaks the truth

$\bar{B}$= Y tell a lie

$P(A) = \frac{3}{3+2}$; $P(\bar{A}) = \frac{2}{3+2}$

$P(B) = \frac{5}{5+3}$; $P(\bar{B}) = \frac{3}{5+3}$

The event C that X and Y contradict each other on an-identical point.

That can happen in two ways

(i) X speaks the truth and Y tell a lie, i.e, $A \cap \bar{B}$

(ii) X tell a lie and Y speaks the truth, i.e, $\bar{A} \cap B$

Solution -3

innovate achieve lead

the events $(A \cap \bar{B})$ and $(\bar{A} \cap B)$ mutually exclusive events

P(C) = P$(A \cap \bar{B})$ + P$(\bar{A} \cap B)$

= $P(A) \cdot P(\bar{B}) + P(\bar{A}) \cdot P(B)$

[A and B are independent events]

= $\frac{3}{5} \cdot \frac{3}{8} + \frac{2}{5} \cdot \frac{5}{8}$

= 0.475

47.5 % of cases are they likely to contradict each on an identical point.

Problem - 4

innovate achieve lead

A certain shop repairs both audio and video components. Let A denote the event that the next component brought in for repair is an audio component, and let B be the event that the next component is a compact disc player (so the event B is contained in A). Suppose that $P(A) = 0.6$ and P$(B) = 0.05$.

What is $P(B/A)$?

Slide 63 of 23

63

Solution - 4

innovate achieve lead

Let A denote the event that the next component brought in for repair is an audio component

and

Let B be the event that the next component is a compact disc player

P(A) = 0.6 and P(B) = 0.05

and given that $B \subseteq A$

From sets operations $A \cap B = B$

then P$(A \cap B)$ = P (B) = 0.05

$P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{0.05}{0.6} = 0.0833$

Slide 64 of 23

64

Problem - 5

innovate achieve lead

Two Vendors X and y are willing to submit their bids for the supply of a large computer system. There is a 50-50 chance for the vender X.
If vender X does not submit the bid, then Y will get the order with a probability ¾.
If vender X submits a bid, then Y will get the order with a probability ¼.
What is the probability that Y will get the order?

Solution - 5

innovate achieve lead

Let A = Event that vender X submits the bid
$\bar{A}$ = Event that vender X not submits the bid
Let B = Event that vender Y will get the order

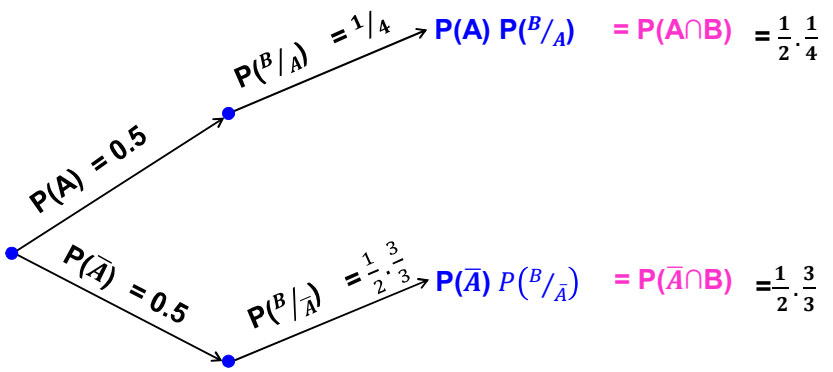
Given that $P(A) = {}^1/_2$ and $P(\bar{A}) = {}^1/_2$
$P({}^B/_{\bar{A}}) = {}^3/_4$ and $P({}^B/_A) = {}^1/_4$
By total probability

$$P(B) = P[(A \cap B) \cup (\bar{A} \cap B)] = P(A \cap B) + P(\bar{A} \cap B)$$ [Mutually disjoint events]

$$= P(A).P({}^B/_A) + P(\bar{A}).P({}^B/_{\bar{A}})$$

$$= \frac{1}{2}.\frac{1}{4} + \frac{1}{2}.\frac{3}{3} = \frac{1}{2}$$

Probability Tree Diagram



P(B) = $P(A).P(B/A) + P(\bar{A}).P(B/\bar{A})$ = P(A∩B) + P(Ā∩B) = $\frac{1}{2}$

Slide 67 of 23

67

Problem - 6

An individual has 3 different email accounts. Most of her messages, in fact 70%, come into account #1, whereas 20% come into account #2 and the remaining 10% into account #3. Of the messages into account #1, only 1% are spam, whereas the corresponding percentages for accounts #2 and #3 are 2% and 5%, respectively. What is the probability that a randomly selected message is spam?

Slide 68 of 23

68

Solution - 6

innovate achieve lead

A_i = { message is from account i} for i = 1,2,3

B = { message is span}

Given that

$P(A_1) = 0.7, P(A_2) = 0.2, P(A_3) = 0.1$

And $P(B/A_1) = 0.01, P(B/A_2) = 0.02, P(B/A_3) = 0.05$

Solution - 6

innovate achieve lead

By Total probability

$P(B) = P[(A_1 \cap B) \cup (A_2 \cap B) \cup (A_3 \cap B)]$

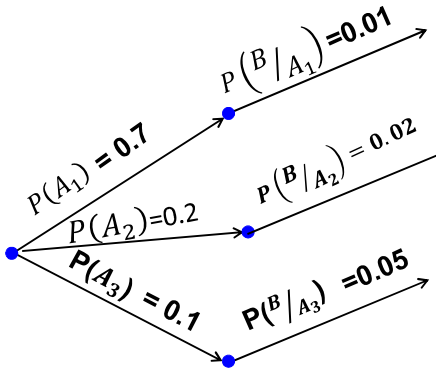
$= P[(A_1 \cap B)] + P[(A_2 \cap B)] + P[(A_3 \cap B)]$ [Mutually disjoint events]

$= P(A_1)P(B/A_1) + P(A_2)P(B/A_2) + P(A_3)P(B/A_3)$

= (0.7)(0.01)+(0.2)(0.02)+(0.1)(0.05)

= 0.016

Probability Tree Diagram



$P(A_1)\,P(B/A_1)$ = (0.7)(0.01)

$P(A_1)\,P(B/A_1)$ =(0.2)(0.02)

$P(A_1)\,P(B/A_1)$ = (0.1)(0.05)

P(B) $= P(A_1)P(B/A_1) + P(A_2)P(B/A_2) + (A_3)P(B/A_3)$ = 0.016

Problem - 7

A company that manufactures video camera as produces a basic model and a deluxe model. Over the years 40% of the cameras sold have been the basic model. Of those buying the basic model 30% purchase an extended warranty, whereas 50% of all deluxe purchasers do so.

What is the probability that a randomly selected customer has an extended warranty, how likely is it that he or she has a basic model?

Solution - 7

Let B be the event of a basic model.

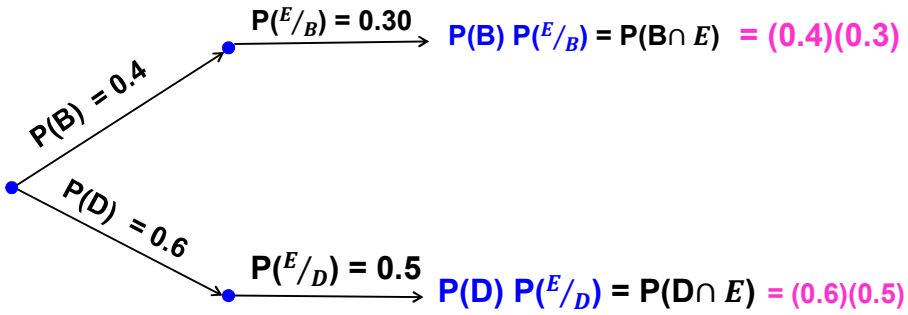
Let D be the event of a deluxe model.

Let E be event of extended warranty.

P(B) = 0.40, P(D)= 0.60,

P(E / B)= 0.30, and P(E / D)= 0.50

Probability Tree Diagram



$P(E) = P(B)\,P(E/B) + P(D)\,P(E/D) = 0.42$

Solution - 7

innovate achieve lead

the probability that a randomly selected customer has an extended warranty, how likely is it that he or she has a basic model is

$$P(B/E) = \frac{P(B \cap E)}{P(E)} = \frac{(0.4)(0.3)}{0.42} = 0.2857$$

Slide 75 of 23

75

Problem - 8

innovate achieve lead

In answering a question on a multiple choice test a student either knows the answer or he guesses.

Let p be the probability that he knows the answer and 1-p be the probability that he guesses. Assume that a student who guesses at the answer will be correct probability is 1/5, where 5 is the number of multiple-choice alternatives.

What is the probability that a student knows the answer to a question given that he answer it correctly?

Slide 76 of 23

76

Solution - 8

innovate achieve lead

Let A be the event the student knows the right answer .

$P(A) = p$

Let B be the event the student guesses the right answer .

$P(B) = 1 - p$

And

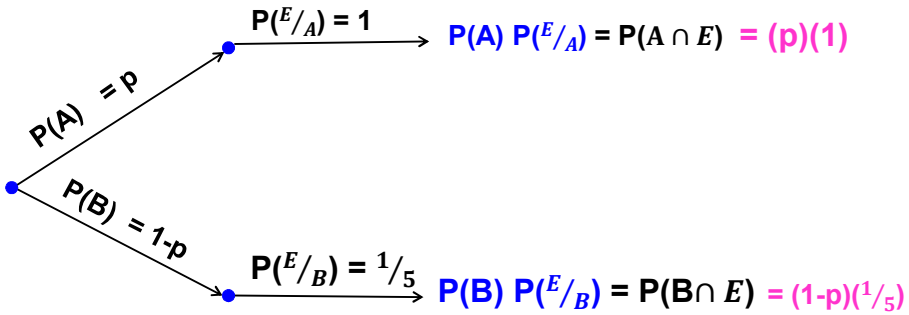
Let E be event the student gets the right answer .

$P(E/A) = 1$ [P(student gets the right answer given that he knows the right answer)]

$P(E/B) = 1/5$

Probability Tree Diagram

innovate achieve lead



P(A) $P(E/A)$ = P(A ∩ E) = (p)(1)

P(B) $P(E/B)$ = P(B∩ E) = (1-p)($1/5$)

P(E) = P(A) $P(E/A)$ + P(B) $P(E/B)$ = $(4p+1)/5$

Solution - 8

The probability that a student knows the answer to a question given that he answer it correctly

$$\mathbf{P}\left(A/E\right) = \frac{P(A \cap E)}{P(E)} = \frac{p}{(4p+1)/5} = \frac{5p}{(4p+1)}$$

Problem - 9

Verify that $P(X) = \frac{x+3}{25}$ $for\ x = 1,2,3,4,5$

Serve as Probability mass function?

Solution:

Given that $P(X) = \frac{x+3}{25}$ $for\ x = 1,2,3,4,5$

$P(X = 1) = \frac{1+3}{25} = \frac{4}{25}$,

$P(X = 2) = \frac{5}{25}$

$P(X = 3) = \frac{6}{25}$

$P(X = 4) = \frac{7}{25}$

$P(X = 5) = \frac{8}{25}$

$\sum_{x=1}^{5} P(X = x) = P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5)$

$= \frac{4}{25} + \frac{5}{25} + \frac{6}{25} + \frac{7}{25} + \frac{8}{25}$

$= \frac{30}{25} > 1$

Given P(X) is not a Probability mass function

[we know that Total probability is 1]

Lec 1 Overview of the course & Descriptive Statistics

"When you can measure what you are speaking about and express in numbers, you know something about it ;but when you cannot measure it, cannot express it in numbers, your knowledge is of meagre and unsatisfactory kind"

Lord Kelvin

"Statistical thinking will be one day as necessary for efficient citizenship as the ability to read and write"

H G Wells

Ex. while crossing a road – statistical thinking based on various parameters & samples

Knowingly or unknowingly we use stats in our daily life

Lies

damn lies

Statistics

Analytics

innovate achieve lead

- The term " Analytics"

- Disciplines
  - Statistics
  - Machine Learning
  - Biology
  - Kernel Methods

Analytics – Data available →suitable mathematical/stats model or technique – based on that take decision (estimation) or validate (hypothesis or assumption) or conclude

Taking decisions using models – systemic computation of data – Large amount of Data

Text book
by Y V K RAVI KUMAR . - Sunday, 22 July 2018, 10:14 AM

T1 mentioned in the Handout

or

The following book is preferred to go through the topics till testing of Hypothesis

Miller & Freund's Probability and Statistics for Engineers 8 Edition (English, Paperback, Richard A. Johnson, Irwin Miller, John Freund)

To start learning R
PRACTICAL DATA SCIENCE WITH R by NINNA ZUMEL & JOHN MOUNT is prefered

Data Science/Analytics – Is it the field for only **cs professionals**? **New field/Finding**?

We can't confine this D.A. to only one stream, we have to use concepts from different disciplines (Statistics – Fourier transform), Machine Learning (SVM, NN), Data Mining – Majority stats concepts

Earlier days – **Data Driven Decision** making – (D.A.) although it's not new approach – it has its base from long time (i.e. Doctor – not feeling fine – past data – past 2-3 days situation, food – based on data provided by patient – conclusion malaria, viral fever, typhoid etc.

Recent times – due to advancements – first reports – data collection then go to doctor, if we fail to notice/ignorance then information is not exact – doctor is not able to find root cause

Why Data Science became so popular – few factors

Instruments/ Tools available, Data Collection is not an issue (various sources), Storage is not an issue, Computation power has increased a lot

Because of these advancements – DDD making → Data Analytics field (center approach)

Importance of Data

- Importance : For any analytical exercise Data are key ingredients
  Replace intuition with data driven decisions

- For example consider the following cases:

  - Medical treatment
  - Industry
  - Power generation
  - Crime detection
  - Cognitive assessment

Industry – based on performance (past) we have to predict lifetime of a machine

Detection of thunderstorm– based on past data (Andhra Pradesh) – notify people

Based on Data available – built a model on top of it – decision taken

IRCTC – ticket confirmation probability – new feature added – very good application

Model _requirements

- Business relevance
- Statistical performance
- Interpretable
- Justifiability
- Operational efficiency
- Economic cost
- Regulation and legislation

What are the basic things/points we have to keep in mind?

Business relevance - At the end of day, model we built based on data we had - it should have relevance with our business or applicability in our industry otherwise it is of no use

Statistical Performance – some stats outcome – estimation or prediction or validation etc

Regulation and legislation – i.e. dealing with medical data – caste/religion data – regulation

Statistics?

Procedures for organising, summarizing, and interpreting information

- ✓ Standardized techniques used by scientists
- ✓ Vocabulary & symbols for communicating about data

Two main branches:

- *Descriptive statistics*
- *Inferential statistics*

Important thing in building a model is to consider parameters which are must – if we ignore those parameters our model is going to fail – we have many parameters which don't have any impact on outcome so want ignore those parameters

i.e. **Fifa 2018 world cup prediction** failure – France won, data → model → prediction

Conclusion – Data prediction is never accurate, some bad parameters are responsible

Data Building requires many things, human thinking, taking right data, no out layers – crucial step – **preprocessing of a data –** build a model then test the model

Tools are available but intuition of a person who is involved makes the difference

Ex. – River crossing – average depth of river – 2 m → whether average is giving true picture

Whether this tool is giving exact picture which we are looking for, which are my reqt? Next step →what are the parameters which are good/imp for our process → Good Model

Descriptive Statistics – economical → start building a model – all types together

Basic tools / concepts in analysis

innovate achieve lead

- Mean
- Median
- Mode
- Range
- Variance / Standard deviation
- Coefficient of variation
- Mean Deviation

Visualization techniques and basic tools used in performance analysis, Probability, Conditional and Bayes theorem, Probability distribution – Normal (Gaussian) Distribution

Sampling and Population concepts – Huge data – computation limitation – take samples – i.e. Opinion polls – they can't take whole population as sample – sampling is crucial here

Correlation and regression, very important to identify variables which are going to affect model and which are not – i.e. Dr checkup – cell phone number irrelevant to disease

Temperature of a person and age of patient's father – correlation coefficient – 0

If any relation and what extend – choose parameters, temperature > 90 → fever

After we have parameters - Building the model for prediction – Regression (linear & logistic)

Time series analysis – time based analysis – forecasting temperature based on last summer – we have past data – in 2017, 2016 – May month temperature → focus on season/month

Inventory – Diwali sale prediction → Consider season (Diwali sale) not whole year & predict

Multivariate analysis, PCA (Principle component analysis), Discrete component analysis

Mean – average – performance of students in a university

10k Data points – for conclusion – need not to go through all data points – take mean

Manufacturing Industry – Electricity – throughout a year - average consumption in a month – suppose it is 40-50 → expectation is 40 → 40 < 47, January – 75, July – 35 etc. Focus on January data which is more than expected

Is the Mean best tool available/very reliable? – **No** - Contradictions are there

Two players – Player A (Avg. last year – 45), Player B (Avg. last year – 48) → Player B is best? → No – May be out layer – Player A performance has not been good last year – that may be the case – we have to consider whole career

So whenever we have doubt or we can't rely on mean – can we re-verify using another tool?

Median –Middle number in sorted list – even number → two numbers – take average

Mode – Most occurred number in set – frequency of a word in sentences

Generally cars, doors & many things are designed for right handers (most number of people)

Shoes, Clothes etc. – some sizes are most frequent – that are mostly used

Languages/Tools – Python (50 People), Excel (10), R (30) – most occurring – Python

Whether mean can be negative? – Yes (Avg. Temperature in J&K is **-10 C** – No restriction

Two data sets – **same Mean** – i.e. **45** – difficult to pick between them – Variance is imp.

Variance – How data points are distributed taking mean as a base

Statistical graphs of data

A picture is worth a thousand words!

- Bar chart / graph
- Histograms
- Pie chart
- Pareto chart / diagram
- Frequency polygons
- Scatter plots
- Time series plot

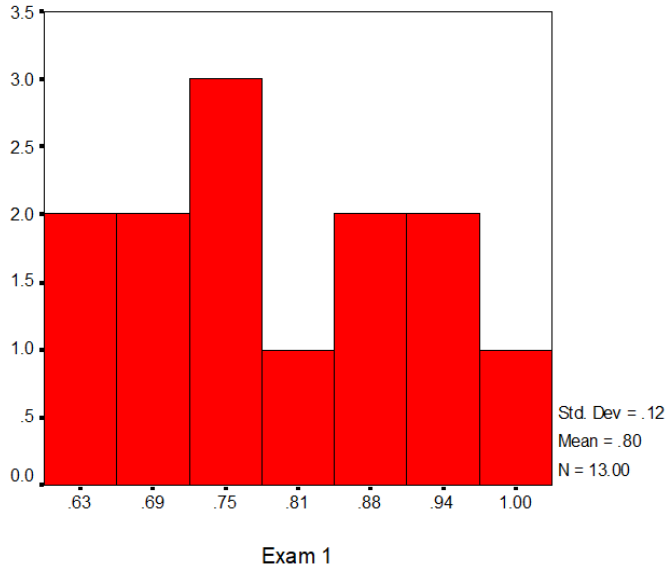
Bar Graphs

innovate achieve lead

Useful for showing two samples side-by-side

Histograms

- Univariate histograms



Histograms

X	f
6	1
5	2
4	2
3	4
2	2
1	1

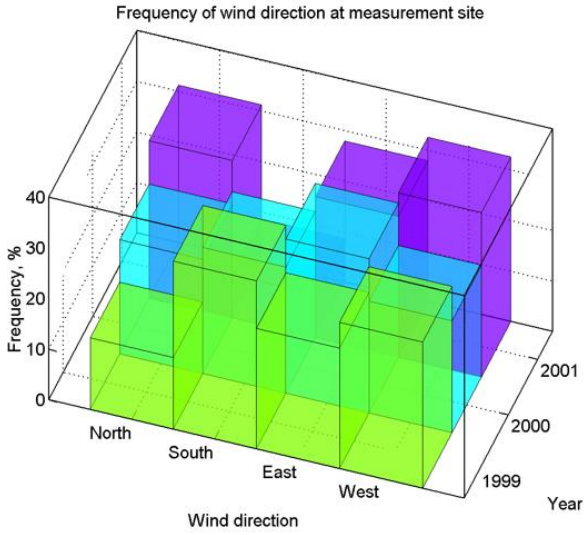
f on *y* axis (could also plot *p* or %)

X values (or midpoints of class intervals) on *x* axis

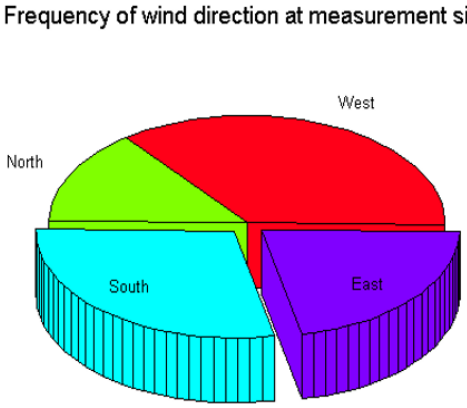
Plot each *f* with a bar, equal size, touching

No gaps between bars

Bivariate histogram



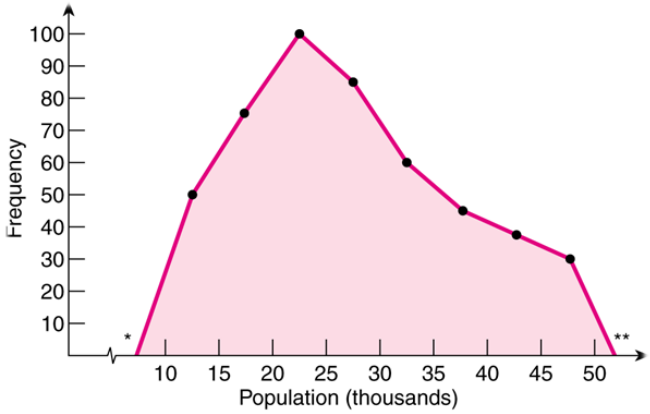
Graphing the data – Pie charts



Frequency Polygons

innovate achieve lead

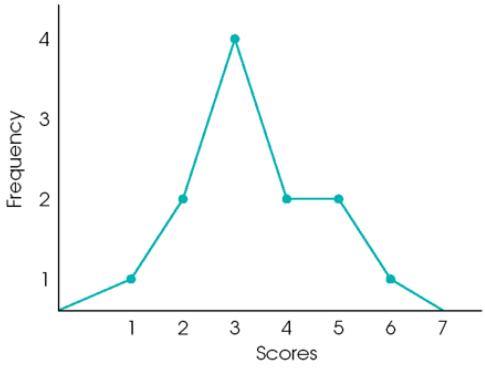
- Frequency Polygons
  - Depicts information from a frequency table or a grouped frequency table as a **line graph**



* 4 cities had populations of less than 10,000.
** 5 cities had populations of 50,000 or greater.

Frequency Polygon

innovate achieve lead



X	f
6	1
5	2
4	2
3	4
2	2
1	1

A smoothed out histogram

Make a point representing *f* of each value

Connect dots

Anchor line on x axis

Useful for comparing distributions in two samples (in this case, plot *p* rather than *f*)

Lec 2 Descriptive Statistics

Today…..

- Recall the past for a while_ Simple tools
- Visualization of data
- Basics of probability
- Discussion & Problems on probability
- Conditional probability

Visualization

- Summary gives an idea about the data

summary(income)

Min	1st QU.	Median	Mean	3rd Qu	Max
- 7.8	12.5	32.0	52.03	67.2	585

- Visualization – why

First we have data set → prepare summary of a data

Is this sufficient to proceed further (without looking into whole data set)? – Yes

Visualizing data – converting data into graphics – we can understand data very clearly

Cricket – 1st team batting/bowling performance – graph → how 2nd team is doing

So visualizing data helps us to **understand the data/process** – sometimes we can take call

FM - Budget speech -- Rs in → Rs Out – complete representation – easy to understand

Data Visualisation

- Line chart
- Bar chart
- Histogram
- Pie chart
- Scatter plot
- Box plot

Main challenge in visualization is which one is opt for the outcome we are looking for.

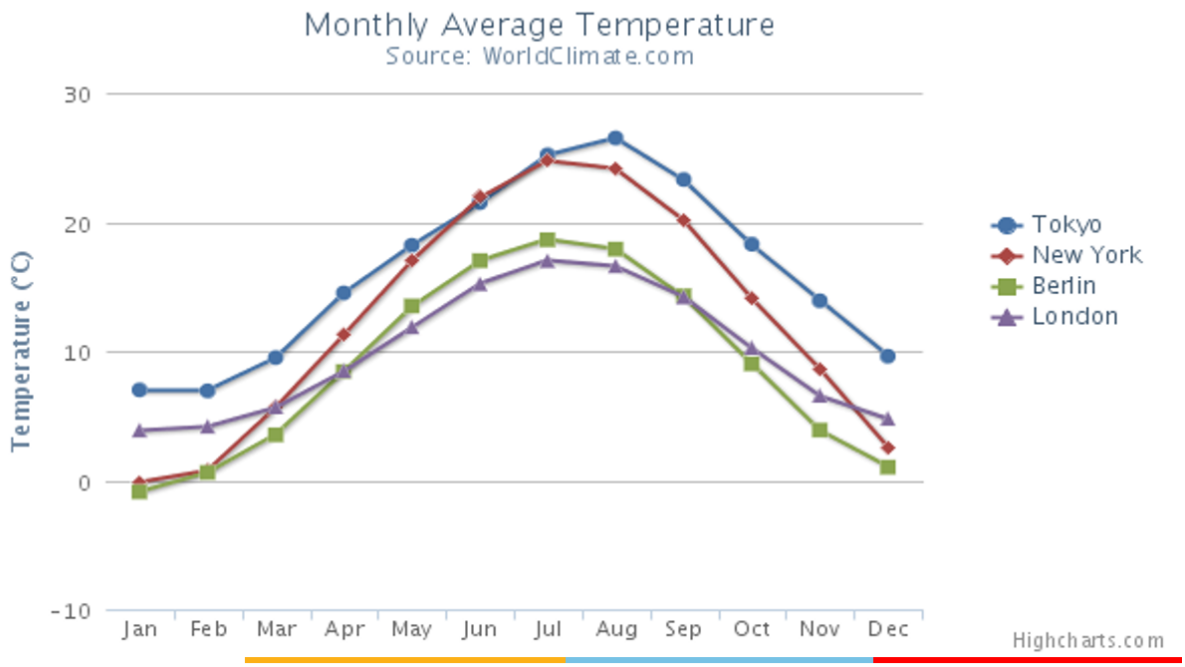
Share market graphs – how dynamically data is moving, detecting anomalies etc.

Instead of looking into whole data set – just looking at representation we can understand

Traffic in Google Maps – visualization – we can plan our journey easily, NSE marker etc.

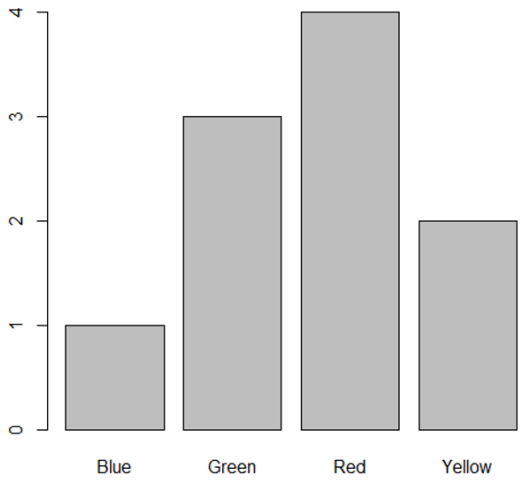
Resource utilization – we use line charts generally

Line Chart



Behind this chart we have csv/excel file contacting data (temperature in this case) – we can have data like which city is hottest/coldest during particular month? – Easily answerable

Bar Chart

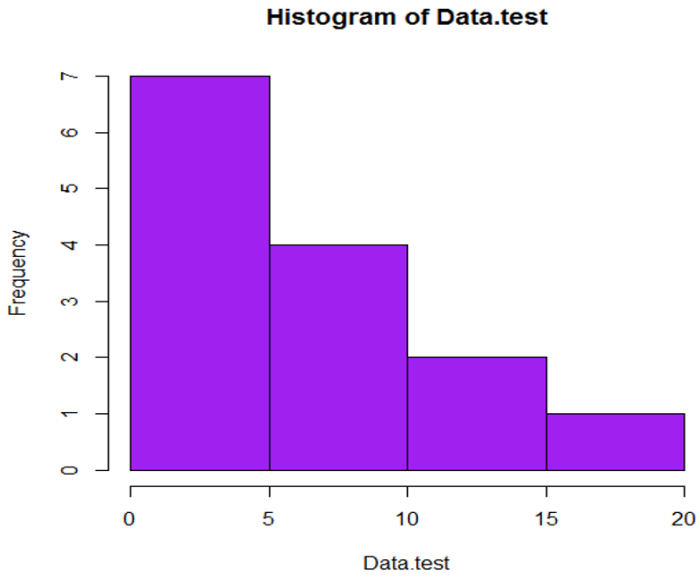


Weather patterns, profits made during quarters – we can observe growth - Grouping

Portfolio managements, number of users region wise, product wise, discrete data etc.

Histograms

innovate achieve lead



Bar chart – they can be categorical values (not numerical values always) – independent

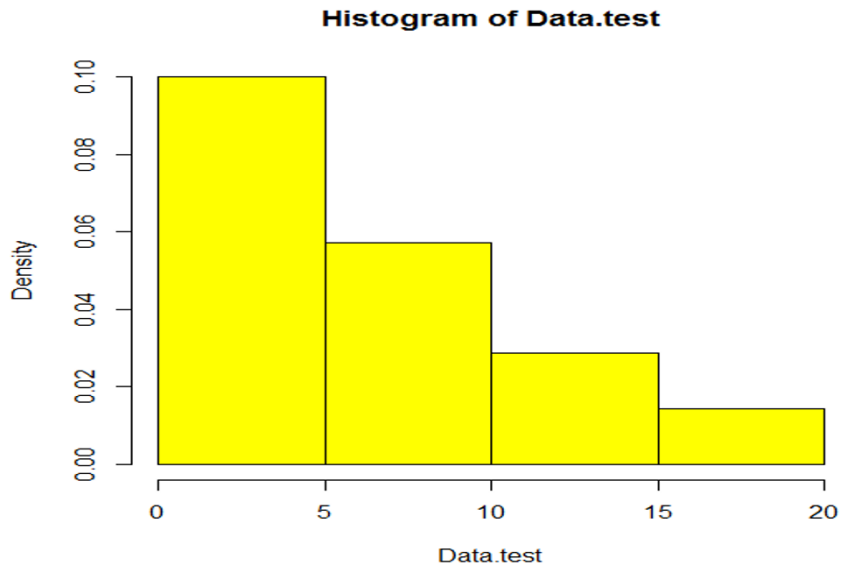
Histograms – numerical – continuous values – in groups → Class intervals – freq. distribution

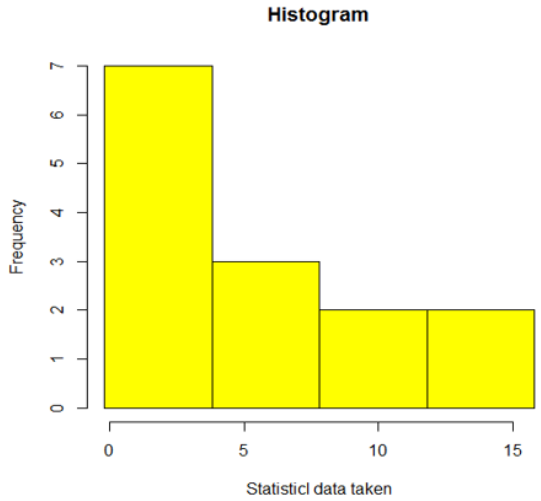
Histogram – Data divided into beans – consecutive ranges/groups

Histogram - how data is distributed in terms of beans or intervals

Histograms

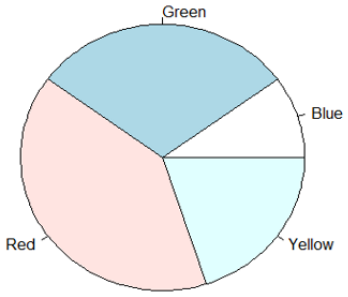
innovate achieve lead





Pie charts

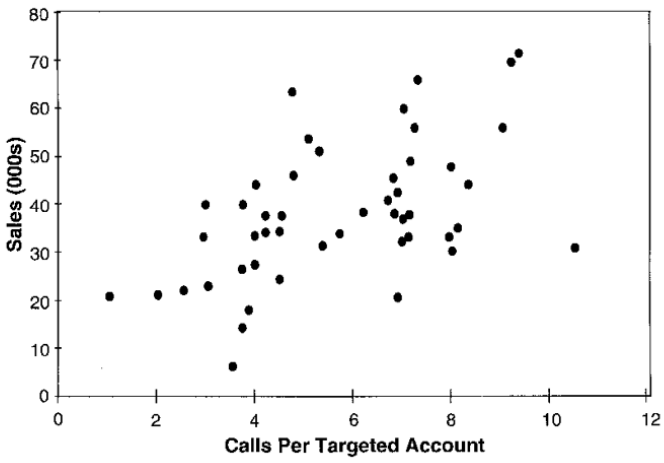
innovate achieve lead



Issues handling, Revenue made by each companies, Major players in mobile market and their shares, financial investment we do – Mutual funds – sectors wise

Scatter Plot

innovate achieve lead



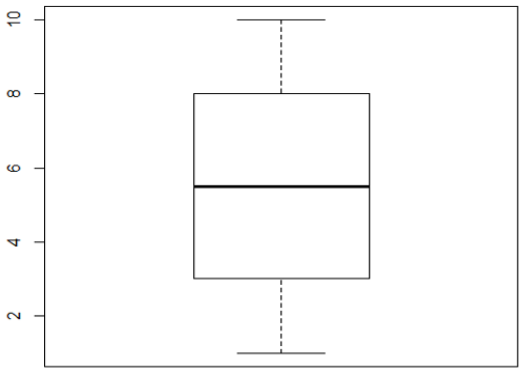
All the data points can be seen as it is and how they are distributed – density of point

Regression – scatter points is helpful to determine which type of regression is suitable

Polling visuals, drawing decision boundaries (ML), Data cleaning – out layers – build model

innovate achieve lead

Box plot



innovate achieve lead

To conclude _ Visualization

- Visualization gives a sense of data distribution and relationship among variables
- Visualization is an iterative process and helps answer questions about the data. Time spent is not wasted during the modelling process and helps to find the optimal model to fit the data

Gives clarity to us before building model – to understand data pattern

innovate achieve lead

!!!!

A famous statistician would never travel by airplane, because she had studied air travel and estimated the probability of there being a bomb on any given flight was 1 in a million, and she was not prepared to accept these odds.

One day a colleague met her at a conference far from home.

"How did you get here, by train?"

"No, I flew"

"What about the possibility of a bomb?"

"Well, I began thinking that if the odds of one bomb are 1:million, then the odds of TWO bombs are (1/1,000,000) x (1/1,000,000) = 10^{-12}. This is a very, very small probability, which I can accept. So, now I bring my own bomb along!"

Random Experiment

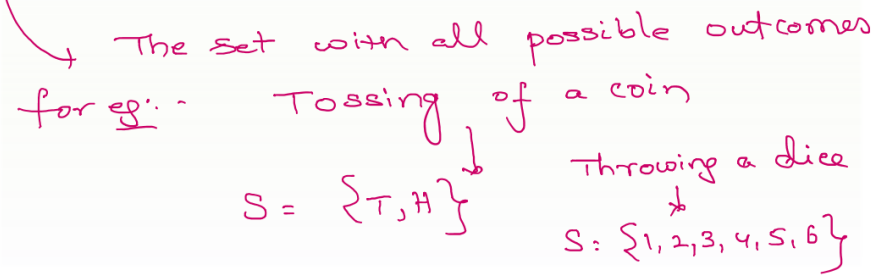
Term "**random experiment**" is used to describe any action whose outcome is not known in advance. Here are some examples of experiments dealing with statistical data:

- Tossing a coin
- Counting how many times a certain word or a combination of words appears in the text of the "King Lear" or in a text of Confucius
- counting occurrences of a certain combination of amino acids in a protein database.
- pulling a card from the deck

Most things done using pen and paper considered as random experiment – outcome is not known/certain in advance, while In Labs - outcome is certain

Sample Space

- Discrete sample spaces.
- Continuous sample spaces



All possible outcomes of an experiment is called sample space. Tossing a coin – S: {H, T}

Discrete – 1, 2, 3, 4…, Continuous – outcomes in ranges/intervals called continuous

Finite Sample space – finite set/elements, Infinite space – infinite numbers etc.

- Event
- Independent events
- Dependent events

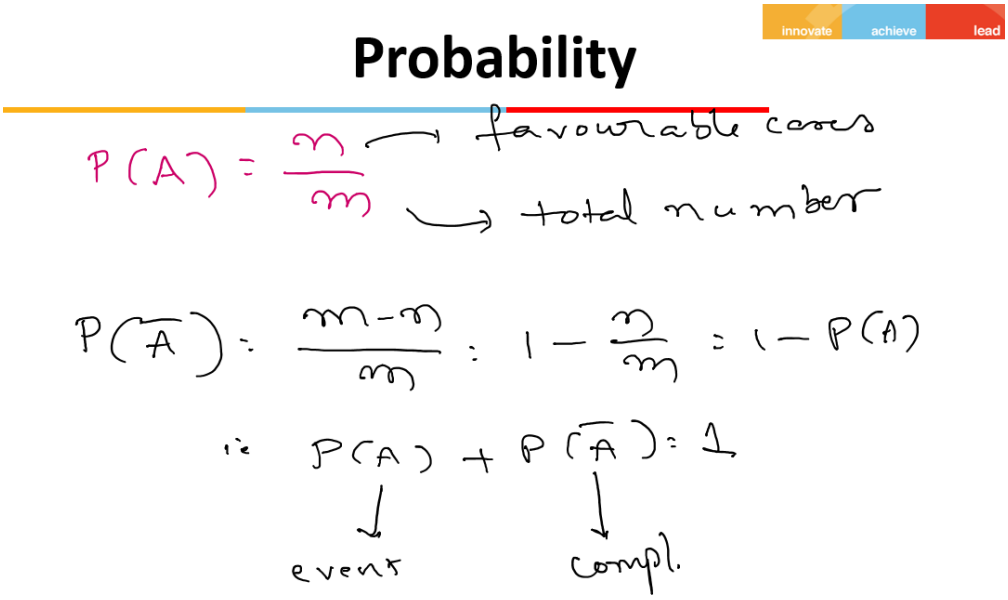
Event is an subset of sample space, getting an outcome from an experiment

Events which are dependent on each other – Dependent Events – i.e. two consecutive head/tails are dependent, same number on both dice

Deck of cards – two events – pulling a card – with replacement (52 cards every time) and w/o replacement (choose from - 51 cards next time)

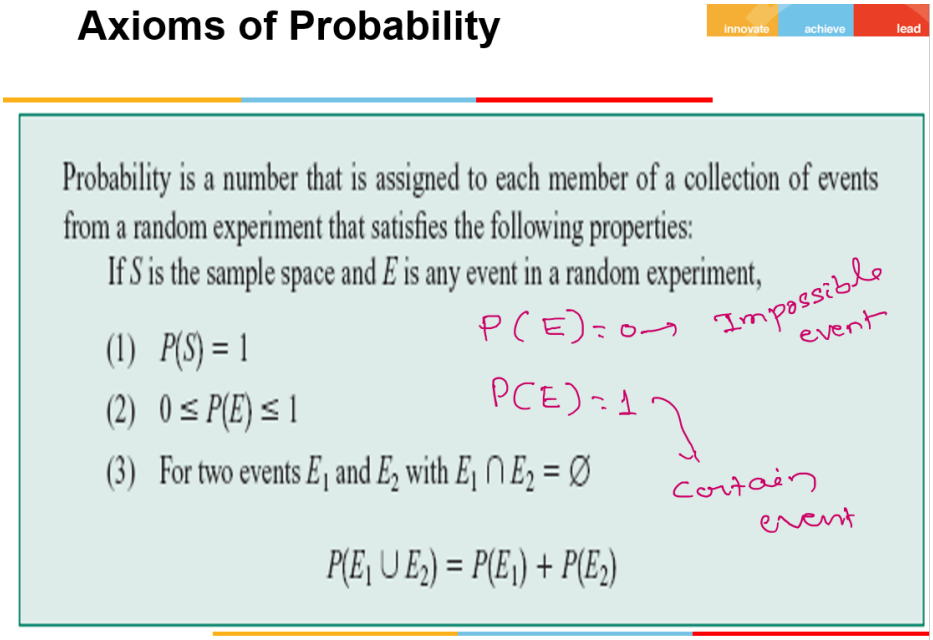
Two students in class – 40 marks to pass an exam, A and B both passing the examination – **independent events**, both are getting 1st rank in university – **Dependent event**

Musical chair game, one person getting a chair depends on whether another person got…



90% chance we'll get this, 50% chances of rain, 70% chance – this team will win

100% - confirmed (total events) → 90% - favorable events



Event – getting number less than or equal to 5 – total numbers less than 5 – 0 to 5 → 6

Throwing a dice - 0/6, 6/6 – probability is always between 0 and 1 (certain – all are in favor)

0 Probability (impossible events) – there is no other element in sample space which favors it

Some event - Probability – 1 → certain event, Mutually exclusive events – P(E1)+P(E2)

The sales manager of an e commerce company says that 80% of those who visit their website for the first time do not buy any mobile. If a new customer visits the website, what is the probability that the customer would buy mobile

i.e $P(A) = \frac{80}{100} = 0.8$

$P(\bar{A}) = 1 - P(A) = 1 - 0.8 = 0.2$

P(A) (Probability is getting Ace) – 52 cards, P(B) getting number less than 4 – 51 cards

Dependent events (w/o replacement) – probability of second events depends on first event

	Blue	Black	Brown	Total
Software prog	35	25	20	80
Project Mgrs	7	8	5	20
Total	42	33	25	100

If an employee is selected at random, what is the probability that he is a software prog?

i.e $\frac{80}{100}$

Sample space: 100 → Soft. Dr 80

....., 42., what is the probability that he is wearing a blue trouser $\frac{42}{100}$

If software prog. is selected – what is the probability of blue trouser – **35/80**

A P(B) 3 P(A) B

innovate achieve lead

A Survey conducted by a bank revealed that 40% of the accounts are savings accounts and 35% of the accounts are current accounts and the balance are loan accounts. → C

- What is the probability that an account taken at random is a loan account ? i.e $P(C) = 1 - (0.4 + 0.35) = 0.25$
- What is the probability that an account taken at random is NOT savings account ? i.e $P(\bar{A}) = 1 - 0.40 = 0.60$
- What is the probability that an account taken at random is NOT a current account $P(\bar{B}) = 1 - 0.35 = 0.65$
- What is the probability that an account taken at random is a current account or a loan account? → Next Session – discuss

Lec 3 Descriptive Statistics & Conditional Probability

Today…..

- Recall the past for a while_ Simple tools
- Visualization of data
- Basics of probability
- Discussion & Problems on probability
- Conditional probability
- Box plot

A Survey conducted by a bank revealed that 40% of the accounts are savings accounts and 35% of the accounts are current accounts and the balance are loan accounts.

- What is the probability that an account taken at random is a current account or a loan account?

C L

$P(C \cup L) = P(C) + P(L)$

$= 0.4 + 0.25$

$= 0.65$

Mutually exclusive events

P(A U B) = P(A) + P(B) – P(A ∩ B) (here 0 for mutually exclusive events)

4

From a Hospital data it is found that 45% of the patients are having high B.P. Also it was found that 35% of these patients having high B P is also having diabetes.

What is the probability that a patient having high BP is also diabetic

4

From a Hospital data it is found that 45% of the patients are having high B.P. Also it was found that 35% of these patients having high B P is also having diabetes.

A → B.P

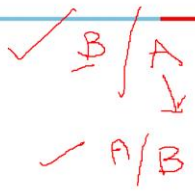
B → diabetes

What is the probability that a patient having high BP is also diabetic

P(A/B) ?

Conditional Probability

The probability of event B given that event A has occurred P(B|A) or, the probability of event A given that event B has occurred P(A|B)



Conditional Probability

B|A – first event happened is A then B, A|B – First B then A (A given B)

P (B|A) and P(A|B) both are same ? **No**

Conditional Probability

innovate achieve lead

Definition

The **conditional probability** of an event B given an event A, denoted as $P(B|A)$, is

$$P(B|A) = P(A \cap B)/P(A) \qquad (2\text{-}9)$$

for $P(A) > 0$.

Example

Consider the random experiment of throwing a dice 15 times and the sample space is

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

Consider A = getting a number < 4 = { 1,3,3,2,3,2,2,3}

B = Getting an even number = { 6,4,2,4,4,2,2,4}

Now Getting a number < 4 which is even = {2,2,2)

Now getting an even number which is < 4 = { 2,2,2}

P(A) = 8/15, P(B) = 8/15

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

Consider A = getting a number < 4 = { 1,3,3,2,3,2,2,3}

B = Getting an even number = { 6,4,2,4,4,2,2,4}

Now Getting a number < 4 which is even = {2,2,2)

Now getting an even number which is < 4 = { 2,2,2}

Here **P(A|B) = P(B|A)** (Because both the sample spaces are same)

Multiplication and Total Probability Rules

innovate achieve lead

Multiplication Rule

$$P(A \cap B) = P(B|A)P(A) = P(A|B)P(B) \qquad (2\text{-}10)$$

Total Probability Rule (two events)

For any events A and B,

$$P(B) = P(B \cap A) + P(B \cap A') = P(B|A)P(A) + P(B|A')P(A') \qquad (2\text{-}11)$$

Independence

innovate achieve lead

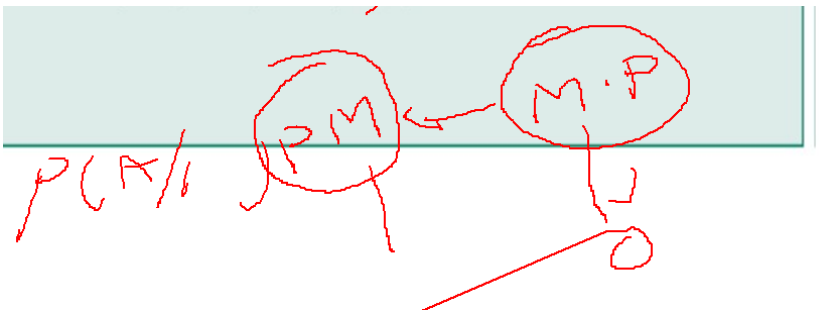
Definition (two events)

Two events are **independent** if any one of the following equivalent statements is true:

(1) $P(A|B) = P(A)$

(2) $P(B|A) = P(B)$

(3) $P(A \cap B) = P(A)P(B)$ (2-13)



B P is also having diabetes.

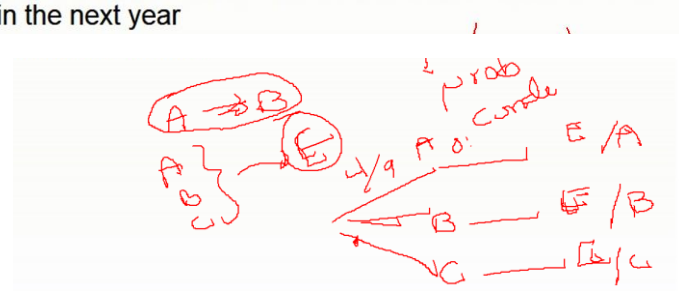
What is the probability that a patient having high BP is also diabetic

What software, which tool, understanding the problem

innovate achieve lead

Example:

In the year 2017,three candidates A,B and C are competing for the post of Vice – Chancellor. The probabilities of getting appointments of these are 4;2;3 respectively. The probability that A if selected would introduce new pay policy is 0.35 and the probabilities of B and C doing the same are 0.52 and 0.80 respectively. What is the probability that there will be new policy in the next year



All three events are mutually exclusive events – only one event will happen

Independence

innovate achieve lead

Definition (two events)

Two events are **independent** if any one of the following equivalent statements is true:

(1) $P(A|B) = P(A)$

(2) $P(B|A) = P(B)$

(3) $P(A \cap B) = P(A)P(B)$ (2-13)

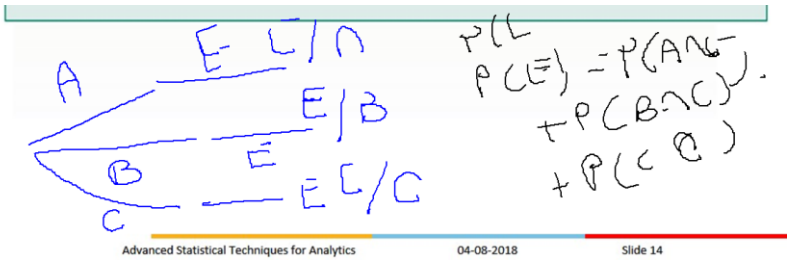
Bayes' Theorem

Definition

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad \text{for} \quad P(B) > 0 \qquad (2\text{-}15)$$

Bayes' Theorem

Definition



$$P(B) = P(E_1 \cap B) + P(E_2 \cap B) + \Lambda + P(E_n \cap B)$$

For each $P(E_i \cap B) = P(B \mid E_i)P(E_i)$

$$\begin{aligned} P(B) &= P(E_1 \cap B) + P(E_2 \cap B) + \Lambda + P(E_n \cap B) \\ &= P(B \mid E_1)P(E_1) + P(B \mid E_2)P(E_2) + \Lambda + P(B \mid E_n)P(E_n) \\ &= \sum_{i=1}^{n} P(B \mid E_i)P(E_i) \end{aligned}$$

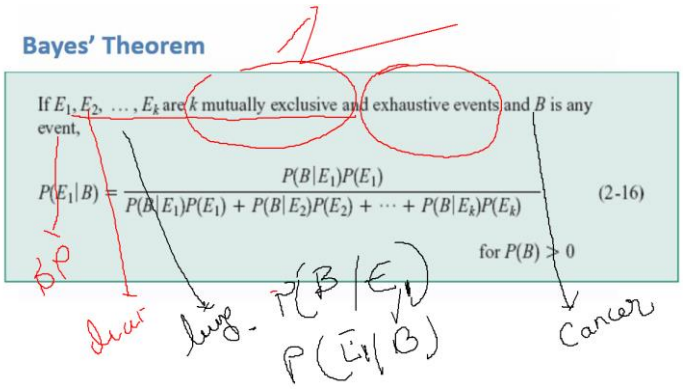
Bayes' Theorem

If $E_1, E_2, \ldots, E_k$ are k mutually exclusive and exhaustive events and B is any event,

$$P(E_1 \mid B) = \frac{P(B \mid E_1)P(E_1)}{P(B \mid E_1)P(E_1) + P(B \mid E_2)P(E_2) + \cdots + P(B \mid E_k)P(E_k)} \qquad (2\text{-}16)$$

for $P(B) > 0$

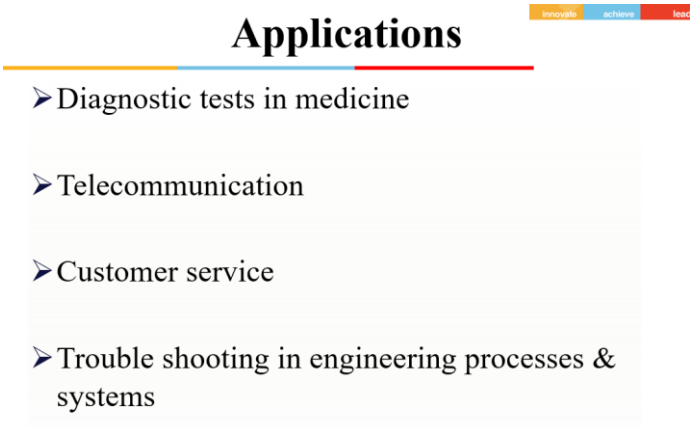
Intersection of any two events is 0 **mutually exclusive** events, all possible events that happen – exhaustive event i.e. Patient – who's having b.p., diabetes, lung disorder – patient is having cancer – take hospital data – who're having B.P., Diabetes, Lung etc.



Applications – Real Life examples – Detecting spam emails, Accident data – understand traffic and avoid accident in future, how many deaths – dunk and drive, reckless driving, road condition – what is the P – accident took place because of D&D, reckless driving etc.

What is the cause behind cancer – genetic disorder, food habits, environment etc.

P(B|E1) → Priori probability → we have to find P(E1|B) – based on data available

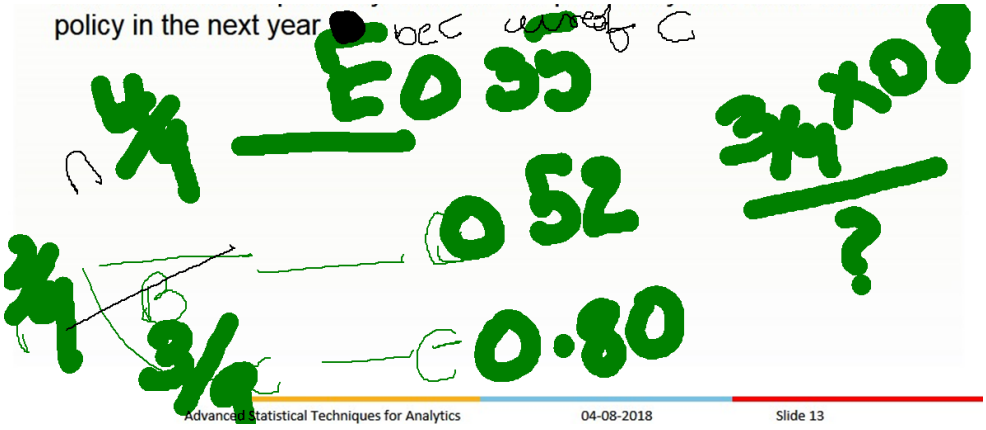


Probability of getting bugs from same team while working in a project

CEO selection – data collected – who will take company to new heights – Probability of company's growth because of **Person A (0.8)**, Person B (0.6), Person C (0.2)

This is how Bayes' theorem helps us to understand situation before taking decision

Denominator – total probability - **P(A|E1)*P(E1) + P(B|E2)*P(E2) + P(C|E3)*P(E3)**



P(C|A),P(C|B,P(C|D) – Prior probabilities we have – we want to find major cause for cancer

Lec 4 Descriptive Statistics

Today…..

- Recall the past for a while_ Conditional probability and Baye's theorem & some examples
- Random variables
- Probability distribution
- Examples

Event wise P(A|B) & P(B|A) are same – outcome of the event – same elements

But probability wise different – P(A|B) – first B happens then A – take B then take A

Conditional Probability
and Baye's theorem —

'Recap'

Conditional probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{or}$$

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

$$\therefore P(A \cap B) = P(A|B)\,P(B) = P(B|A)\,P(A)$$

P(A|B) – sample space – B → on which event happened

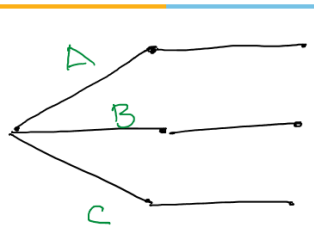
Bayes' – sequence of events - Various events happened before happening some event

$$P(A|E) = \frac{P(A \cap E)}{P(E)} = \frac{P(E|A)\,P(A)}{\sum_i P(E|A_i)\,P(A_i)}$$

'Baye's theorem'

P(E) – total probability, Baye's theorem – take decision

Baye's theorem



Total prob $= \sum_{A,B,C} P(E|A)P(A)$

ie $P(A/E) = \dfrac{P(E|A)\,P(A)}{\sum P(E|A)\,P(A)}$ ✓

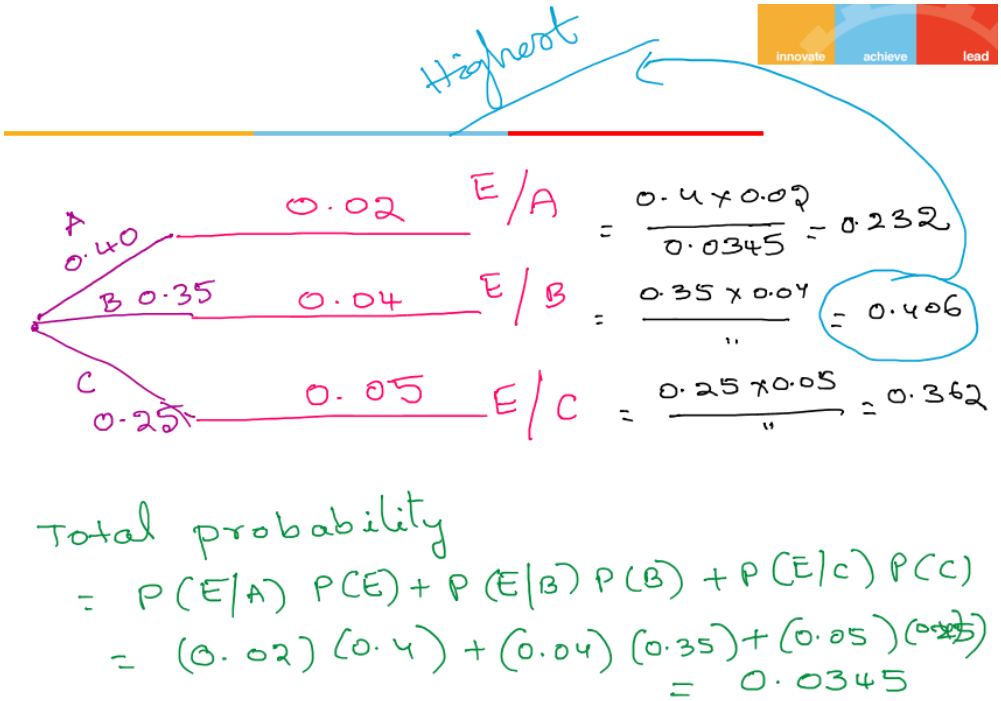
or $P(B|E) = \dfrac{P(E|B)\,P(B)}{\sum P(E|A)\,P(A)}$ ✓

Expecting some examples
or case studies
or applications from
Participants

Spam checkup, Statistical Machine Translation – using Corpus – translate one language to another, Medical Application, Communication, past data – reservation probability

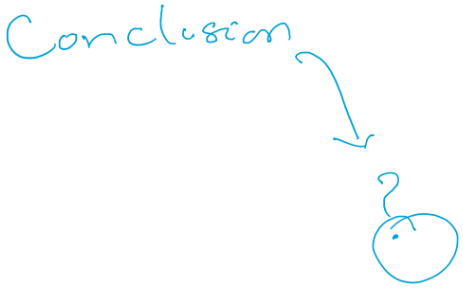
Example :-

In a factory, three machines A, B and C manufactures 40%, 35% & 25% of the total output. From the past records, it is observed that of their out-put 2, 4, 5 percents are defective. A product is drawn at random and is found to be defective. what is the prob. that it was manufactured by A | B | C ? What is the observation?



True Negative, False positive – Posterior probability

Baye's Application - Motive here is to improve manufacturing performance, Tool – give data – applied – outcome – how to interpret outcome



P(B|E) – probability of defected items produced & it's from machine B – machine B needs fix

Consider Cricket match scenario – lost match – reasons – 1) poor batting 2) poor fielding 3) poor bowling – we have prior probability before experiment - posterior probability (after an event) – probability – we lost match due to fielding is 0.4357 – need to improve fielding

Random Variable

Generalization of values – tossing of a coin 3 times or 3 coins – **Xc** = 1 – 1 Head, Xc = 2 → 2 Head – Xc = number of heads – discrete values – **Discrete** random variable, **Continuous** values – temperate – infinite number of times

Random Variables

We now introduce a new term

Instead of saying that the possible outcomes are 1,2,3,4,5 or 6, we say that ***random variable*** X can take values {1,2,3,4,5,6}. **A random variable is an expression whose value is the outcome of a particular experiment.**

The random variables can be either *discrete* or *continuous.*

It's a convention to use the upper case letters (X, Y) for the names of the random variables and the lower case letters (x, y) for their possible particular values.

Random Variables

Definition

A **discrete** random variable is a random variable with a finite (or countably infinite) range.

A **continuous** random variable is a random variable with an interval (either finite or infinite) of real numbers for its range.

Random Variables

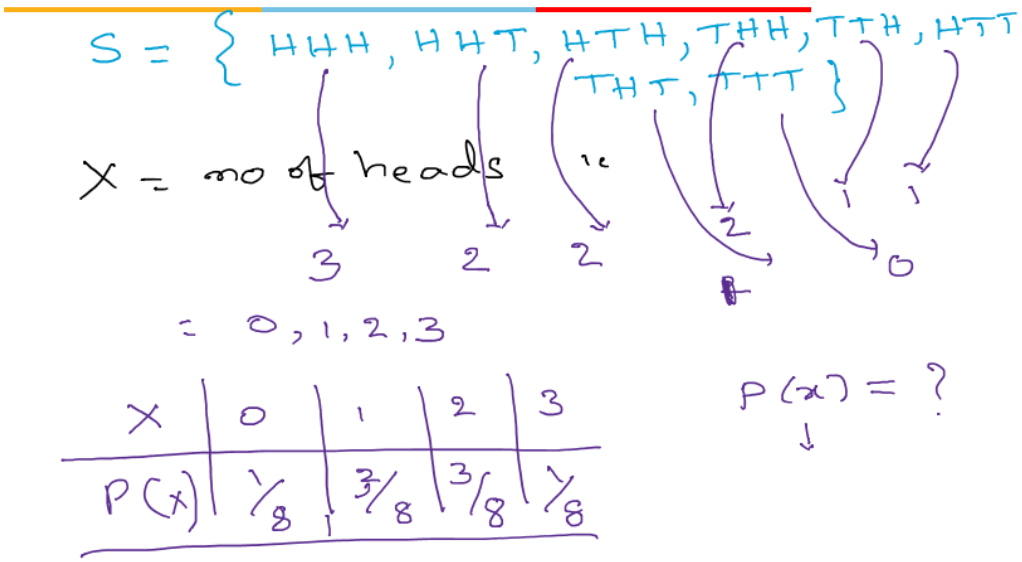
Examples of Random Variables

Examples of **continuous** random variables:

electrical current, length, pressure, temperature, time, voltage, weight

Examples of **discrete** random variables:

number of scratches on a surface, proportion of defective parts among 1000 tested, number of transmitted bits received in error



We can have function – **fn(S) → x** – returns – number of Heads – range – [0, 3]

Probability – P(x) – getting n Heads – we can do & represent without using Random variables – random variables here helps to represent & organize it

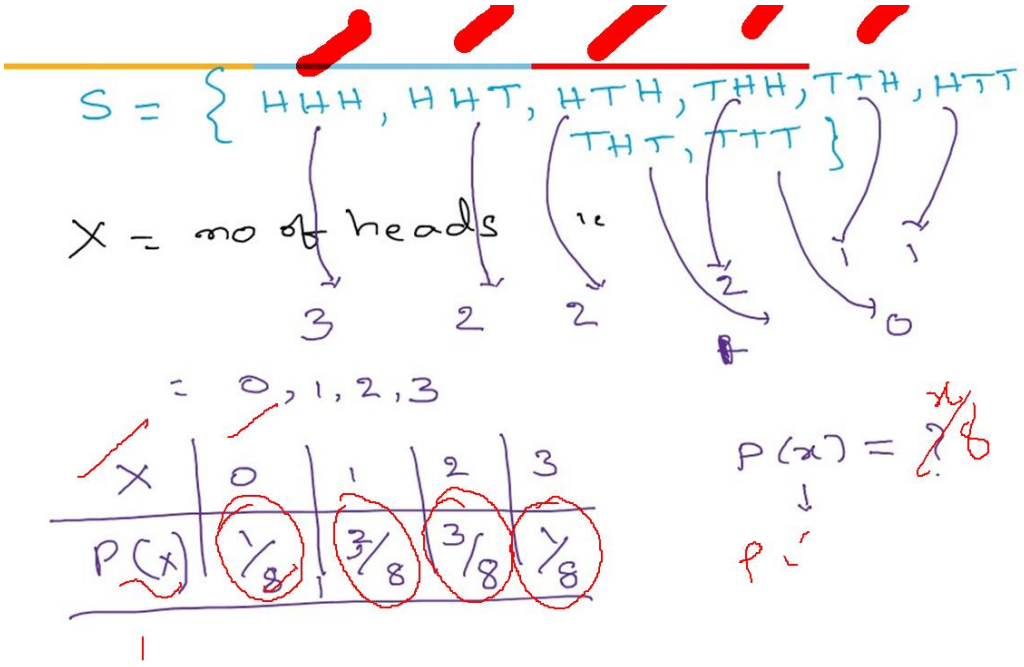
P(x) – Probability distribution function – like frequency distribution – how probabilities are distributed – at x = 0, x = 1 …

P(x) = x/8 → then we can find probability for any x – in terms of variable

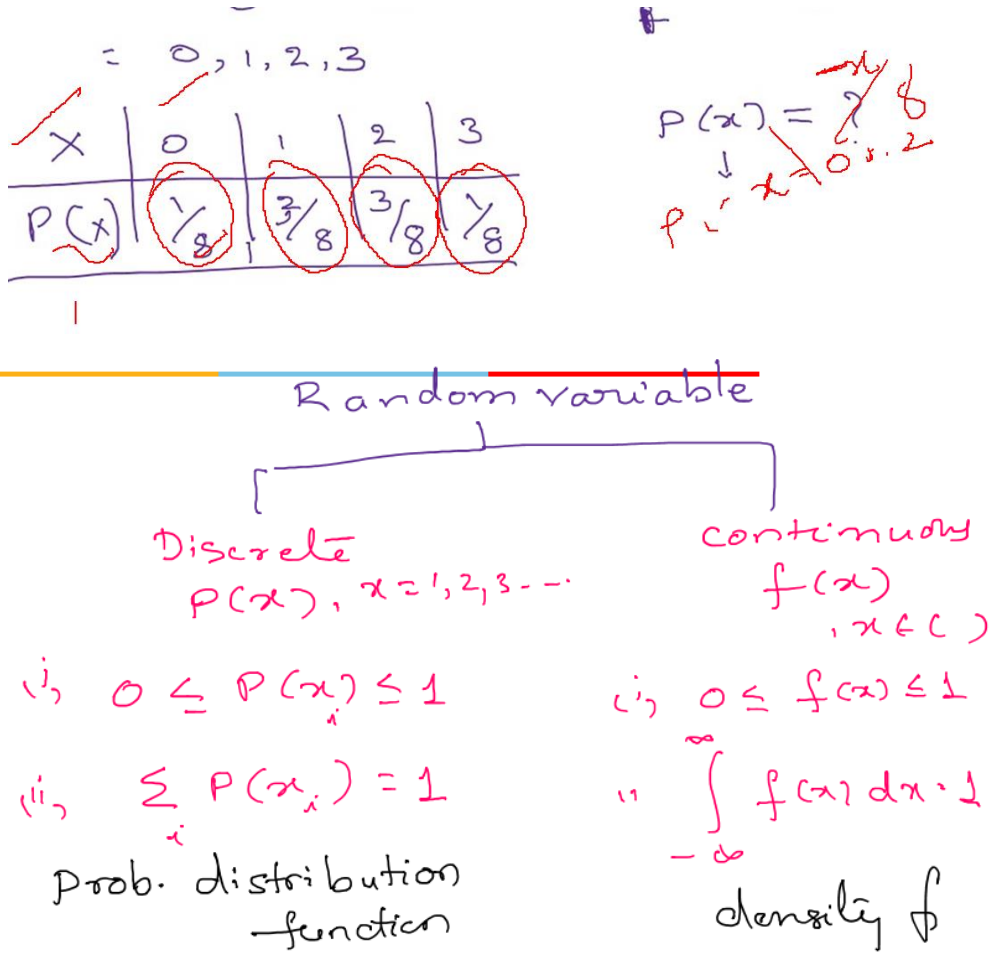
Observation – All values lies between 0 & 1 – all Heads & No Heads – are equally distributed

Frequency distribution - Distribution of marks – students – range - 10-20, range – 20-30 etc. – class strength = 60 → all frequency sum - 60

Probability distribution – probabilities are distributed – **total probability = 1**



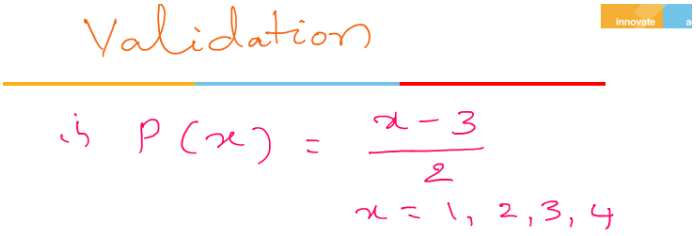
Validate – if P(x) is valid expression or not – check – probability always lies between [0, 1]



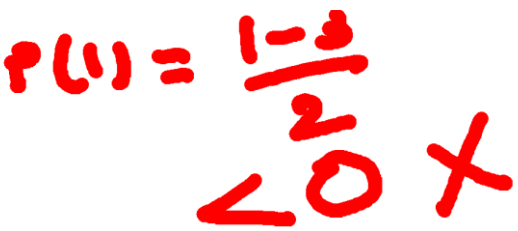
f(x) – probability density function, P(x) – Probability distribution function

P(x) = 0 → impossible event – rarest event – for any value of x → P(x) = 0

f(x) – related with continuous random variable, P(x) – discrete – also we can use P(x) for discrete random variable



Check for all values of x → if any value is negative or value > 1 – not valid



If valid for all individual values – then check for distribution – total value = 1

another

$$P(x) = \frac{x^2}{5}, \quad x = 0,1,2,3,4$$

$\sum_{x=0}^{3} P(x) > 1$

Example:-

x	0	1	2	3	4	5	6	7
$P(x)$	0	K	$2K$	$2K$	$3K$	K^2	$2K^2$	$7K^2+K$

(i) K value: ?

(ii) $P(x<6)$

(iii) $P(x \geq 6)$

(iv) $P(3<x\leq 6)$

x	0	1	2	3	4	5	6	7
$P(x)$	0	K	$2K$	$2K$	$3K$	K^2	$2K^2$	$7K^2+K$

(i) K value: ? $\sum P(x)=1$

(i) K value: ? $\sum P(x) = 1$

$0+K+2K+\cdots = 1$

$10K^2 + 9K - 1 = 0$

$K = -1, \frac{1}{10}$

(ii) $P(x<6)$

(iii) $P(x \geq 6)$

(iv) $P(3<x\leq 6)$

innovate achieve lead

Example:-

x	0	1	2	3	4	5	6	7
$P(x)$	0	k	$2k$	$2k$	$3k$	k^2	$2k^2$	$7k^2+k$

i) k value: ? $\sum P(x) = 1$

$\Rightarrow 0 + k + 2k + \dots = 1$

$10k^2 + 9k - 1 = 0$

$k = -1, \frac{1}{10}$

ii) $P(x < 6)$

iii) $P(x \geq 6)$

iv) $P(3 < x \leq 6)$

innovate achieve lead

Example:-

x	0	1	2	3	4	5	6	7
$P(x)$	0	k	$2k$	$2k$	$3k$	k^2	$2k^2$	$7k^2+k$

i) k value: ? $\rightarrow k = \frac{1}{10}$

ii) $P(x < 6) \rightarrow P(0) + \dots P(5) = k^2 + 8k = \frac{81}{100}$

iii) $P(x \geq 6) \rightarrow P(6) + P(7) = \frac{19}{100}$

iv) $P(3 < x \leq 6) \rightarrow P(4) + \dots = \frac{33}{100}$

x	0	1	2	3	4	5	6	7
$P(x)$	0	k	$2k$	$2k$	$3k$	k^2	$2k^2$	$7k^2+k$

$P(x >= 6) = 1 - P(x < 6)$

Example:

$$f(x) = \begin{cases} Kx^2, & 0 < x < 3 \\ 0, & \text{otherwise} \end{cases}$$

$K = ?$

$P(1 < x < 2)$

Example:

$$f(x) = \begin{cases} Kx^2, & 0 < x < 3 \\ 0, & \text{otherwise} \end{cases}$$

$$\int f(x)\,dx = 1 \Rightarrow \int_0^3 Kx^2\,dx = 1$$

$$\Rightarrow K\left[\frac{x^3}{3}\right]_{x:0}^{3} = 1$$

$$\Rightarrow K/3\,[27 - 0] = 1 \quad \text{i.e.} \quad K = 1/9$$

Example:

$$\int x^n\,dx = \frac{x^{n+1}}{n+1}$$

$$f(x) = \begin{cases} Kx^2, & 0 < x < 3 \\ 0, & \text{otherwise} \end{cases}$$

$K = ? \rightarrow 1/9$

$P(1 < x < 2) \rightarrow = \int_1^2 Kx^2\,dx$

$= 1/9\left[\frac{x^3}{3}\right]_{x:1}^{2}$

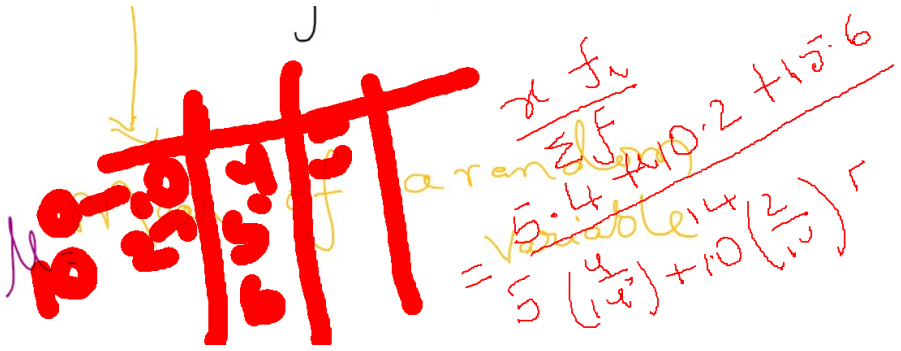
$= 1/27\,[8 - 1] = 7/27$

Expectation of a random Variable

$$E(x) = \sum_i x_i p(x_i)$$

$$= \int x f(x)\, dx$$

μ = Mean of a random variable



4 / 14, 2/14 – probabilities – 4,2 etc. in corresponding space – 14 – sample space

Variance of a r.v

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

$$= E(x^2 + \mu^2 - 2\mu x)$$

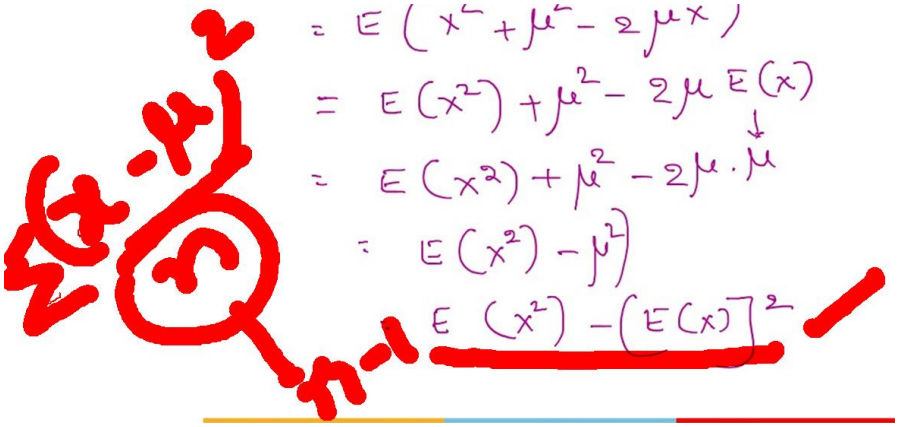
$$= E(x^2) + \mu^2 - 2\mu E(x)$$

$$= E(x^2) + \mu^2 - 2\mu \cdot \mu$$

$$= E(x^2) - \mu^2$$

$$= E(x^2) - [E(x)]^2$$

E(x) = expectation of x – mean – Mu



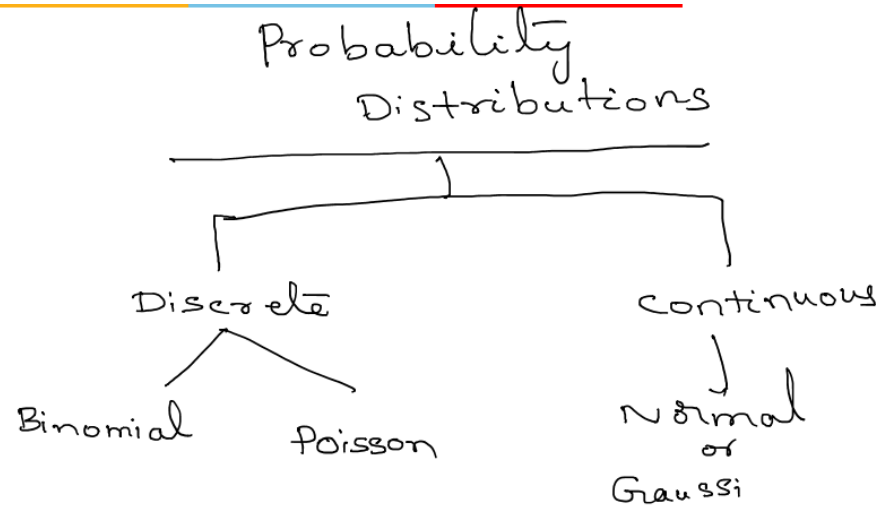
Sampling – n → (n-1)

ie Mean $\mu = E(x) = \sum x\, p(x)$
$= \int x\, f(x)\, dx$

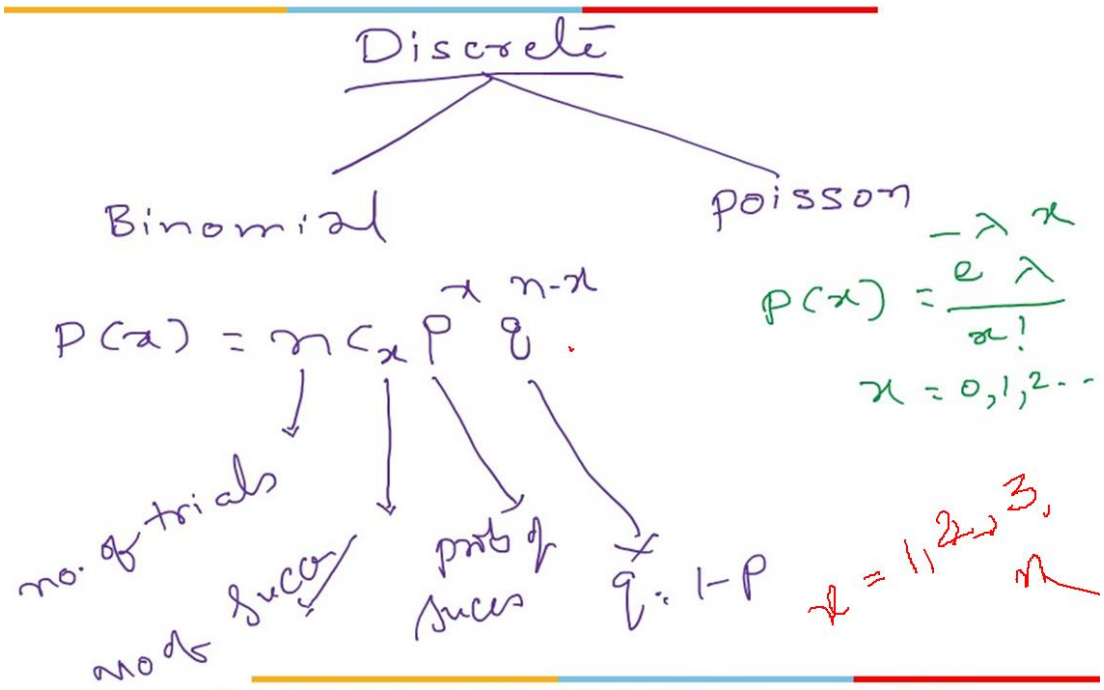
Variance $\sigma^2 = E[(x-\mu)^2]$
$= E(x^2) - [E(x)]^2$

$E(x^2) = \sum x^2 p(x)$ → discrete

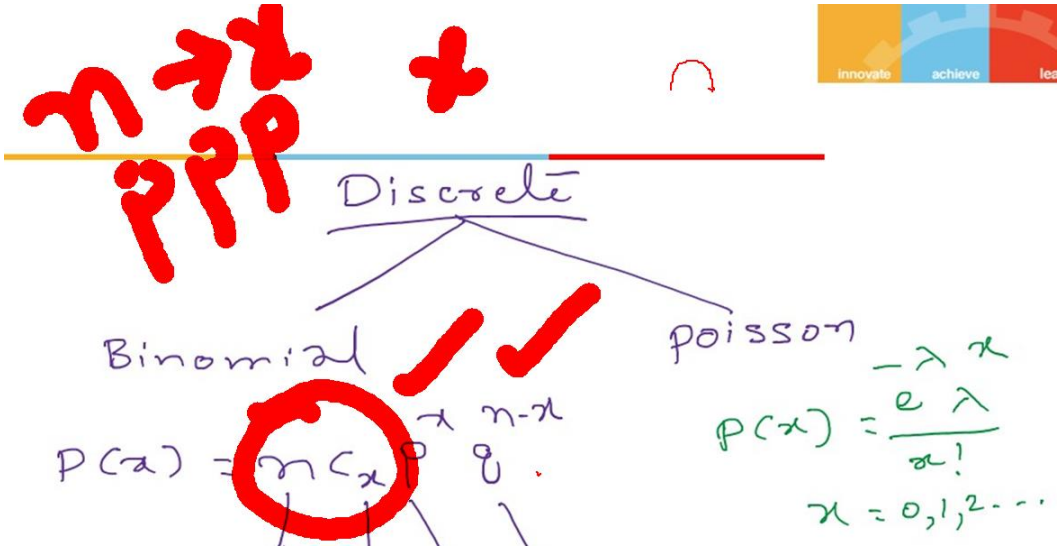
$\int x^2 f(x)\, dx$ → continuous



Communication – Gaussian distribution



Binomial trials – outcome is binomial always – success or failure, Probability of getting success/failure – independent of trials – constant – Head/Tail – Trial 1 → ½, Head/Tail → Trial 2 → ½ - independent & constant



n times – p, (n-x) times – q

P(x) – validation – lies between [0,1], total probability = 1, Sum(P(x))

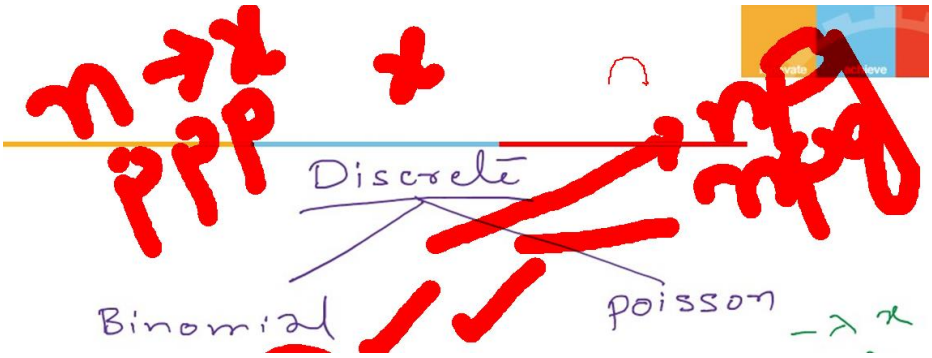
Binomial Expansion – (p+q) ^ n = (1) ^ n = 1

Poisson distribution – rarest of the rare events – n is very large, p is very small – almost 0

Lambda – some parameter, n is very large, p is very small → Binomial to Poisson

Manufacturing units – six sigma standards (few follow), Bugs in million lines of code, Testing & getting defects

Binomial – **mean = n * p, variance - = n * p * q**, variance > mean

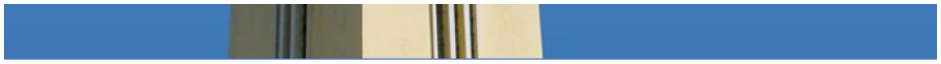


Binomial – n is very large, p is very small – n*p – almost constant → **Lambda –** Poisson

Poisson – mean = variance = Lambda

Which distribution to use – data set – to model it to next level → find mean & variance – if mean > variance → Binomial else if mean == variance – Poisson is suitable

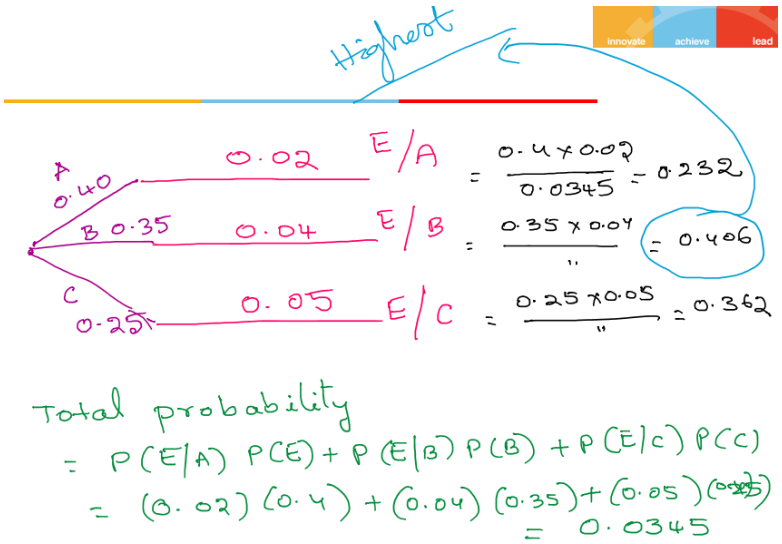
Lec 5 Descriptive and inferential statistics



L- 5: Descriptive and inferential statistics

Agenda

- Quick Review of the topics covered in previous class
- Normal Distribution
- Sampling
- Testing of Hypothesis



Number of products defected – A → 0.4 * 0.232, similar for B & C

Example

Technicians regularly make repairs when breakdowns occur on an automated production line. Janak, who services 20% of the breakdowns, makes an incomplete repair 1 time in 20.Tarun ,who services 60% of the breakdowns ,makes an incomplete repair 1 time in 10 Gautham, who services 15% of the breakdowns, makes an incomplete repair 1 time in 10 and Prasad ,who services 5% of the breakdowns, makes an incomplete repair 1 time in 20.For the next problem with the production line diagnosed as being due to an initial repair that was incomplete, what is the probability that this initial repair was made by Janak?

Solution

Let A be the event that the initial repair was incomplete
B_1 that the repair was made by Janak
B_2 that it was made by Tarun ,
B_3 that it was made by Gautham,
B_4 that it was made by Prasad,

$$P(B_1/A) = \frac{P(B_1)P(A/B_1)}{P(B_1)P(A/B_1)+P(B_2)P(A/B_2)+P(B_3)P(A/B_3)+P(B_4)P(A/B_4)}$$

$$= \frac{(0.20)(0.05)}{(0.20)(0.05)+(0.60)(0.10)+(0.15)(0.10)+(0.05)(0.05)}$$

$$= 0.114$$

Better to go for tree like representation – for Baye's theorem

Problem

On the average, five cars arrive at a particular car wash every hour. Let X count the number of cars that arrive from 10AM to 11AM. (mean = 5).What is the probability that no car arrives during this period?

Poisson dist. $\lambda = 5$ $P(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$

$P(x=0) = \frac{e^{-5}5^{0}}{0!} = e^{-5}$ ✓

No. of cars – discrete random variable – Either Binomial or Poisson – which one – Binomial Distribution (Binomial Trials – success & failure – independent) – here Poisson is more appropriate – n & p is not known, x = 0 – no car arrives

Given a problem – choosing correct model is very important & challenging task

Suppose the car wash is in operation from 8AM to 6PM, and we let Y be the number of customers that appear in this period.(lambda = 50).

What is the probability that there are between 48 and 50 customers, inclusive?

$\lambda = 5$,

$P(48 \le x \le 50) = P(48) + P(49) + P(50)$

$= \frac{e^{-5}5^{48}}{48!} + \frac{e^{-5}\cdot 5^{49}}{49!} + \frac{e^{-5}\cdot e^{50}}{50!}$

No clue about – total number of trials – most of the time - we can go for Poisson distribution

Binomial – each trial – either success or failure – probability remains same – constant – H/T

Poisson– n is very large, p is very small – n*p is almost constant or don't have **n** value clearly

Normal distribution

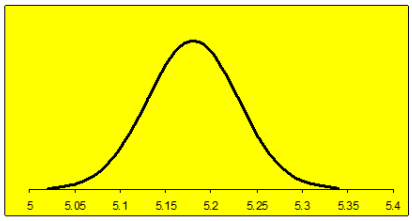
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

Naïve based algorithm – numerical data, any sample – repeatedly – becomes Normal Distribution, Signal processing – communication – Gaussian curve – important

Magnetic field – density vs space, dealing with team – performance of team members – Normal curve is used – Bell curve

Normal Distribution

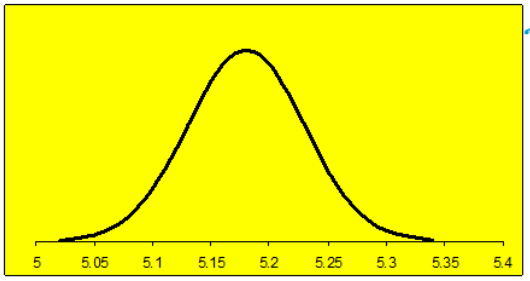
Probability density function - $f(X)$



$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1/2(X-\mu)^2}{\sigma^2}}$$

Bell/Normal/Gaussian Curve – symmetric – never touches axis – continuous – infinite

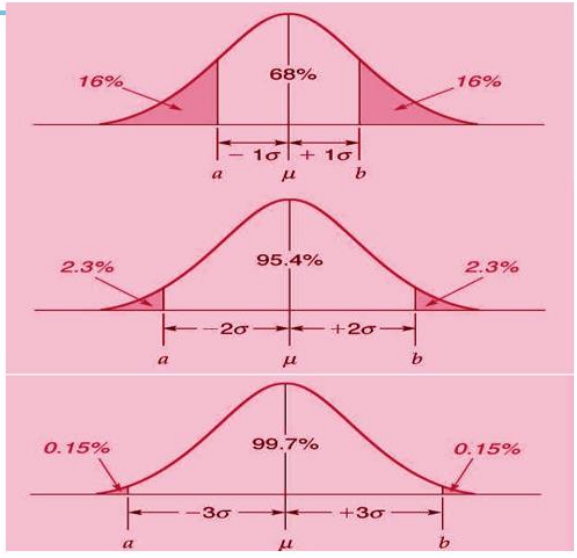
Probability density function - $f(X)$



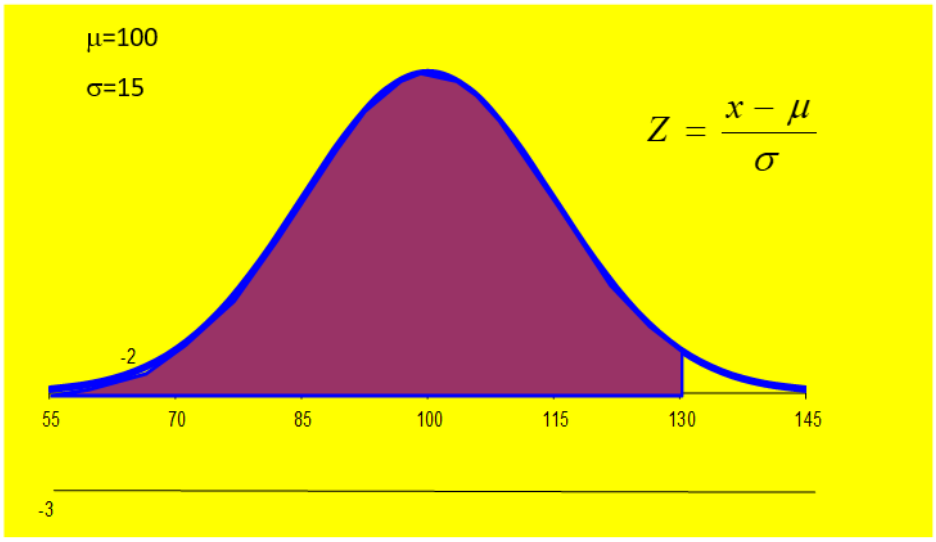
→ Normal Curve or Gaussian curve

Three Common Areas Under the Curve

Three Normal distributions with different areas



Standard Normal Distribution



How to find

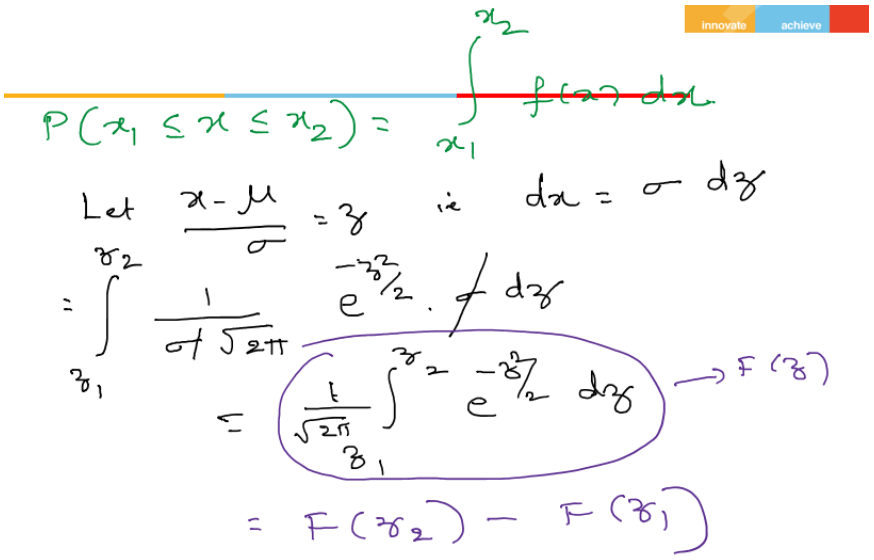
$$P(2 < x < 5) = \int_{x=2}^{5} f(x)\,dx = \int_{x=2}^{5} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx$$

Integration ????

To evaluate this integral – complex task

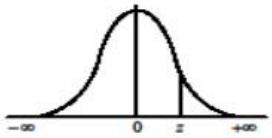
Note

Since the normal density cannot be integrated in between every pair of limits a and b , probabilities relating to normal distributions are usually obtained from special tables (see tables)

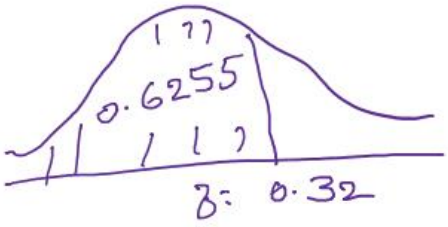


Probability of x lying between x1 & x2 – convert them into standard normalization form

NORMAL DISTRIBUTION TABLE



	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998



z = 0.32 → 0.6255 → limit is - infinity to 0.2 → Integration is 0.6255 – area under curve

0.8413 → - infinity to 1 → area under curve

Area under cureve – both side – positive & negative – probability – 0.5 – two halves - similar

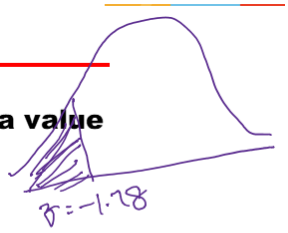
Normal distribution will take on a value

1) **to the left of z = -1.78**

2) **to the right of z = -1.45**

3) **corresponding to -0.80 ≤ z ≤ 1.53**

4) **to the left of z = -2,52 and to the right of z = 1.83**



P(z) < -1.78 → -1.78 – left side of the curve

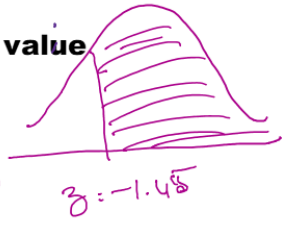
Normal distribution will take on a value

1) **to the left of z = -1.78**

2) **to the right of z = -1.45**

3) **corresponding to -0.80 ≤ z ≤ 1.53**

4) **to the left of z = -2,52 and to the right of z = 1.83**



Left side of value – direct value from log, right side – (1 – value from log)

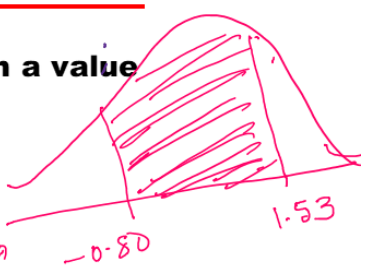
Normal distribution will take on a value

1) **to the left of z = -1.78**

2) **to the right of z = -1.45**

3) **corresponding to -0.80 ≤ z ≤ 1.53**

4) **to the left of z = -2,52 and to the right of z = 1.83**



ie $F(1.53) - F(-0.80)$

F(1.53) = 0.9370, F(0.80) = 0.5319, Now a days table has negative values

This problem is static – only values varies → - infinity – lower limit fixed – upper limit varies

F(z) = 1 – F(z) → F(-1.0) = 1 – F(1.0) = 1 – 0.8413 = 0.1587

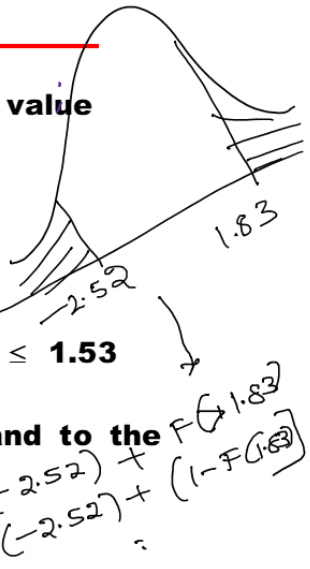
Normal distribution will take on a value

1) to the left of z = -1.78

2) to the right of z = -1.45

3) corresponding to -0.80 ≤ z ≤ 1.53

4) to the left of z = -2,52 and to the right of z = 1.83



Calculation of probabilities using a normal distribution

Problem

The mean and standard deviation of a normal variate are 8 and 4 respectively

Find 1) P [5≤ X ≤10]
2)P [X ≥5]

Solution

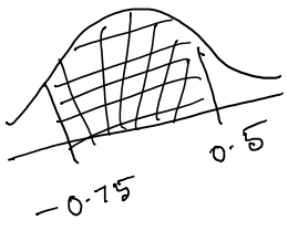
1) $\mu = 8$

$\sigma = 4$

We know that $Z = \frac{X-\mu}{\sigma} = \frac{X-8}{4}$

When X=5 $Z = \frac{5-8}{4} = -0.75$

When X=10 $Z = \frac{10-8}{4} = 0.5$



$P [5 \le X \le 10] = P [-0.75 \le Z \le 0.5]$

Normalization helps to convert any values to [0, 1]

= F (0.5) – F (- 0.75)

= 0.6915 – .22663 = 0.4649

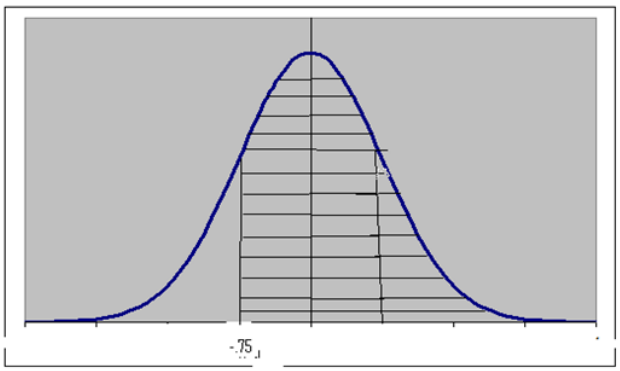
Probability of students having marks between 5 & 10 – 0.4649 – total 100 students in class = 100 * 0.4649 → 49 students

2)P [X≥5] = P [Z ≥ -0.75] = 1- F (-0.75)

= F(0.75)

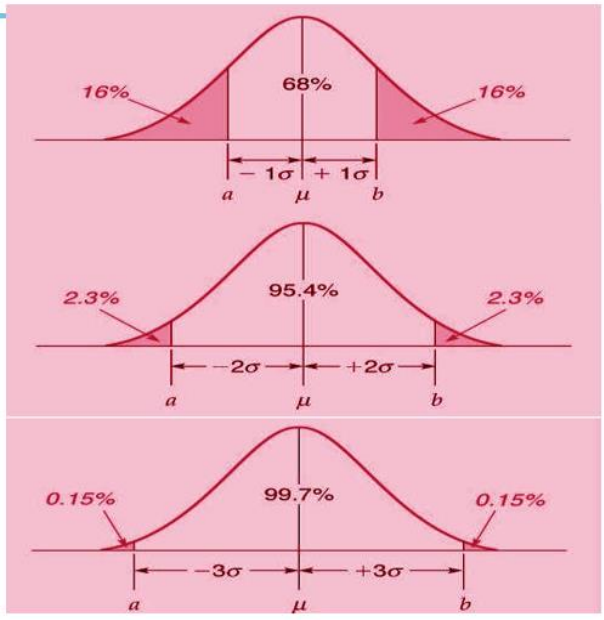
0.7734

= 0.7734



Three Common Areas Under the Curve

Three Normal distributions with different areas



Example:

In a test conducted on 1000 candidates, the average score is 42 with a S.D of 24. Assuming normal distribution, find

a) no of candidates whose score exceeds 58

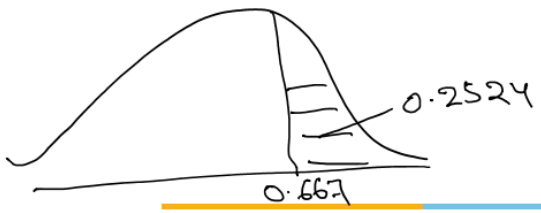
b) no of candidates whose scores lies b/w 30 and 66

a) Score exceeds 58 i.e $x > 58$

$P(x > 58)$ $\quad z = \frac{x-\mu}{\sigma}$

$= \frac{58-42}{24}$

$P(z > 0.667)$ $\quad = 0.667$



1000×0.2524

$\downarrow$

252 candidates

b) $P(30 \le x \le 66)$

$\downarrow$ $\quad \downarrow$

$\frac{30-42}{24}$ $\quad \frac{66-42}{24}$

$= -0.5$ $\quad 1$

$P(-0.5 \le z \le 1)$

$= F(1) - F(-0.5)$

$= 0.5328$ $\quad 0.5328 \times 1000$

$= 532.8 \rightarrow 533$

Suppose find P(100<x<1000) → we have to do like this – P(100)+ P(101) + … + P(1000)

Suppose the car wash is in operation from 8AM to 6PM, and we let Y be the number of customers that appear in this period.(lambda = 50).

What is the probability that there are between 48 and 50 customers, inclusive?

$\lambda = 5$,

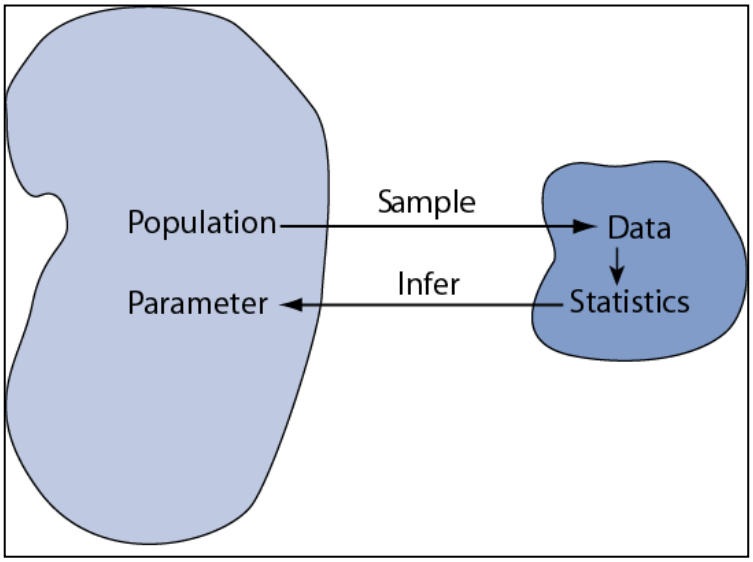
$$P(48 \le x \le 50) = P(48) + P(49) + P(50)$$

$$= \frac{e^{-5} 5^{48}}{48!} + \frac{e^{-5} \cdot 5^{49}}{49!} + \frac{e^{-5} \cdot e^{50}}{50!}$$

Time consuming & complex – Soln. - try to approximate Binomial distribution to Normal

Inferential Statistics

- Sampling
- Sample
- Random sampling
- Central Limit theorem



1 Lakh data sets – want to build model – liner regression – equation joining two variables – take sample – build model – analysis

Huge Data set – take sample – try to build model – analysis – sample inferences – over the entire population

Ex. Election results – exit polls – 40 Cr voters – not feasible – sample – various parts of India

Govt. Schemes – how many people are happy, Also happiness index levels

Drug industry – R&D – before manufacturing – Clinical Trials - testing on some population

Chemistry lab – characteristics of component – 100 ml sample – tests – properties

Cooking – Dish – sample – half cooked or fully cooked

For good outcomes – sample should not be biased – selection of sample from collection should be random – probability to take sample from every population should be same

Sampling size makes difference – Population size – 10L – sample – 100 → not reasonable – sample size should be reasonable – accurate & effective results

Sample data – should be authentic

Build a model – take a sample – build a model – before implementation – check whether model works well or not – test data – should be different from train data – random - should be picked from population – 10L → 1L train data, other than train data – 10k – test data

Statistical Inferences

Theory of statistical inference is divided into two major areas

- Estimation
- Tests of hypothesis

Estimation – once sampling is completed – infer it for entire population – based on sampling outcome estimate it for entire population – i.e. student data – 10k population – sample 1k – various tools – mean – standard deviation – marks – sampling is done – generalize - find for entire population – estimate average marks of 10k students

New vehicle – launched – average/mileage of vehicle – i.e. 25 kmpl – select few vehicles – test it under different conditions – calculate the average

Testing of hypothesis – claim – mileage of bike – 70 kmpl – Hypothesis test – correction of estimation

Go for sampling – sampling outcome & claim – check the validity of the claim - hypothesis

Hypothesis Testing

Goal:

Make statement(s) regarding unknown population parameter values based on sample data

Lec 6 Inferential statistics

Agenda

- Quick Review of the topics covered in previous class
- Testing of Hypothesis

For better sampling – Randomness of sample – Not biased selection

Nature of the population – Heterogeneous population – take state or entire country as population – suppose India – entire population is heterogeneous – have segments – which are homogenous – some proportional representation should be there in it – equal homogeneous clusters – i.e. class – upper, middle & lower – equally represented in sample

Size of sample matters – some proportion should be from population → 10L population – 100 sample – not justifiable → **Select ratio** – mathematical representation

Appropriate tools – **make sampling more accurate & reliable**

Actual analytics start from – Test of hypothesis

Once sampling is completed – generalize it – entire country – 10L students – sample – build a model – analysis – mean – 50 marks → generalize – estimate for entire population

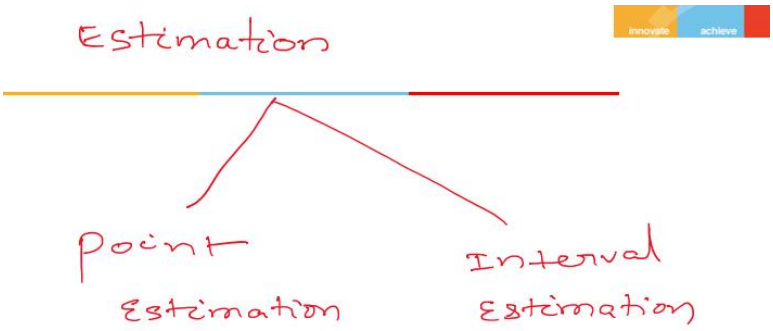
Appraisal – 5 point rating – 7 point rating → guess - 99% chance – confidence – probability

Inferential Statistics

- Sampling
- Sample
- Random sampling
- Central Limit theorem

Theory of statistical inference is divided into two major areas

- Estimation
- Tests of hypothesis

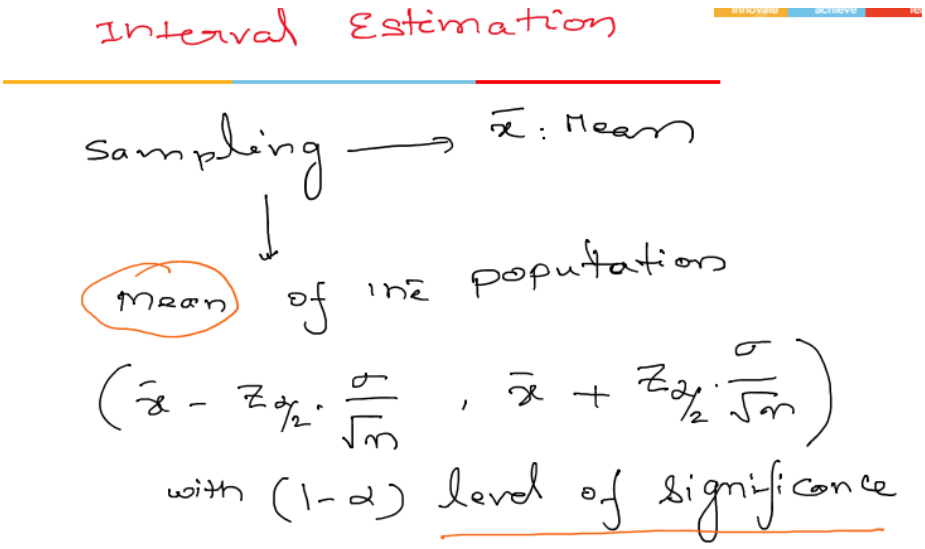


Point estimation – with point of confidence – 50% chance of this, 80% chance of that – relevance of the number – important to take the decision

I can score 80+ marks out of 100, Speculating promotion

If there's 99.99% chance or 50-50% chance – hike in salary – different – 99% - plan in advance - 50% we will wait until confirmation

Similarly – operation success – chances – or chances of failure



Mostly we do discussion related interval estimation, Mean of the population estimated based on sampling technique, **s.d.** of the population – if no clue – we can replace it by **s** – standard deviation of the sample

Based on distribution – we proceed with estimation

(1 – Alpha) – Level of confidence or level of significance → if Alpha is 95% → leftover is 5% → alpha/2 = 2.25% of the population – z*alpha / 2

Two Distribution – Z distribution – normal/standard normal/Gaussian distribution & T distribution – when to go for Z or T

Large Sample – Z distribution, Small sample (n < 30) – T distribution

T distribution – less skewed – in comparison to Normal distribution

T distribution – highly complicated integral – Need to calculation - distribution table – **degrees of the freedom** – n sample size → **n-1** →degrees of the freedom

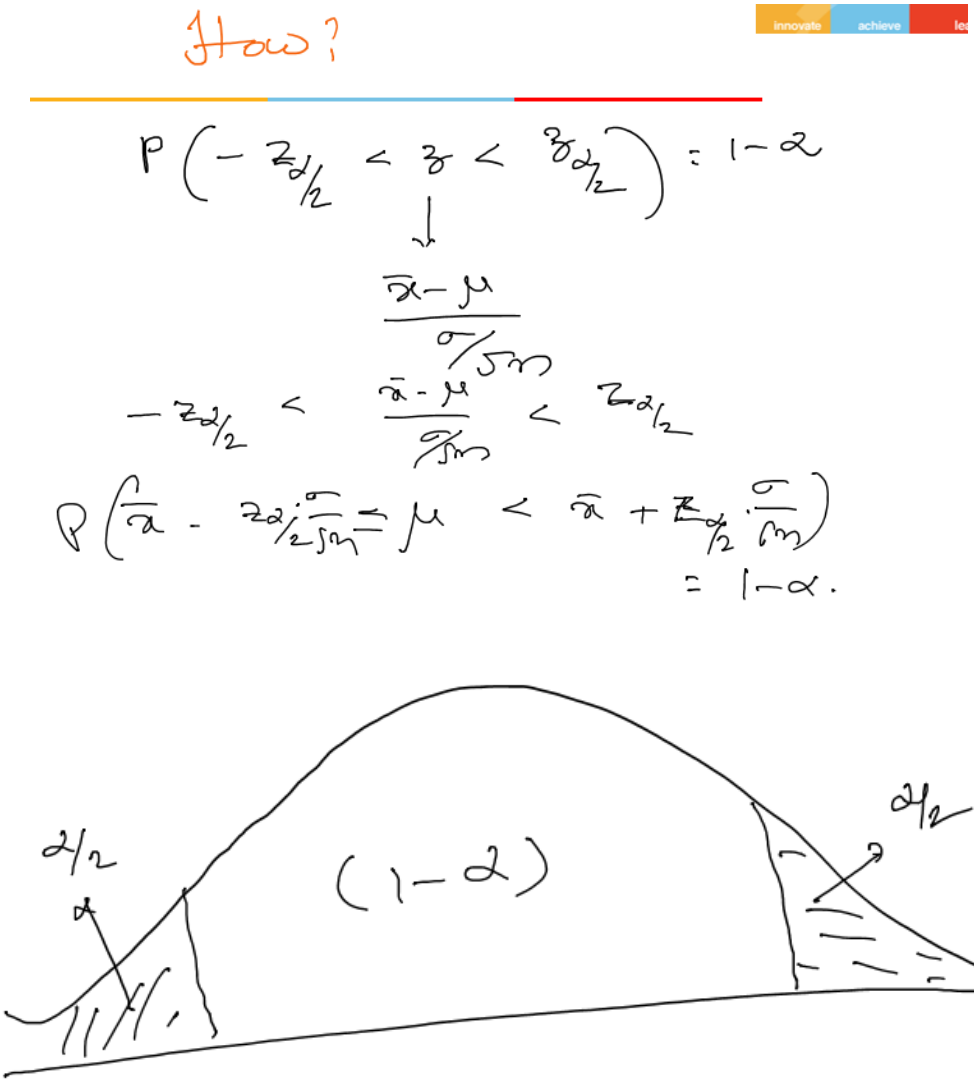
n < 30 – observed – skewness is not there – lesser comparison to bell curve – not like the normal curve or Gaussian curve – after 30 bell curve – 30 border points

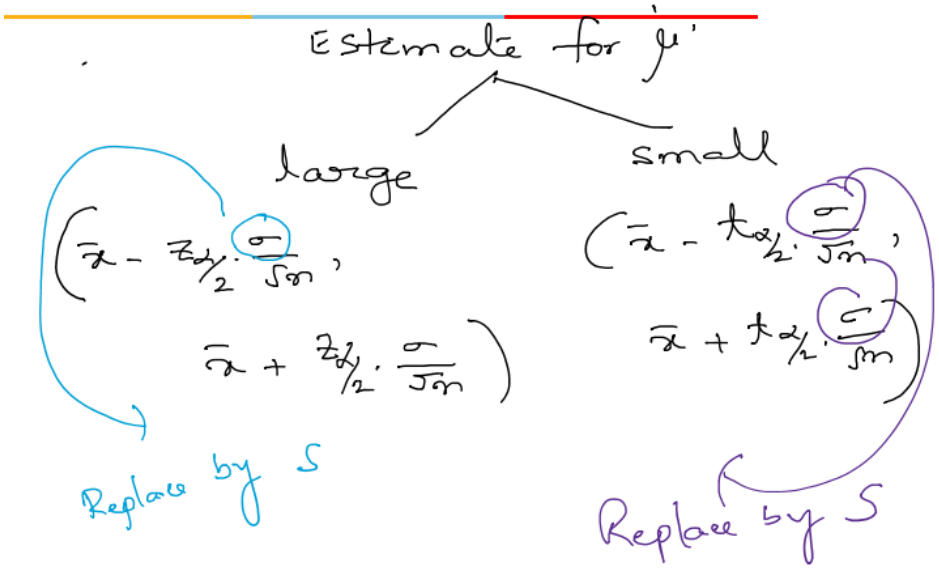
if x + y = 5 → what are the unique values of x & y that satisfy this equation – when we can have unique selection – if x = 3 (known or restricted) → one restricted is degree of the freedom – how many parameters we have to fix to find solution

Design of the system or simulation of the system – degree of the freedom is important

Assuming probability – **z** lies between **(-zα/2, zα/2)** is **1 - α**

In generic terms – **sampling mean** – **x** & **σ/√n**





Confidence = 95%, **α = 5%, α/2 = 2.25%**

Example:

A company wants to estimate the average life of the product. The S.D is known to be 100 hours. A random sample of 50 gave a sample average life of 350 hours.

Estimate the confidence interval for the mean.

Example:

A company wants to estimate the average life of the product. The S.D is known to be 100 hours. A random sample of 50 gave a sample average life of 350 hours.

Estimate the confidence interval for the mean.

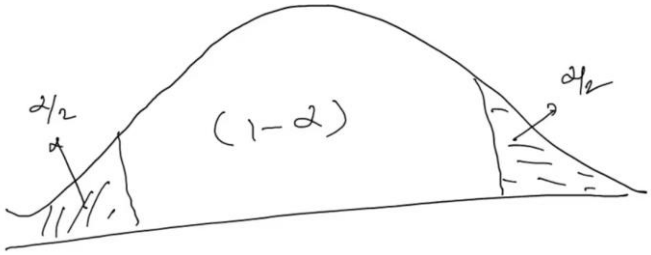
95% is LOS suppose

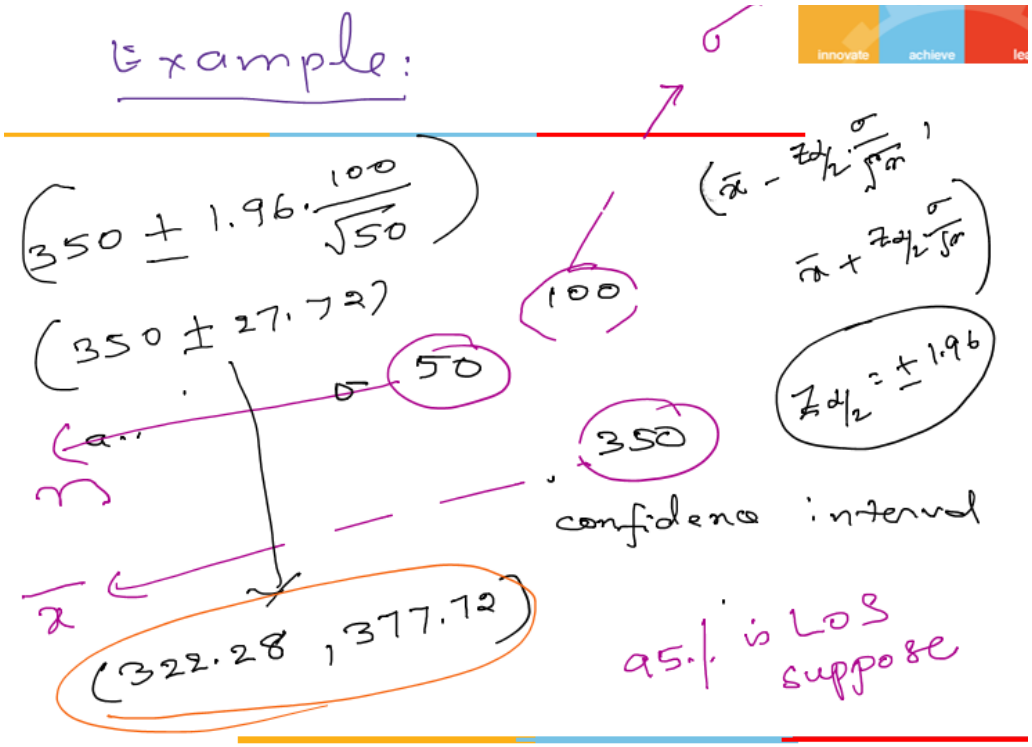
Average life of a product – sample – generalize it – take a call – before production

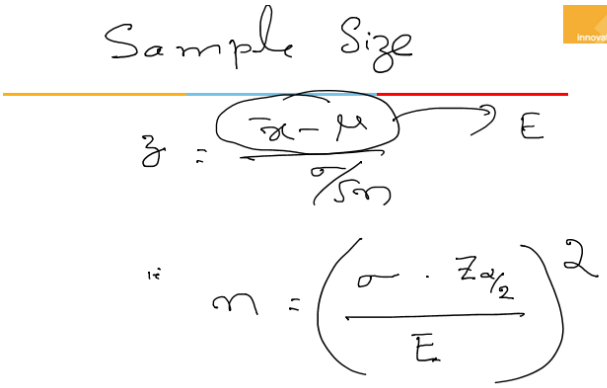
If bike production – sample – mileage 80 kmpl with s.d. → generalize – entire population estimation – almost same - we can proceed further

If sample – 50 kmpl → population – 20 kmpl – we have to check any flaw in sampling

α/2 = 2.25% = 0.025 – area under curve







With help of sample - x (mean) – we are estimating → Mu (Population) – difference – **error**

If we have error – limit – then we can decide sample size – difference – fix – sample size est.

Hypothesis Testing

Goal:

Make statement(s) regarding unknown population parameter values based on sample data

Hypothesis Testing

- ✓ Is also called *significance testing*
- ✓ Tests a claim about a parameter using evidence (data in a sample

Example

Drug company has new drug, wishes to compare it with current standard treatment

Federal regulators tell company that they must demonstrate that new drug is better than current treatment to receive approval

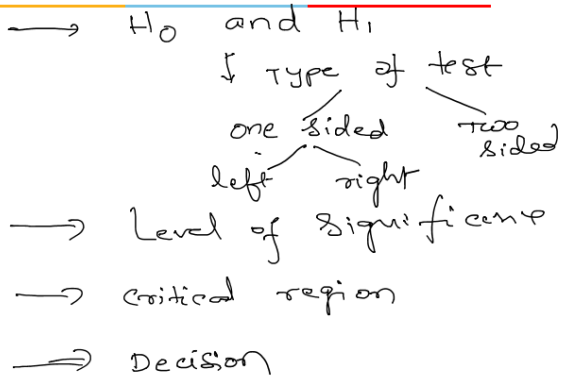
Firm runs clinical trial where some patients receive new drug, and others receive standard treatment

Numeric response of therapeutic effect is obtained (higher scores are better).

Parameter of interest: $m_{New} - m_{Std}$

Hypothesis Testing Steps

- ➢ Null and alternative hypotheses
- ➢ Test statistic
- ➢ P-value and interpretation
- ➢ Significance level (optional)



Null hypothesis – is it related to mean? Is it related with proportion? Is it related to various observations/values- accordingly to we have to choose statistics – mean related discussion – mean related distribution & mean related tools available

Depending on Null hypothesis – choose alternative – if claim failed then what is alterative

LOS – how much allowed in our system – entire discussion at which level

Claim – Product is the best – if test fails – product is not the best – alternative

Example

Null hypothesis H_0: $\mu = 170$

The **alternative hypothesis** can be either H_1: $\mu > 170$ **(one-sided test)**

or

H_1: $\mu \neq 170$ **(two-sided test)**

A. Hypotheses:

H_0: μ = 100 versus

H_a: μ > 100 (one-sided)

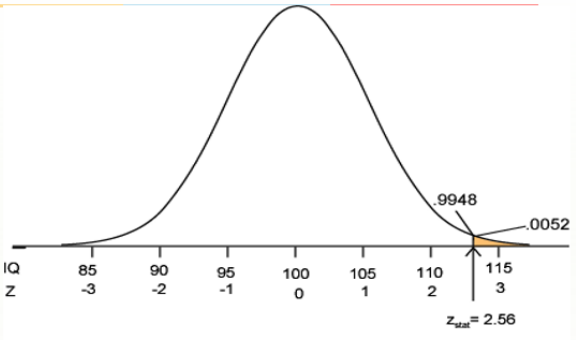
H_a: μ ≠ 100 (two-sided)

B. Test statistic:

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{9}} = 5$$

$$z_{\text{stat}} = \frac{\bar{x} - \mu_0}{SE_{\bar{x}}} = \frac{112.8 - 100}{5} = 2.56$$

C. *P*-value: $P = \Pr(Z \geq 2.56) = 0.0052$



P =.0052 ⇒ it is unlikely the sample came from this null distribution ⇒ strong evidence against H_0

Hypothesis Testing

Test Result –	H_0 True	H_0 False
True State H_0 True	Correct Decision	Type I Error
H_0 False	Type II Error	Correct Decision

$\alpha = P(Type\ I\ Error)$ $\beta = P(Type\ II\ Error)$

- Goal: Keep α, β reasonably small

Problem

A trucking firm is suspicious of the claim that the average life time of certain tyres is at least 28,000 miles. To check the claim, the firm puts 40 of these tyres on its trucks and get a mean life of 27,463miles with a standard deviation of 1,348 miles. What can it conclude if the probability of Type I error is to be at most 0.01

Claim validity, tool used here – average/mean

Claim - Smokers – prone to lung cancer, Regular students – good performance

Claim has to nothing do with mean or value, Validity of claim - Sampling – conclusion

Solution

1.Null hypothesis : H_0 : $\mu \geq 28{,}000$ miles

2. Alternate hypothesis: H_1: $\mu < 28{,}000$ miles

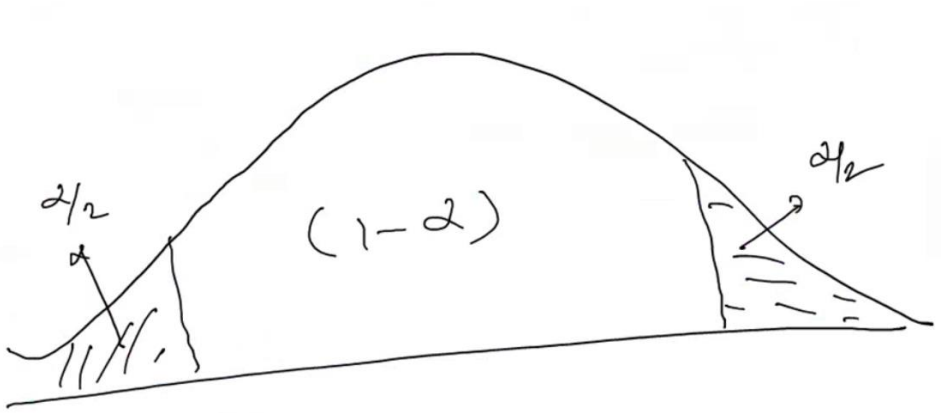
Claim = Assumption = Hypothesis

Level of significance – depends on system design – how much error allowed

Civil Engineering application – pipe – leakage – 5% of chance – we can fix it,

Biomedical Application – 5% is big error – we can't go it – take decision call – LOC, LOS

Based on alternative hypothesis → here μ < 28000 – lies one side → one tailed test, if we have μ = 28000, alternative μ != 28000 →μ < 28000 & μ > 28000 – two tailed test → Critical region – it defines **α/2** – Critical region or region of acceptance



Two tailed test – right & left test → **α/2,** one tail test – either left/right – **α**

Left side area – region of rejection, Right side area – region of acceptance

3. Level of significance: $\alpha = 0.01$

4. Critical region

This is a left tailed test (large sample)

If $Z = Z_{cal} < -Z_{\alpha}$ we reject null hypothesis

If $Z = Z_{cal} < -Z_{\alpha} = -Z_{0.01} = -2.33$ we reject null hypothesis

5.Computation

Test statistic

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{27,463 - 28,000}{\frac{1,348}{\sqrt{40}}} = -2.52$$

6.Conclusion

Since $Z = Z_{cal} = -2.52 < -2.33$, we reject null hypothesis at level of significance 0.01. In other words the trucking firm's suspicion that $\mu < 28,000$ miles is confirmed.

-2.52 in **region of rejection** – we have to reject Null Hypothesis – claim is wrong

Hypothesis concerning one mean (small sample)

Procedure

1. Null hypothesis $H_0 : \mu = \mu_0$

2.Alternate Hypothesis $H_1 : \mu \neq \mu_0$ (Two tailed test)

Or

$H_1 : \mu > \mu_0$ (Right tailed test)

Or

$H_1 : \mu < \mu_0$ (left tailed test)

3. Level of significance : α

4. Critical region

For two tailed test $H_1 : \mu \neq \mu_0$

Reject H_0 if $t < -t_{\frac{\alpha}{2}}$ or

$t > t_{\frac{\alpha}{2}}$ with (n-1) degrees of freedom

For right tailed test $H_1 : \mu > \mu_0$

Reject H_0 if $t > t_{\alpha}$ with (n-1) degrees of freedom

For left tailed test $H_1 : \mu < \mu_0$

Reject H_0 if $t < -t_{\alpha}$ (n-1) degrees of freedom

5. Test statistic

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$ with (n-1) degrees of freedom

6. Calculation

7. Decision

Problem

It is claimed that a random sample 49 tyres has a mean life of 15200 kms. This sample was drawn from a population whose mean is 15150 kms and a standard deviation of 1200kms. Test the significance at 0.05 level.

Solution:

1. Null hypothesis $H_0 : \mu = 15200$
2. Alternate hypothesis $H_1 : \mu \neq 15200$
3. Level of significance $\alpha = 0.05$
4. critical region :- This is a two tailed test (large sample). So reject H_0 if $(Z_{cal} = Z) < -Z_{\frac{\alpha}{2}}$ or $(Z = Z_{cal}) > Z_{\frac{\alpha}{2}}$

Here $\alpha = 0.05$

$$\frac{\alpha}{2} = \frac{0.05}{2}$$

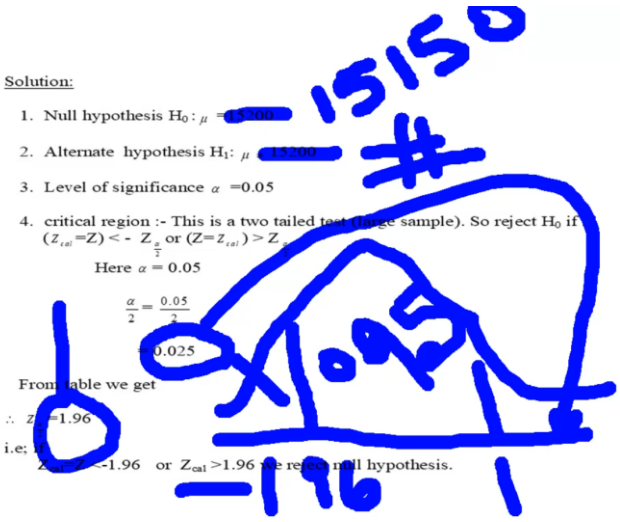
$$= 0.025$$

From table we get

$\therefore Z_{\frac{\alpha}{2}} = 1.96$

i.e; if

$Z_{cal} = Z < -1.96$ or $Z_{cal} > 1.96$ we reject null hypothesis.



6. Computation :

Test statistic

$$Z_{cal} = Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{15200 - 15150}{\frac{1200}{\sqrt{49}}}$$

$$= 0.2916$$

7. Decesion:

Since $Z_{cal} = 0.2916 < 1.96$ we accept the mull hypothesis.

First check hypothesis for 90% - pass → 95% → pass → 99% - confidence

We can change the claim & confidence → tool available – check → Accept or reject

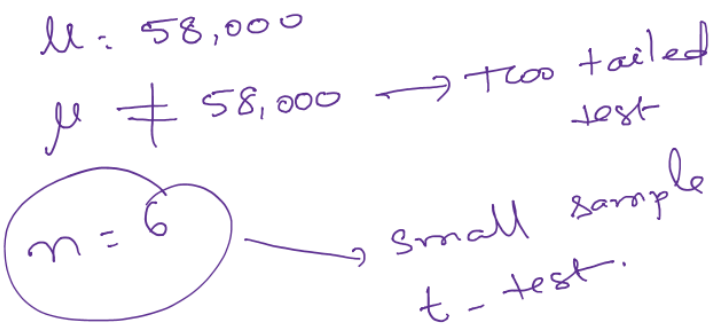
Example:

A random sample of 6 steel beams has a mean compressive strength of 58,392 p.s.i (pounds per square inch) with a standard deviation of 648 p.s.i . use this information at the level of significance $\alpha = 0.05$ to test

whether the true average compressive strength of steel from which the sample came is 58,000 p.s.i

A random sample of 6 steel beams has a mean compressive strength of 58,392 p.s.i (pounds per square inch) with a standard deviation of 648 p.s.i . use this information at the level of significance $\alpha = 0.05$ to test

whether the true average compressive strength of steel from which the sample came is 58,000 p.s.i



L- 7: Inferential statistics & Predictive Analytics

Agenda

- Central limit theorem
- Type I, Type II Errors
- Testing of Hypothesis – continuation from previous session
- Covariance
- Correlation
- Introduction to regression

Central Limit Theorem

If $\bar{X}$ is the mean of a sample of size n taken from a population having the mean μ and variance σ^2, then

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$ is a random variable whose distribution function approaches that of the standard normal distribution as $n \to \infty$.

Z → related to Standard Normal Distribution, Central limit theorem allows us to use normal distribution irrespective nature of the distribution – if sample is known & sample size is large

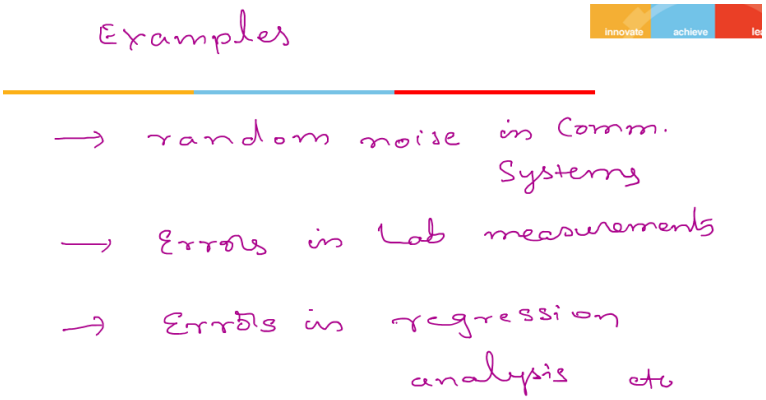
the mean μ and variance σ, then

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$ is a random variable

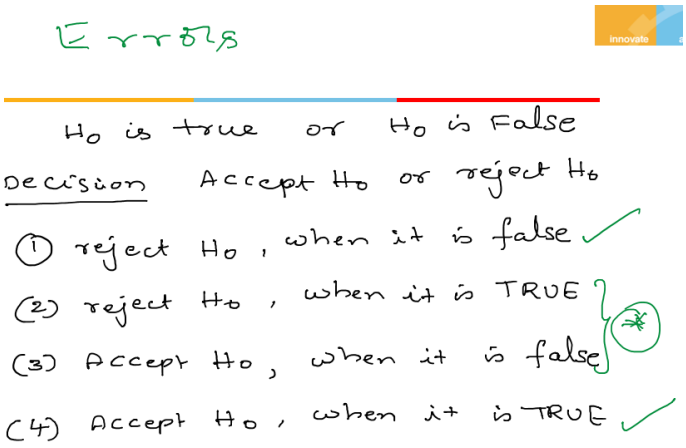
→ It does not matter what the distribution of X_i's is

→ In many real applications, the random variable is a sum of independent random variables. In all such cases, CLT helps to use normal distribution.

Sampling distribution follows Normal distribution irrespective of nature of the population distribution – for large sample → n > 30

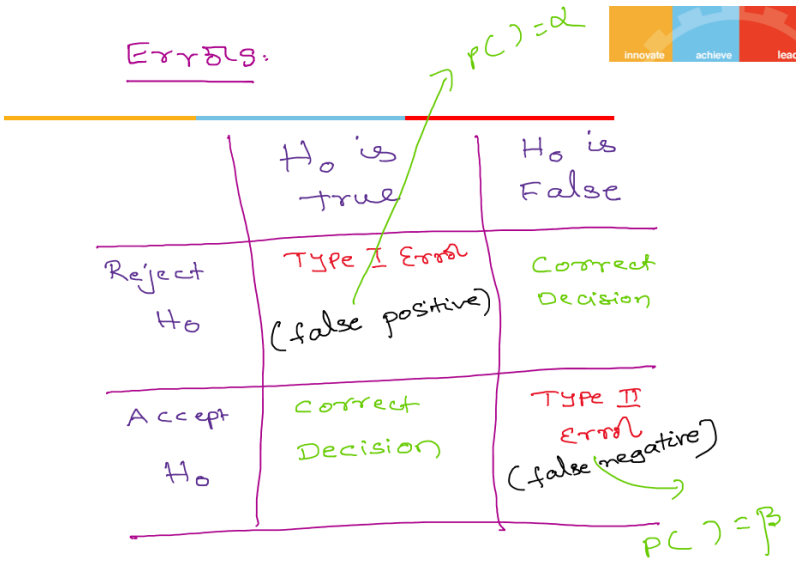


Lab experiment – errors – discrete in nature – large sample – we can use Normal



Sample outcome – validate – Null Hypothesis → 1) & 4) – decision is right

2) & 3) – important → 2) more important – hypothesis – fever – Doctor – tests – no fever – rejecting claim when it is true – **False positive (Type I) – error** is there still we are rejecting
3) Not having fever – Doctor – test – have fever – **True Negative (Type II)**



P (Type I error) → Alpha – LOS – level of significance → False Positive → very important – even though it's true – rejecting it → we can have another level of testing – before rejection

P (Type II error) → Beta – LOC – level of confidence – True Negative

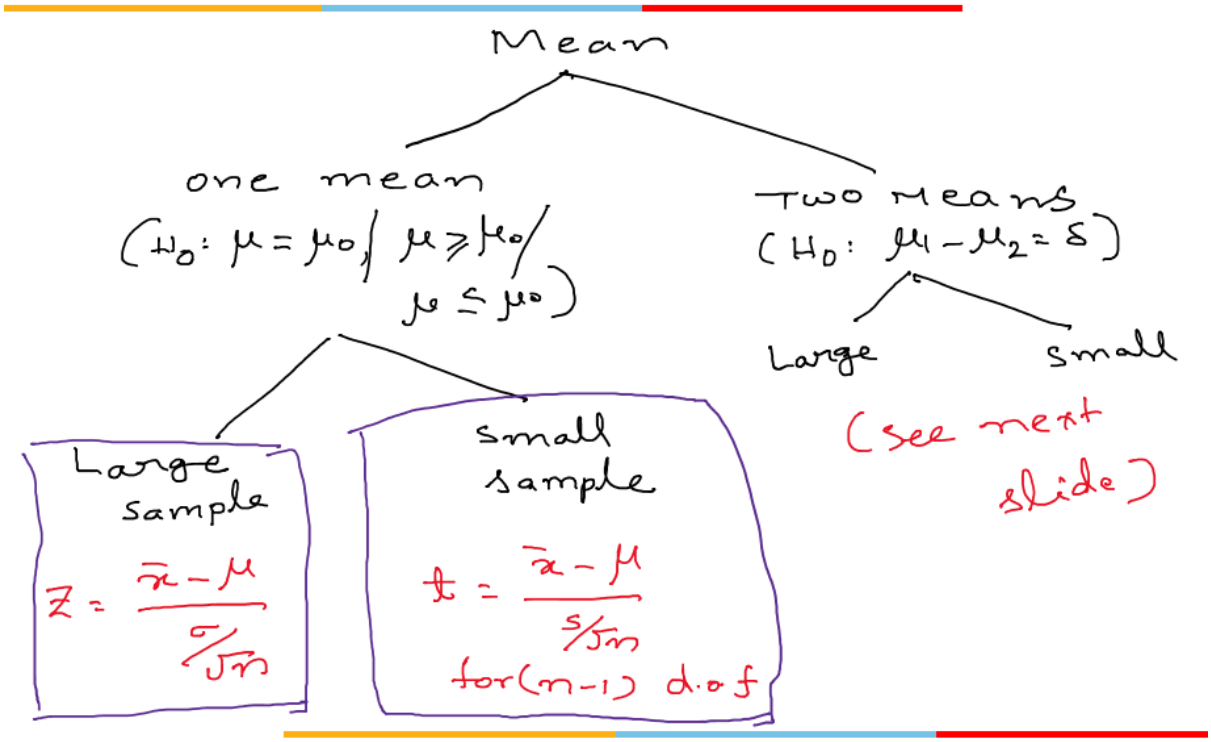
Decrease Alpha → Beta increases – vice & versa → determine the power of the adjust **α** & **β**

Testing of Hypothesis

or

Hypothesis testing

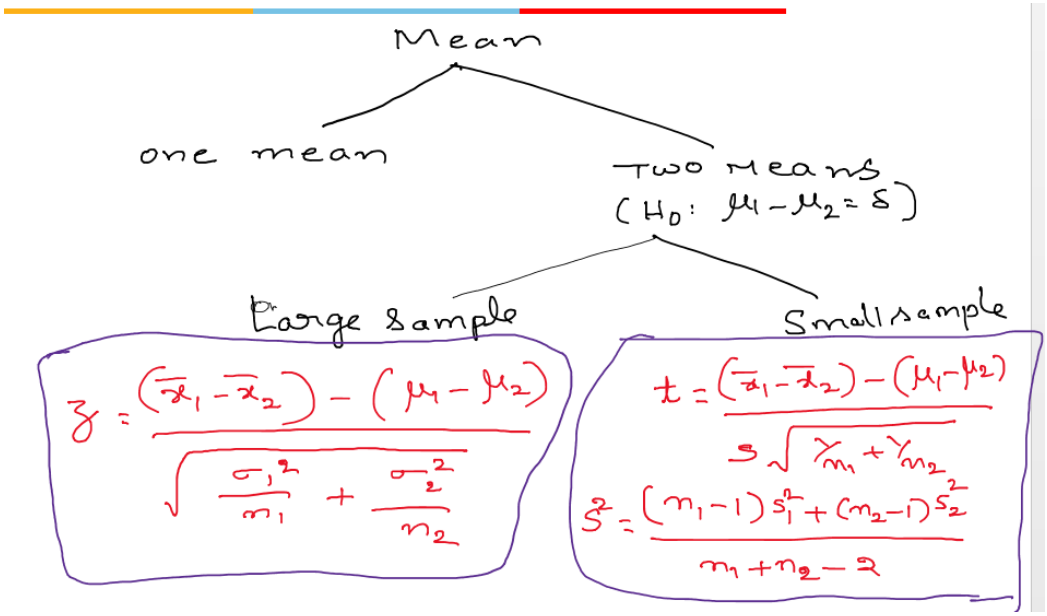
Claim related mean – two means or one mean



x – Mean of the sample, **μ** – mean of the population, **σ** – standard deviation of the population, if σ – no clue → replace σ with **s**

First we have to check is it mean related – yes – one mean or two mean → large or small sample – z or t distribution

In this case – two means - Degrees of the freedom → n1+n2 – 2



Testing of Hypothesis

Example - 1

Example :- (1)

Can it be concluded that the average life span of Indians is more than 70yrs. If a random sample of 100 Indians has average life span of 71.8 years with a S.D of 8.9 years.

One mean or two mean → here one mean – large sample

Example :- (contd)

Can it be concluded that the average life span of Indians is more than 70yrs. If a random sample of 100 Indians has average life span of 71.8 years with a S.D of 8.9 years.

Sample
$n = 100$
$\bar{x} = 71.8$
$s = 8.9$

Validation

$H_0 : \mu > 70$

↓ population

Can it be concluded that the average life span of Indians is more than 70yrs. If a random sample of 100 Indians has average life span of 71.8 years with a S.D of 8.9 years.

→ one mean problem

→ $n = 100$: Large sample, so Z-test

∴ $Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ or $\frac{\bar{x} - \mu}{s/\sqrt{n}}$

$H_0 : \mu > 70$

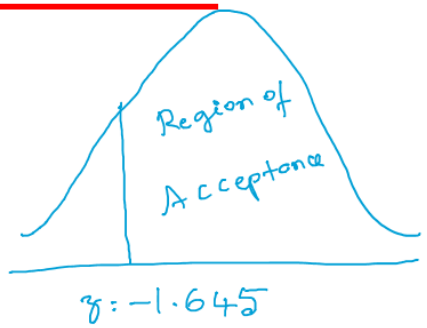
$H_1 : \mu \leq 70$ → Left tailed test

$\alpha = 5\%$ (Let)

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{71.8 - 70}{8.9/\sqrt{100}}$$

$= 2.022$

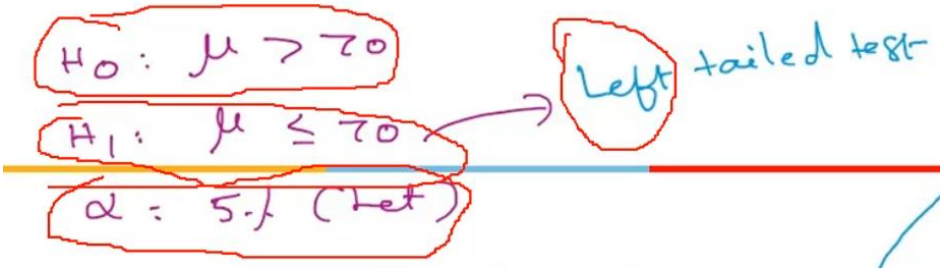
Lies in the region of acceptance



∴ H_0 is accepted

i.e Avg life is more than 70 years

LOS – not given – we are taking 5% → use – gives critical region where z value lies

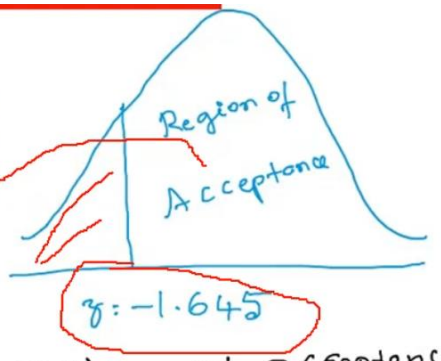


Large Sample – **Z** distribution → no degree of freedom

$\alpha = 5\%$ (Let)

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{71.8 - 70}{8.9/\sqrt{100}}$$

$$= 2.022$$



Lies in the region of acceptance

∴ H_0 is accepted

i.e Avg life is more than 70 years

Testing of Hypothesis

Example – 2.

Example - 2

A machine which produces Mica insulating washers for use in electronic devices said to have a thickness of 10mm.

A sample of 10 washers has an average thickness of 9.52 mm with a S.D of 0.6mm. Whether the sample is drawn from the given population

(use 5% Level of significance)

Example - 2 (small sample)

A machine which produces Mica insulating washers for use in electronic devices said to have a thickness of 10mm (mean). A sample of 10 washers has an average thickness of 9.52 mm with a S.D of 0.6mm. Whether the sample is drawn from the given population (use 5% Level of significance)

$\bar{x}$ → 9.52 mm

S → 0.6 mm

$H_0 : \mu = 10$

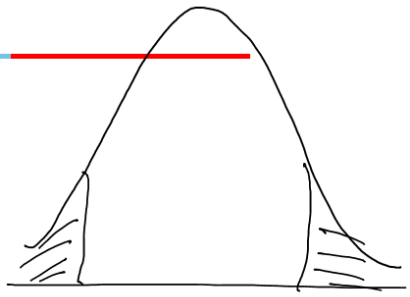
$H_1 : \mu \neq 10$

$\alpha = 0.05$

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

$$= \frac{9.52 - 10}{0.6/\sqrt{10}}$$

$$= -2.52$$



→ Reject H_0

$H_0 : \mu = 10$

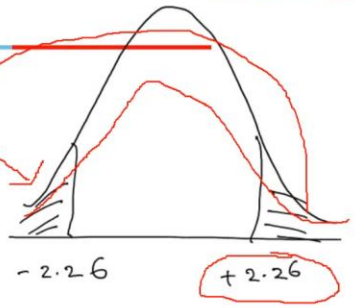
$H_1 : \mu \neq 10$

$\alpha = 0.05$

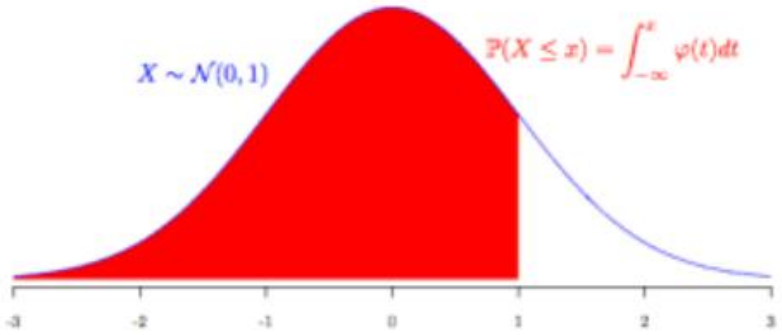
$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}}$$

$$= \frac{9.52 - 10}{0.6/\sqrt{10}}$$

$$= -2.52$$

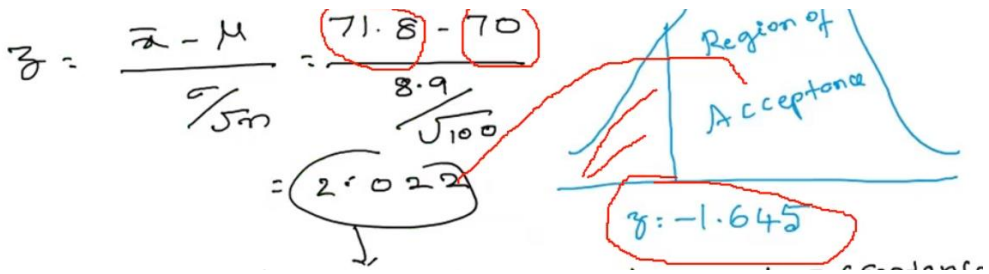


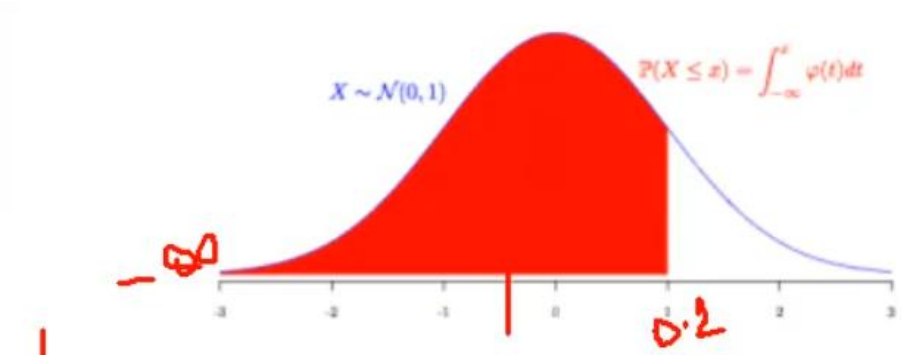
→ Reject H_0



	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Testing of hypothesis – probability is known – find z value – reverse





	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133

Testing of Hypothesis

Example – 3

(consider this example (ignore the one given in slides)

Example 3 – modified problem

A random sample of 40 items produced by a company A have a mean life time of 647 hours with S.D 27 hours. While a sample of 40 items by company B has a mean life time of 638 hours with S.D of 31 hours.

Does this substantiate the claim of the company A that their items are ~~superior to those~~ produced by company B in same as those terms of mean life

If one sample is large i.e. 35, other one is small i.e. 15 → n1+n2 – 2 = 50 – 2 = 48 → **z**

Example 3 – Solution ?

A random sample of 40 items produced by a company A have a mean life time of 647 hours with S.D 27 hours. While a sample of 40 items by company B has a mean life time of 638 hours with S.D of 31 hours.

Does this substantiate the claim of the company A that their items are ~~superior to those~~ same as those produced by company B in terms of mean life

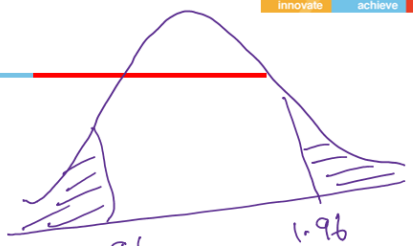
i.e. $\mu_1 = \mu_2$

Solution :-

$H_0: \mu_1 - \mu_2 = 0$

$H_1: \mu_1 - \mu_2 \neq 0$

$\alpha = 0.05$



$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(647 - 638) - 0}{\sqrt{\frac{(27)^2}{40} + \frac{(31)^2}{40}}}$$

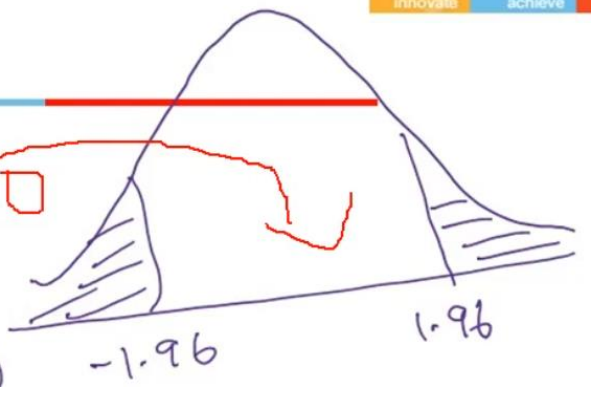
$= 3.73$ → Reject H_0

Solution :-

$H_0: \mu_1 - \mu_2 = 0$

$H_1: \mu_1 - \mu_2 \neq 0$

$\alpha = 0.05$



Example 3 — Earlier problem as it is

A random sample of 40 items produced by a company A have a mean life time of 647 hours with S.D 27 hours. While a sample of 40 items by company B has a mean life time of 638 hours with S.D of 31 hours.

Does this substantiate the claim of the company A that their items are superior to those produced by company B.

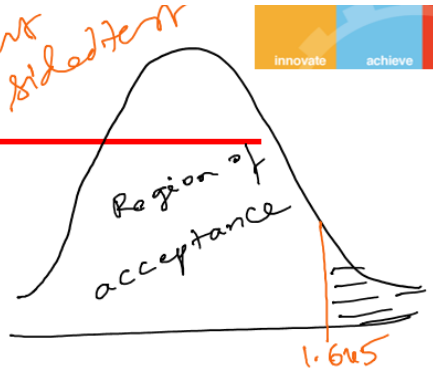
i.e $\mu_1 > \mu_2$

Solution:-

$H_0: \mu_1 - \mu_2 = 0$

$H_1: \mu_1 - \mu_2 > 0$

$\alpha = 0.05$



$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(647 - 638) - 0}{\sqrt{\frac{(27)^2}{40} + \frac{(31)^2}{40}}}$$

$= 3.73$

$3.73 > 1.645$ → Reject H_0

i.e Accept alternate Hypothesis.

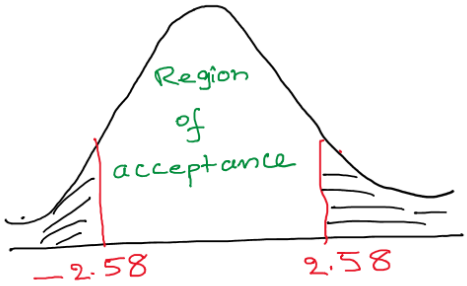
Critical region for some 'α's (Z-distribution)

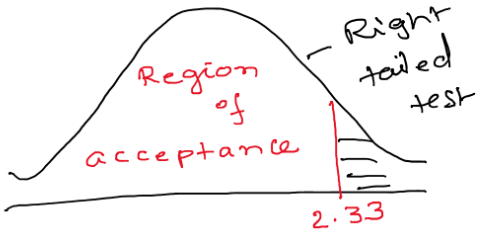
useful table values (Z- distribution)

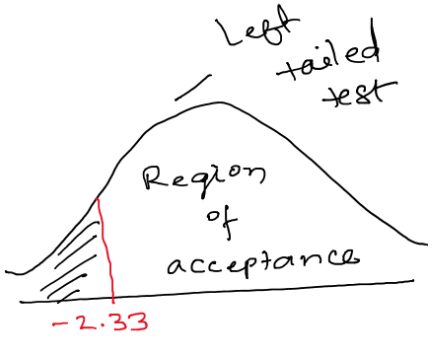
	Level of significance		
	0.01	0.05	0.1
Two - tailed test	± 2.58	± 1.96	± 1.645
Right tailed test	2.33	1.645	1.28
Left tailed test	-2.33	-1.645	-1.28

$\alpha = 1\%$ (or 0.01)

Two tailed test

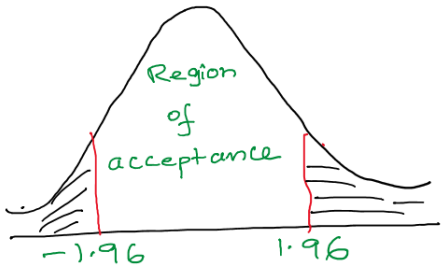


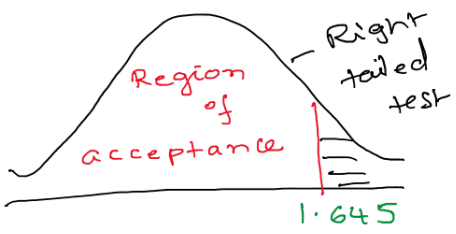


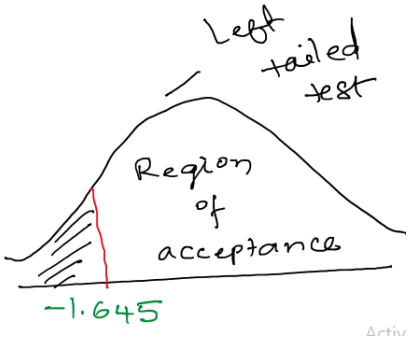


$\alpha = 5\%$ (or 0.05)

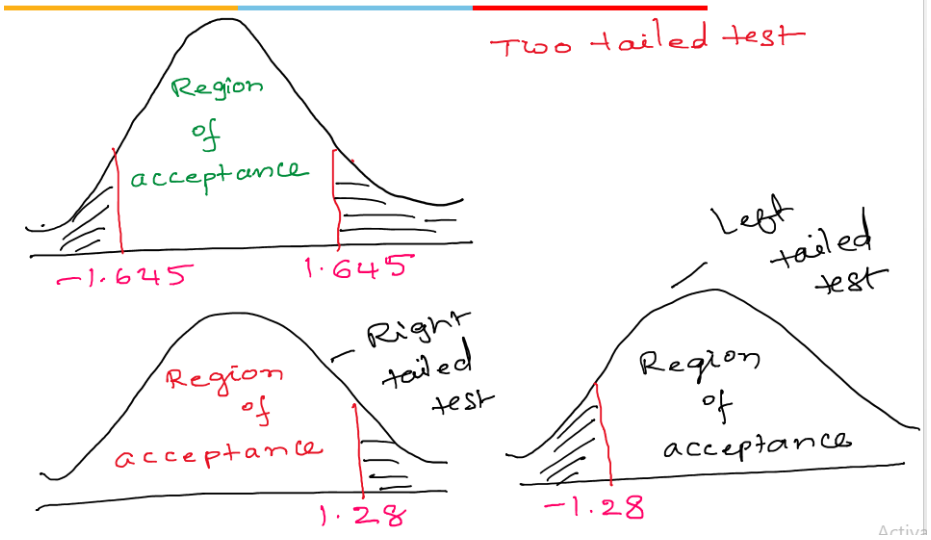
Two tailed test







$\alpha = 10\%$ (or 0.1)



t distribution – critical region not fixed – degrees of freedom - based on size of the sample

Note :-

In case of 't' distribution, the critical region depends on the degrees of freedom ($\nu = n-1$, where n is the size of the sample)

Testing of Hypothesis

Example – 4

Example- 4 :-

A Company believes that the advertisement A is more effective than advt. B. To test this sampling is done.

In a random sample of 60 customers who saw advertisement A, 18 tried the product. In a random sample of 100 customers, who saw advt B, 22 tried the product.

Does this indicate that advt A is more effective than advt B.

Nothing like as mean – output – also not mean

Example- 4 :-

A Company believes that the advertisement A is more effective than advt. B. To test this sampling is done.

In a random sample of 60 customers who saw advertisement A, 18 tried the product. In a random sample of 100 customers, who saw advt B, 22 tried the product.

Does this indicate that advt A is more effective than advt B.

Sample A: 18 out of 60
Sample B: 22 out of 100
} Advt (A) > Advt (B) ????

Testing of Hypothesis

Example – 5

Example:

consider the following data

Travel time	Stress: High	Moderate	Low	Total
< 20 min	9	5	18	32
20 - 50 min	17	8	28	53
> 50 min	18	6	7	31
Total	44	19	53	116

Proportions – out of 100 → 80

Based on this data, Can we conclude that stress levels depends on travel time

? ? ?

1 proportion, two proportions, several proportions → very low, low, moderate etc.

If we observe examples (4) & (5), the hypothesis / claim is not related to mean.

We can observe that the hypot, is based on proportion

Sample A: 18 out of 60
Sample B: 22 out of 100
} ⇒ Hypothesis is Sample A > Sample B

$\frac{x}{n} = \bar{P} \longrightarrow P_0$

One proportion: (large sample)

$H_0: P = P_0 \,/\, P \geq P_0 \,/\, P \leq P_0$

$$Z = \frac{\bar{P} - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}$$

Two proportions: (large sample)

$H_0: P_1 - P_2 = \delta \,/\, P_1 - P_2 \geq \delta \,/\, P_1 - P_2 \leq \delta$

$$Z = \frac{(\bar{P}_1 - \bar{P}_2) - \delta}{\sqrt{\bar{P}(1-\bar{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad \text{where } \bar{P} = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2}$$

One Proportion – large or small, Two – large or small

Example- 4 :-

A Company believes that the advertisement A is more effective than advt. B. To test this sampling is done.

In a random sample of 60 customers who saw advertisement A, 18 tried the product. In a random sample of 100 customers, who saw advt B, 22 tried the product.

Does this indicate that advt A is more effective than advt B.

Sample A: 18 out of 60
Sample B: 22 out of 100
} Advt (A) > Advt (B) ????

Sample A : 18 out of 60

Sample B : 22 out of 100

$\bar{P}_1 = \frac{18}{60} = 0.3$, $\bar{P}_2 = \frac{22}{100} = 0.22$ Let α : 5%

$H_0 : P_1 - P_2 = 0$

$H_1 : P_1 > P_2$

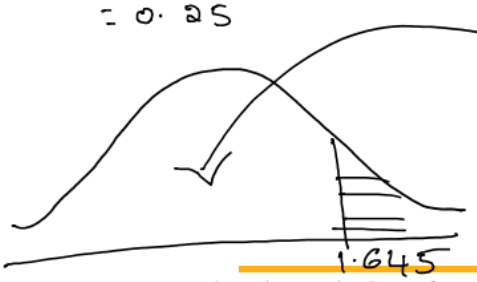
$$\bar{P} = \frac{n_1\bar{P}_1 + n_2\bar{P}_2}{n_1 + n_2}$$

$= 0.25$

$$z = \frac{(\bar{P}_1 - \bar{P}_2) - \delta}{\sqrt{\bar{P}(1-\bar{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$= \frac{(0.3 - 0.22) - 0}{\sqrt{(0.25)(0.75)\left(\frac{1}{60} + \frac{1}{100}\right)}}$$

$= 1.131$



Accept H_0

∴ no significant diff in the effectiveness of advertisements

Example:

consider the following data

Travel time	Stress: High	Moderate	Low	Total
< 20 min	9	5	18	32
20-50 min	17	8	28	53
> 50 min	18	6	7	31
Total	44	19	53	116

Based on this data, Can we conclude that stress levels depends on travel time

? ? ?

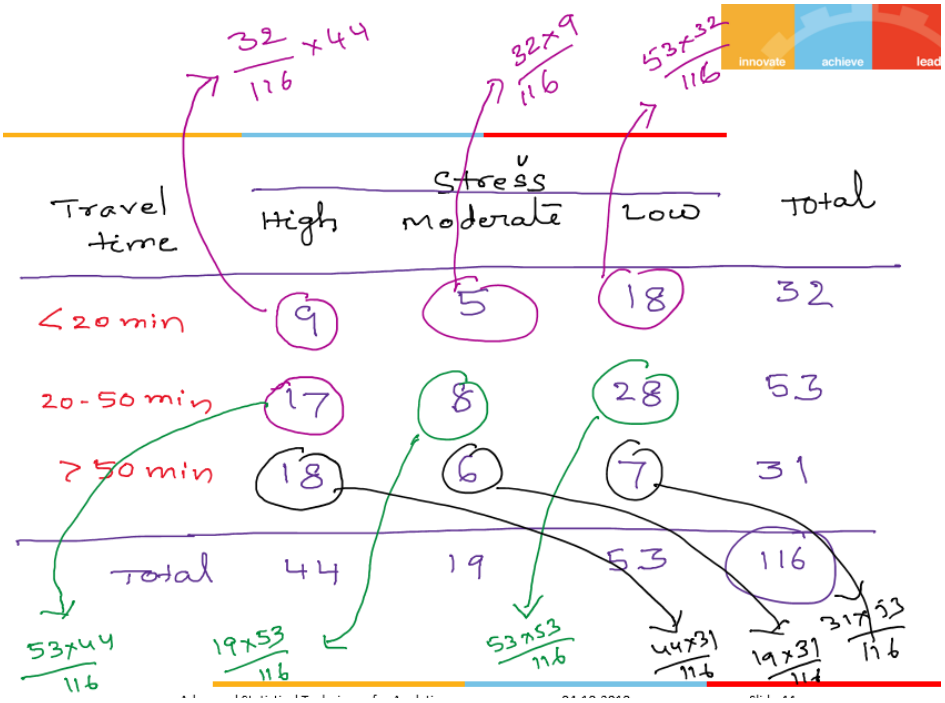
Chi-Square (χ^2) distribution

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

O: observed frequencies

E: Expected frequencies

for $(r-1) \times (c-1)$ degrees of freedom



$$\chi^2 = \sum \frac{(O-E)^2}{E} = \sum \frac{(9-12.14)^2}{12.14} + \dots$$

$$= 9.836$$

Travel time	Stress: High	Moderate	Low	Total
< 20 min	9/12.14	5/5.24	18/14.62	32
20-50 min	17/20.10	8/8.68	28/24.22	53
> 50 min	18/11.75	6/5.08	7/14.17	31
Total	44	19	53	116

Calculated $\chi^2 = 9.836$

Let $\alpha = 0.01$

d.o.f: $(r-1) \times (c-1)$

$= (3-1) \times (3-1) = 4$

$\chi^2_{0.01,4} = 13.30$

$\chi^2_{cal} = 9.836 < 13.30$

H_0 accepted

Example

A tobacco company claims that there is no relationship between smoking and lung ailments. ↓ H_0

	Lung ailment	Non-lung ailment	Total
Smokers	75	105	180
Non-smokers	25	95	120
	100	200	300

Based on this data, can we accept/reject the claim

observed frequency

$E = \frac{180}{300} \times 100 = 60$

$E = \frac{180}{300} \times 200 = 120$

	Lung ailment	Non-lung ailment	Total
Smokers	75	105	180
Non-smokers	25	95	120
	100	200	300

$\frac{120}{300} \times 100 = 40$

$\frac{120}{300} \times 200 = 80$

$$\chi^2 = \frac{(75-60)^2}{60} + \frac{(105-120)^2}{120} + \cdots$$

$$= 14.063$$

From χ^2-tables at 0.05 LOS

d.o.f $= (r-1) \times (c-1)$
$= (2-1) \times (2-1)$
$= 1$

$= 3.841$

	Lung ailment	Non-lung ailment	Total
Smokers	75 (60)	105 (120)	180
Non-smokers	25 (40)	95 (80)	120
	100	200	300

$\chi^2 = 14.063 > 3.841$

Reject H_0 $\rightarrow$

Correlation

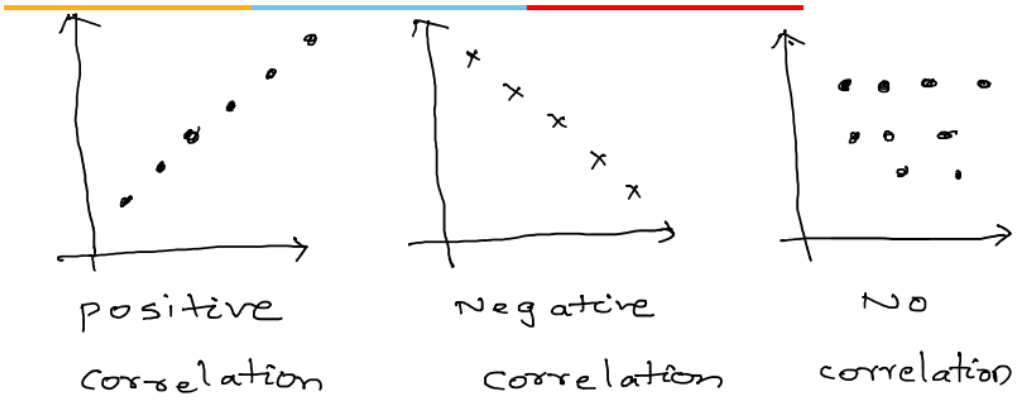
Correlation

- $\rightarrow$ Sales of a company and Expenditure on advertisement
- $\rightarrow$ Price and Demand of a product
- $\rightarrow$ Inflation and Gold Price
- $\rightarrow$ IQ and performance in Entrance.

Relation between two properties – height & weight → increases together

Income of the family & Nutrition level – free meals scheme in school

Recruitment & College rank – Tier 1, Tier 2 etc. → relation between these two



Positive – one variable increases – another also increases, Negative – one increases where another decreases

Positive - Height – Weight, Salary – Expenditure, Negative - Salary – Poverty, Age - Memory

No correlation – experience & performance of teacher

Coefficient of Correlation

$r = 1 \Rightarrow$ Perfect and positive relation

$r = -1 \Rightarrow$ " " negative relation

$r = 0 \Rightarrow$ No relation

$0 < r < 1 \Rightarrow$ Partial positive relation

$-1 < r < 0 \Rightarrow$ " negative "

$$r = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{\Sigma dx\, dy}{\sqrt{\Sigma dx^2 \, \Sigma dy^2}}$$

where $dx = x - \bar{x}$

$dy = y - \bar{y}$

$dx^2 = (x - \bar{x})^2$

$dy^2 = (y - \bar{y})^2$

Example - 1

X	1	2	3	4	5	6	7	8	9
Y	10	11	12	14	13	15	16	17	18

$$\bar{x} = \frac{\Sigma x}{n} = \frac{45}{9} = 5$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{126}{9} = 14$$

Whether they are correlated or not → how correlated & what is one value – another given

Correlation – two variables – are related or not, relation - Regression – prediction of value

X	dx = x-5	dx^2	Y	dy = y-14	dy^2	dxdy
1	-4	16	10	-4	16	16
2	-3	9	11	-3	9	9
3	-2	4	12	-2	4	4
4	-1	1	14	0	0	0
5	0	0	13	-1	1	0
6	1	1	15	1	1	1
7	2	4	16	2	4	4
8	3	9	17	3	9	9
9	4	16	18	4	16	16
		60			60	59

$$r = \frac{\Sigma dx\, dy}{\sqrt{dx^2 \Sigma dy^2}} = \frac{59}{\sqrt{60 \times 60}} = 0.9833$$

Coefficient of Determination

r is coeff. of Correlation

r^2 is coeff of determination

↓

Indicates the extent to which variation in one variable is explained by the variation in the other.

$r = 0.9 \Rightarrow r^2 = 0.81$

i.e. 81% of the variation in y due to variation in 'x'

remaining 19% is due to some other factors.

Regression :-

x	1	2	3	4	5
y	1	4	9	16	25

when x : 7 : y = ?

x	1	2	3	4	5
y	1	6	2	5	4

when x : 7, y : ?

Lec 8 Predictive Analysis

L- 8: Predictive Analytics

Agenda

- Covariance
- Correlation
- Introduction to regression
- Method of least squares
- Simple linear regression

Covariance of X and Y

$$Cov(X,Y) = E\left[(X-\mu_X)(Y-\mu_Y)\right]$$

$$= \sum\sum (x-\mu_x)(y-\mu_y)\,P(x,y) \quad \text{if discrete}$$

$$= \int\int (x-\mu_x)(y-\mu_y)\,f(x,y)\,dx\,dy \quad \text{if continuous}$$

Consider the following

⇒ Whether spending on advertising of a company is related to overall sales of the company.
→ If it is related, how it is related

⇒ Forecasting the sales, given the budget for advertising

⇒ Whether spending on advertising of a company is related to overall sales of the company.
→ If it is related, how it is related

⇒ Forecasting the sales, given the budget for advertising

Correlation

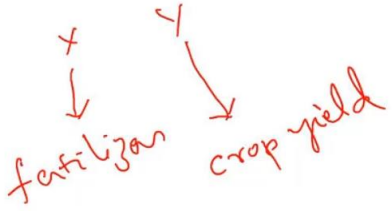
Regression

And also

⇒ Farmer has an impression that if he uses more fertilizers, then the crop yield increases.

We need to validate this?

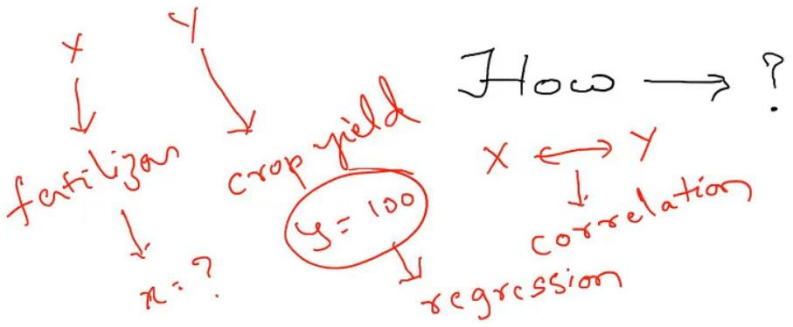
How → ?



How → ?

X ⟷ Y

Correlation – crop yield & fertilizer → relation between them



Predicting based on related → Future prediction – Regression

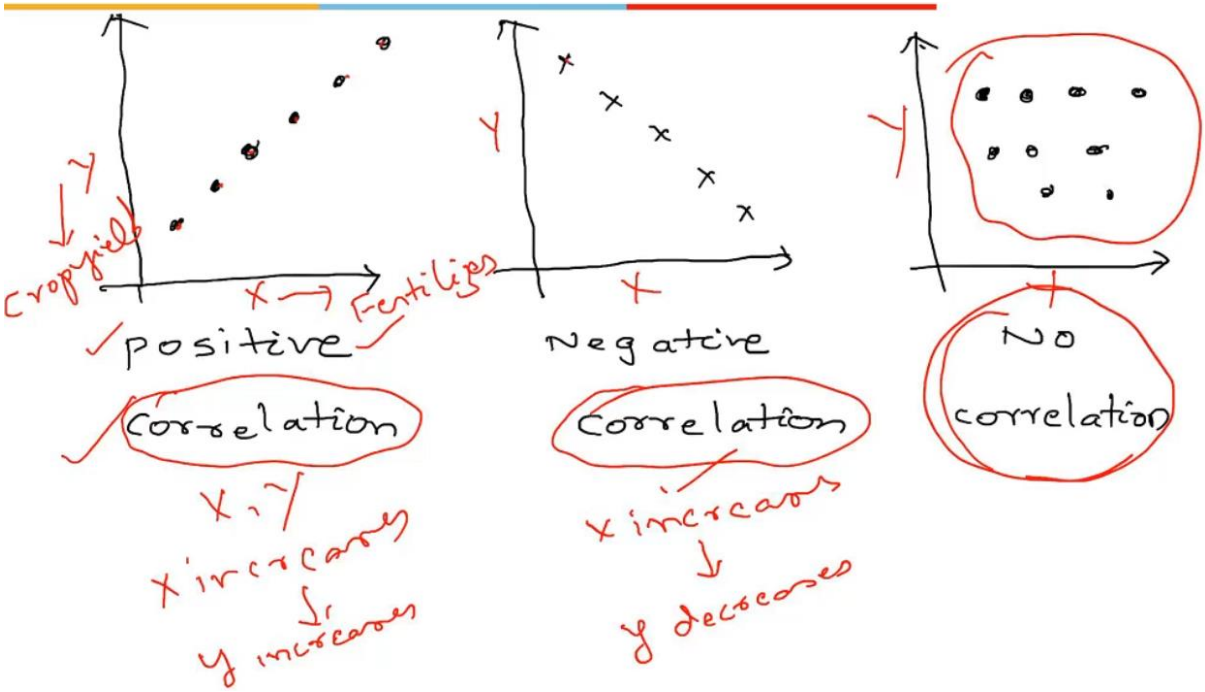
Correlation

→ Sales of a company and Expenditure on advertisement

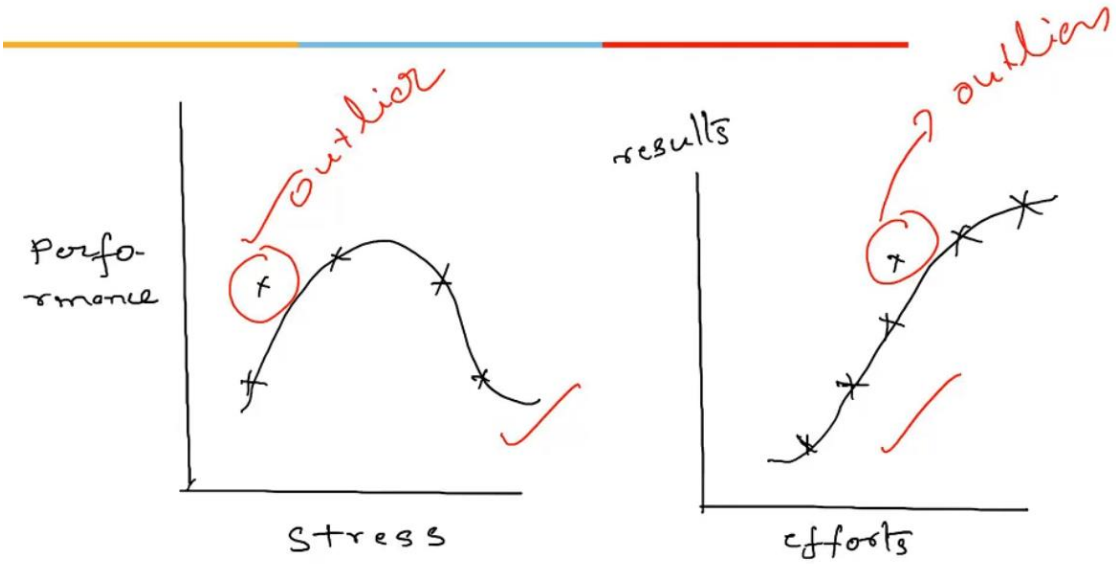
→ Price and Demand of a product

→ Inflation and Gold Price

→ IQ and performance in Entrance.



Quantification → based on visualization – we can conclude – is there any Mathematical tool – to conclude – if no. of points are more – it's difficult to come out with conclusion – require Mathematical representation – helps to come out with conclusion



Coefficient of correlation:

$$r = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \cdot \Sigma Y^2}}$$

where $X = x - \bar{x}$ ($\bar{x}$: mean of x)

$Y = y - \bar{y}$ ($\bar{y}$: mean of y)

$X^2 = (x - \bar{x})^2$

$Y^2 = (y - \bar{y})^2$

Coefficient of Correlation

innovate achieve lead

$r = 1 \Rightarrow$ Perfect and positive relation

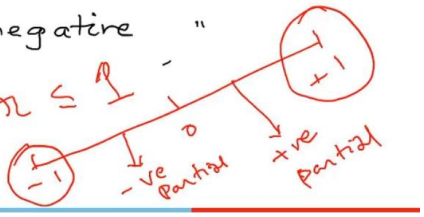
$r = -1 \Rightarrow$ " " negative relation

$r = 0 \Rightarrow$ No relation

$0 < r < 1 \Rightarrow$ Partial positive relation

$-1 < r < 0 \Rightarrow$ " negative "

$-1 \le r \le 1$



Example - 1

innovate achieve les

x	1	2	3	4	5	6	7	8	9
y	10	11	12	14	13	15	16	17	18

$$\bar{x} = \frac{\Sigma x}{n} = \frac{45}{9} = 5$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{126}{9} = 14$$

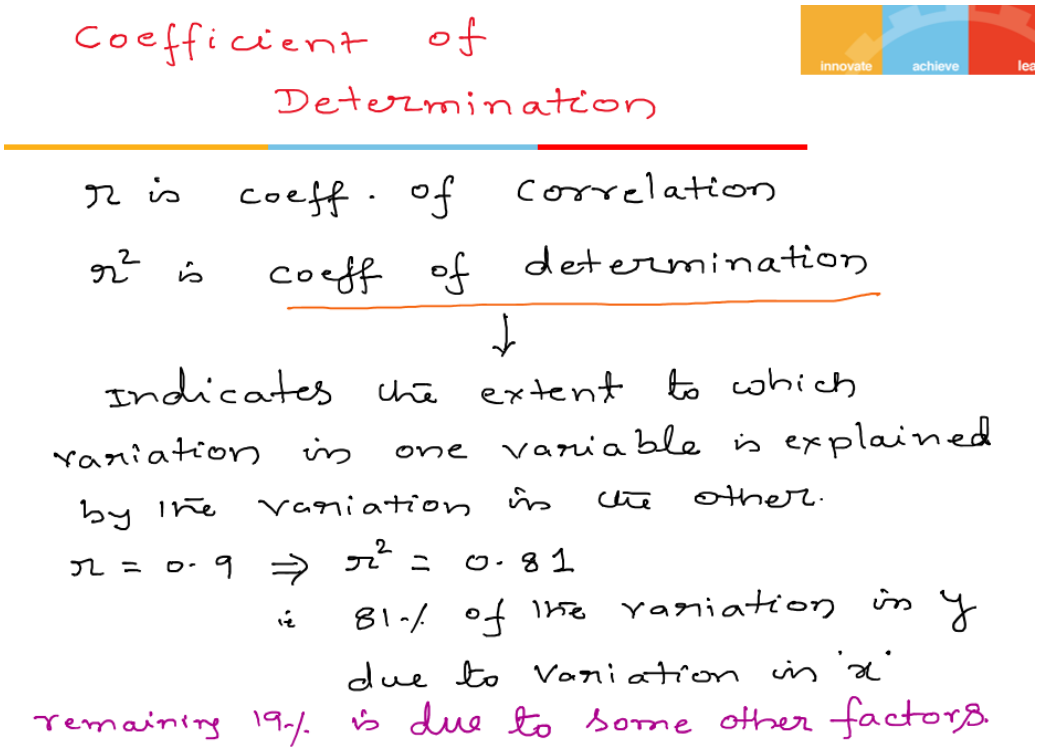
x	$X = x-5$	X^2	y	$Y = y-14$	Y^2	XY
1	-4	16	10	-4	16	16
2	-3	9	11	-3	9	9
3	-2	4	12	-2	4	4
4	-1	1	14	0	0	0
5	0	0	13	-1	1	0
6	1	1	15	1	1	1
7	2	4	16	2	4	4
8	3	9	17	3	9	9
9	4	16	18	4	16	16
		60			60	59

$$r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \Sigma Y^2}}$$

$$= \frac{59}{\sqrt{60 \times 60}}$$

$$= 0.9833$$

Coefficient of determination – r^2 - At which extent – x expands y

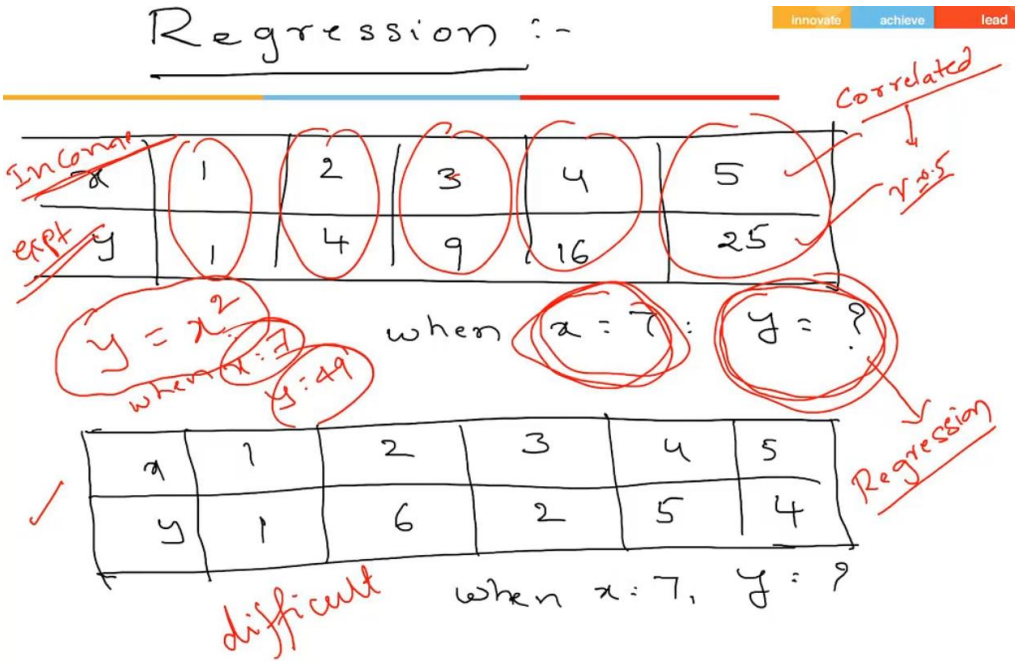


Coefficient of correlation – relation between two variables – normalized – [-1, 1] – change in variable based on another variable

Covariance → relation between two variables – not normalized form

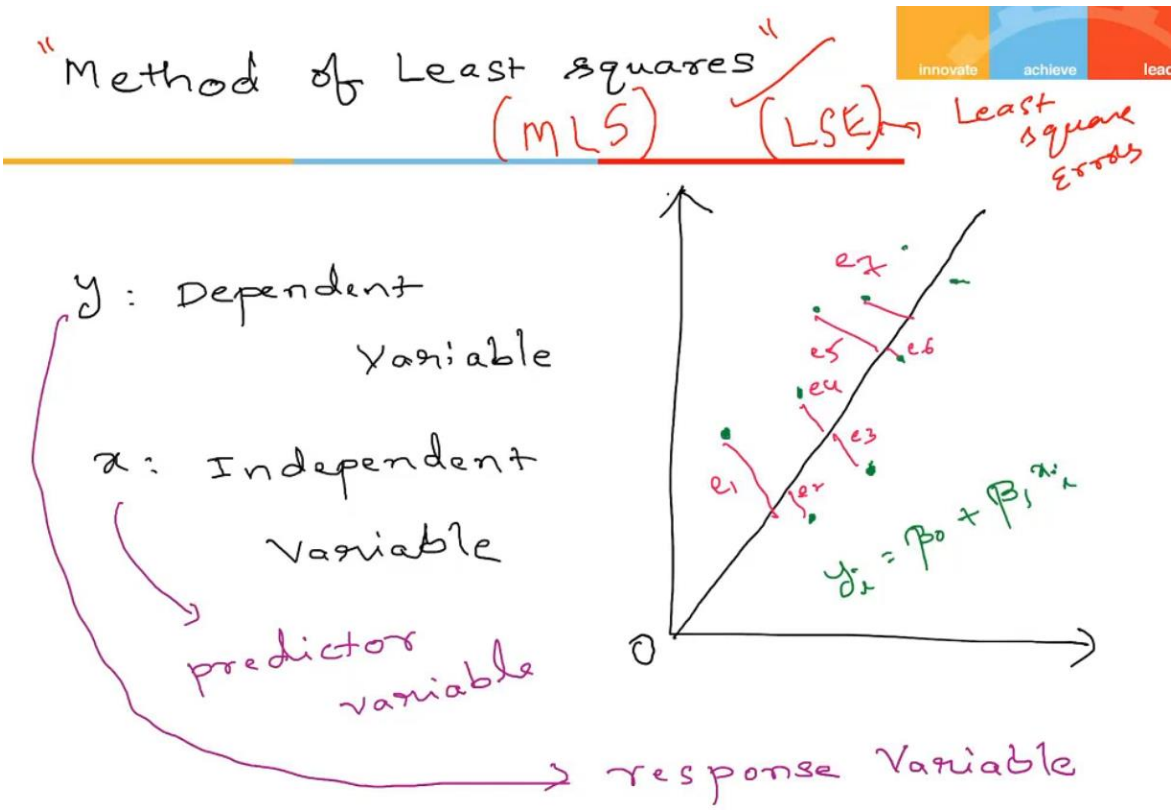
Team performance – Manager - Opinion → punctual – 9 to 6 – for good performance – employees – proposal – work from home (higher travel time) – work from home – testing period – 1 month – these two are related or not

Regression



First find relation – having relation or not – Correlation → Forecasting → regression – exact relation is required to predict

To establish a relation – what exact relation – Regression, To find out what relation and whether they are related or not → Correlation

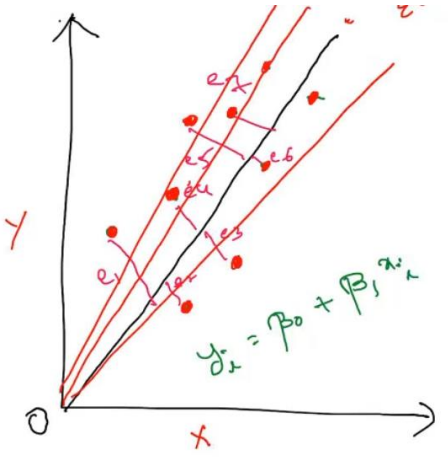


Try to draw straight line – so most points are covered

y : Dependent Variable

x : Independent Variable

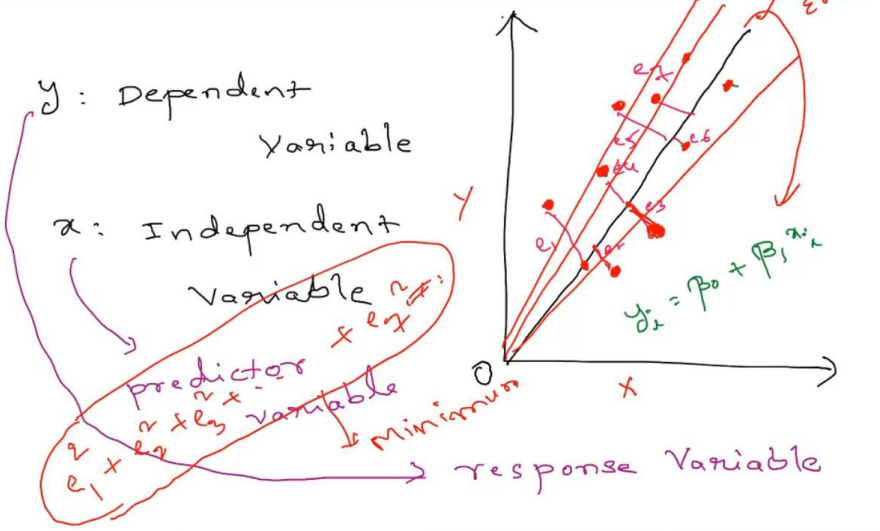
predictor variable



Which line is optimal? – Error approximation is minimal – LSE

"Method of Least squares"

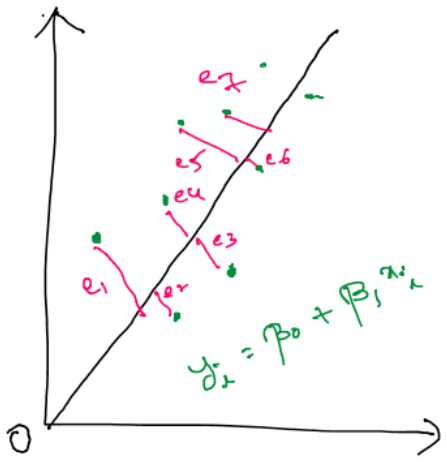
(MLS) (LSE) → Least square Errors



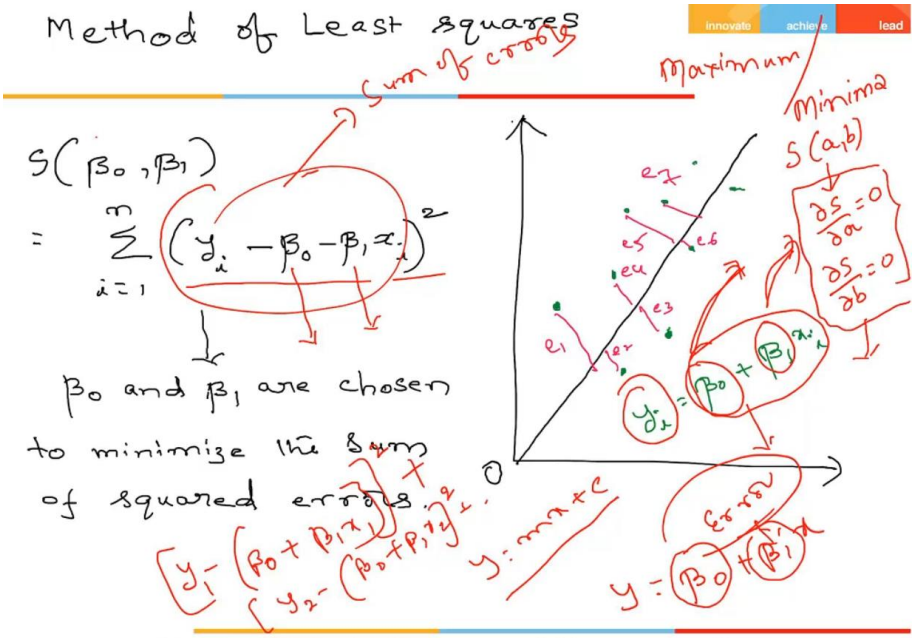
Method of Least squares

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

β_0 and β_1 are chosen to minimize the sum of squared errors.



Method of Least squares



Method of Least squares

$$S(\beta_0,\beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{\partial S}{\partial \beta_0} = 0 \Rightarrow 2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)(-1)$$

$$\Rightarrow \sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i$$

$$\frac{\partial S}{\partial \beta_1} = 0 \Rightarrow 2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)(2)(-x_i)$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i = \beta_0 \sum x_i + \beta_1 \sum x_i^2$$

on solving these, we get β_0 & β_1 which minimizes error.

Method of Least squares

$$S(\beta_0,\beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

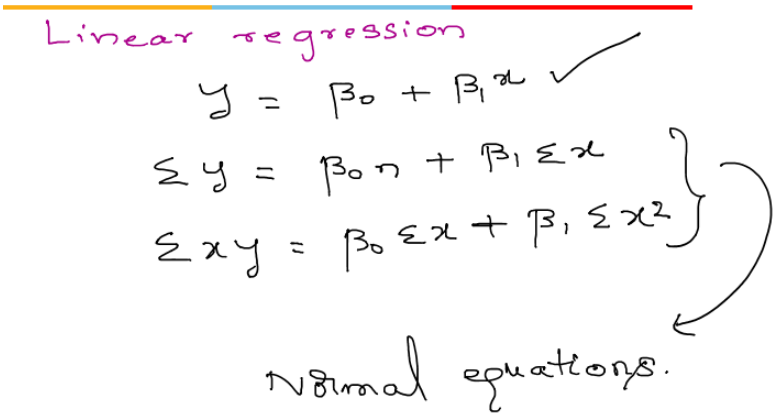
$$\frac{\partial S}{\partial \beta_0} = 0 \Rightarrow 2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)(-1) = 0$$

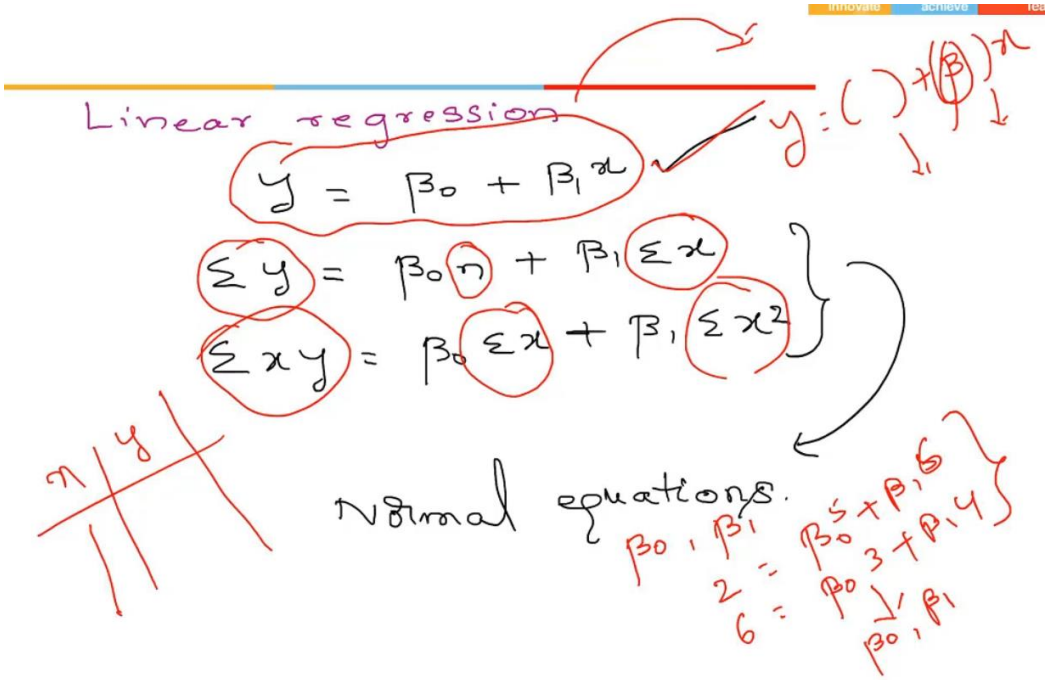
$$\Rightarrow \sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i$$

$$\frac{\partial S}{\partial \beta_1} = 0 \Rightarrow 2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)(2)(-x_i)$$

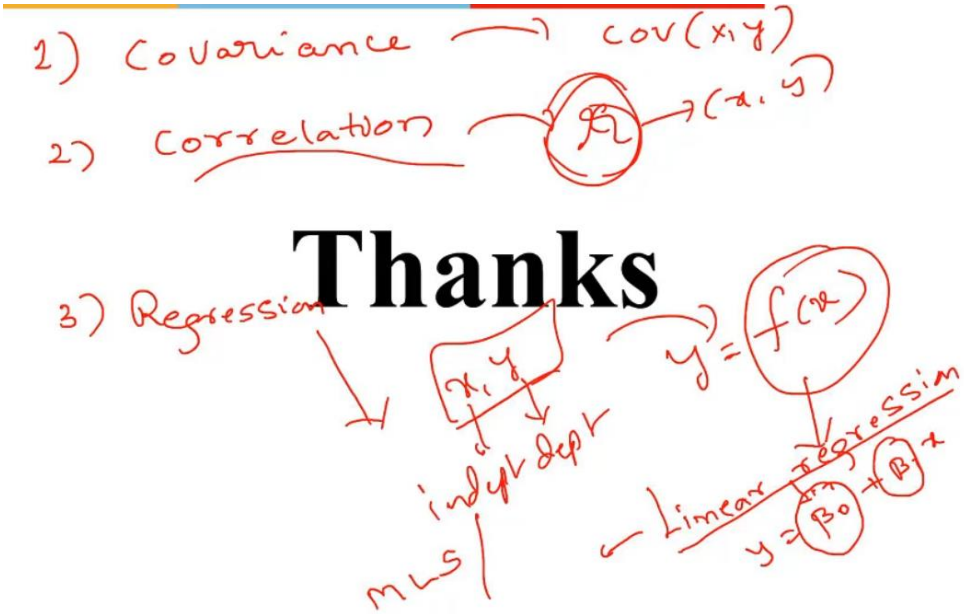
$$\Rightarrow \sum_{i=1}^{n} x_i y_i = \beta_0 \sum x_i + \beta_1 \sum x_i^2$$

on solving these, we get β_0 & β_1 which minimizes error.





Find β0,β1 → so most of the points are nearer → optimal case



Matrix Approach:

Let $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}$

observations $y_i = 1, 2, \ldots n$ → by a vector Y

unknowns $\beta_0, \beta_1 \cdots \beta_{p-1}$ → " " β

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,p-1} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,p-1} \end{bmatrix}$$

$$\underset{n\times 1}{\hat{Y}} = \underset{n\times p}{X}\ \underset{p\times 1}{\beta}$$

Find β to minimize

$$S(\beta) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \cdots \right)^2$$

$$= \| Y - X\beta \|^2 = \| Y - \hat{Y} \|^2$$

Diff S wrt to each β we get Linear eqns

$$X^T X \hat{\beta} = X^T Y \rightarrow \text{normal eqns}$$

If $X^T X$ is non-singular, the soln is

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Computationally, it is sometimes unwise even to form the normal equations because the multiplications involved in forming $X^T X$ can introduce undesirable round-off error.

Linear regression (Multiple regression)

Example:-

x_0	Size	No of rooms	No of floors	Age of home	price Lakhs
1	2000	5	2	45	4000
1	1400	3	1	40	2000
1	1600	3	2	30	3000
1	800	2	1	35	2000
		x_1	x_2	x_3	x_4	y

Linear regression (Multiple regression)

Example:-

$$X = \begin{bmatrix} 1 & 2000 & 5 & 2 & 45 \\ 1 & 1400 & 3 & 1 & 40 \\ 1 & 1600 & 3 & 2 & 30 \\ 1 & 800 & 2 & 1 & 35 \end{bmatrix} \quad y = \begin{bmatrix} 4000 \\ 2000 \\ 3000 \\ 2000 \end{bmatrix}$$

$$\beta = (X^T X)^{-1} X^T y$$

Example:

Consider the following data

x	1	2	4	0
y	0.5	1	2	0

Fit a linear regression line

Estimate y when $x = 5$.

x	y	xy	x^2
1	0.5	0.5	1
2	1	2	4
4	2	8	16
0	0	0	0
$\Sigma = 7$	$\Sigma 3.5$	$\Sigma 10.5$	$\Sigma 21$

$y = \beta_0 + \beta_1 x$

$\Sigma y = n\beta_0 + \beta_1 \Sigma x$

$\Sigma xy = \beta_0 \Sigma x_1 + \beta_1 \Sigma x^2$

$3.5 = 4\beta_0 + \beta_1 (7)$

$10.5 = 7\beta_0 + \beta_1 (21)$

on solving these

$\beta_0 = 0$

$\beta_1 = 0.5$

i.e. $y = 0 + (0.5)x$

when $x = 5$, $y = (0.5)5 = 0.25$

Lec 9 Predictive Analytics & Revision

L- 9: Predictive Analytics & Revision

Agenda

- Review of last session
- Introduction to regression
- Method of least squares
- Simple linear regression

Covariance of X and Y

$E(x) = \sum x P(x)$ or $= \int x f(x)\,dx$

$$Cov(X,Y) = E\left[(X-\mu_X)(Y-\mu_Y)\right]$$

$$= \sum_x \sum_y (x-\mu_x)(y-\mu_y) P(x,y)$$ if discrete — joint p.d.f

$$= \iint (x-\mu_x)(y-\mu_y) f(x,y)\,dx\,dy$$ if continuous — joint prob. density fn

$$Cov(x, y) : \frac{\sum_{i=1}^{n} (x - \mu_x)(y - \mu_y)}{n - 1}$$

Correlation

X → income

Y → expenditure

$(n-1)$

n

And also

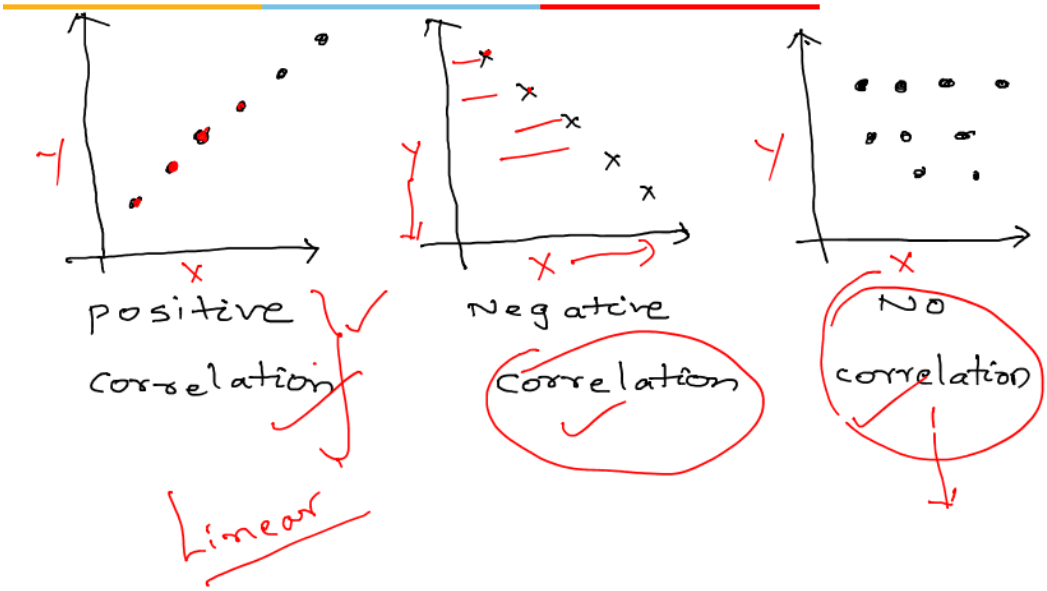
innovate achieve

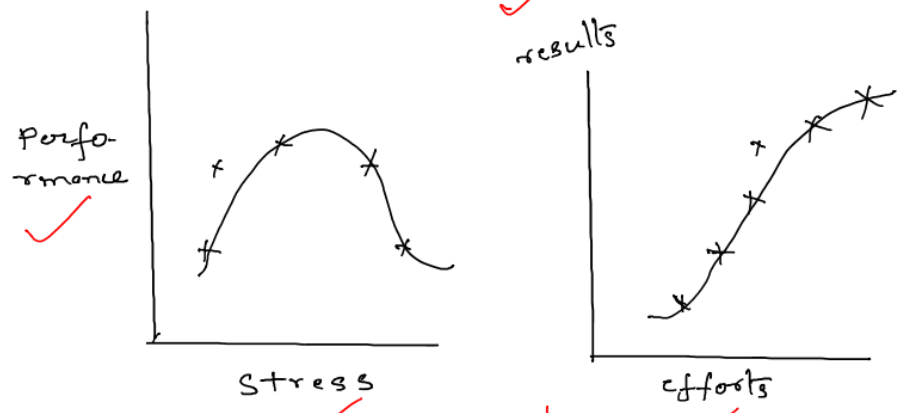
⇒ Farmer has an impression that if he uses more fertilizers, then the crop yield increases.

We need to validate this?

How ⟶ ?

innovate achieve le





non-linear

Coefficient of correlation:

$$r = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \cdot \Sigma Y^2}}$$

where $X = x - \bar{x}$

$Y = y - \bar{y}$

$X^2 = (x - \bar{x})^2$

$Y^2 = (y - \bar{y})^2$

Coefficient of Correlation

$r = 1 \Rightarrow$ Perfect and positive relation

$r = -1 \Rightarrow$ " " negative relation

$r = 0 \Rightarrow$ No relation

$0 < r < 1 \Rightarrow$ Partial positive relation

$-1 < r < 0 \Rightarrow$ " negative "

$-1 \leq r \leq 1$

Example - 1

x	1	2	3	4	5	6	7	8	9
y	10	11	12	14	13	15	16	17	18

$$\bar{x} = \frac{\Sigma x}{n} = \frac{45}{9} = 5$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{126}{9} = 14$$

x	$X = x-5$	X^2	y	$Y = y-14$	Y^2	XY
1	−4	16	10	−4	16	16
2	−3	9	11	−3	9	9
3	−2	4	12	−2	4	4
4	−1	1	14	0	0	0
5	0	0	13	−1	1	0
6	1	1	15	1	1	1
7	2	4	16	2	4	4
8	3	9	17	3	9	9
9	4	16	18	4	16	16
		60			60	59

$$r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \Sigma Y^2}}$$

$$= \frac{59}{\sqrt{60 \times 60}}$$

$$= 0.9833$$

$r = 0.9833$

→ Corrected

→ +ve correlated

x	$X = x-5$	X^2	y	$Y = y-14$	Y^2	XY
1	−4	16	10	−4	16	16
2	−3	9	11	−3	9	9
3	−2	4	12	−2	4	4
4	−1	1	14	0	0	0
5	0	0	13	−1	1	0
6	1	1	15	1	1	1
7	2	4	16	2	4	4
8	3	9	17	3	9	9
9	4	16	18	4	16	16
		60			60	59

$$\therefore \text{Cov}(x,y) = \frac{\Sigma XY}{n-1} = \frac{59}{8} = 7.375$$

Advanced Statistical Techniques for Analytics 05-10-2018 Slide 13

innovate achieve lead

Coefficient of Determination

r is coeff. of correlation

r^2 is coeff of determination

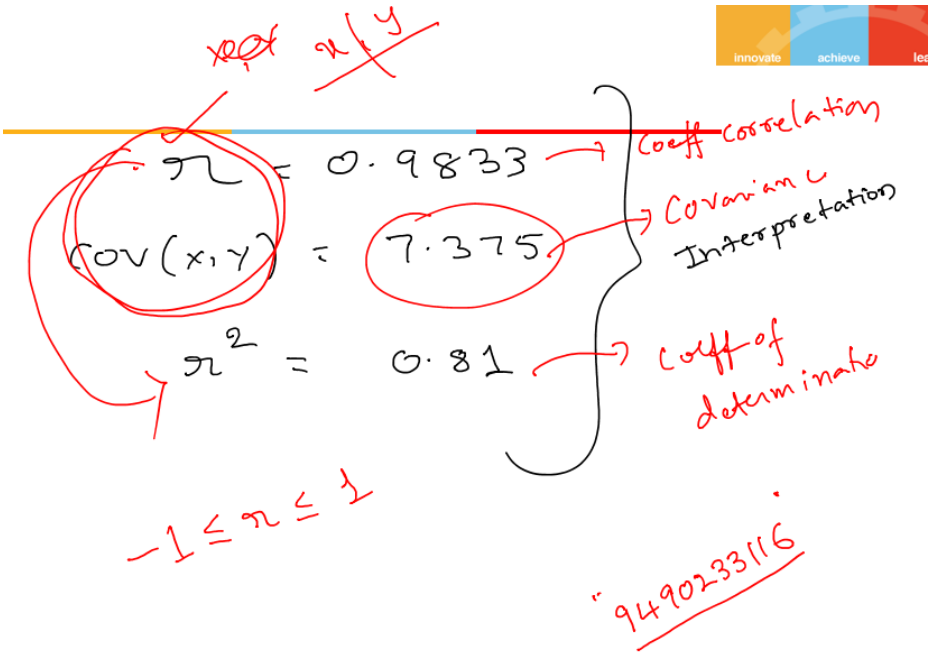
↓

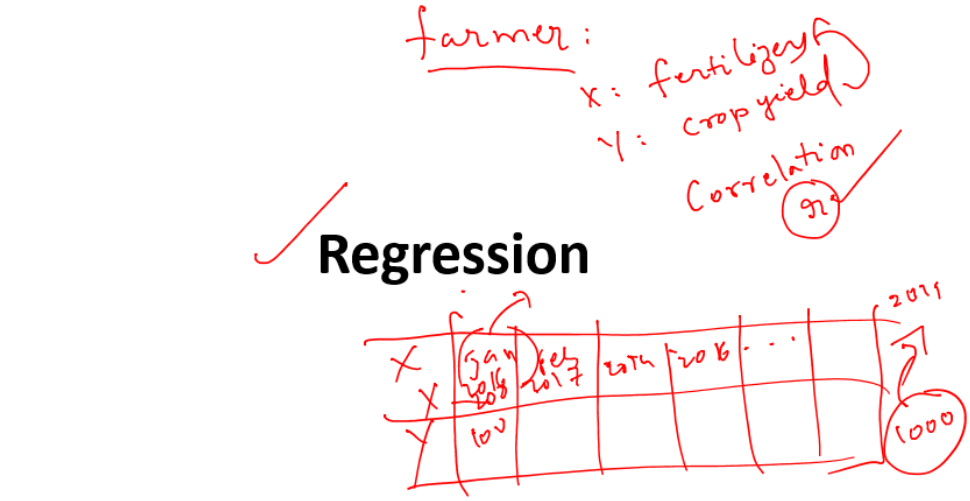
Indicates the extent to which variation in one variable is explained by the variation in the other.

$r = 0.9 \Rightarrow r^2 = 0.81$

i.e. 81% of the variation in y is due to variation in 'x'

remaining 19% is due to some other factors.

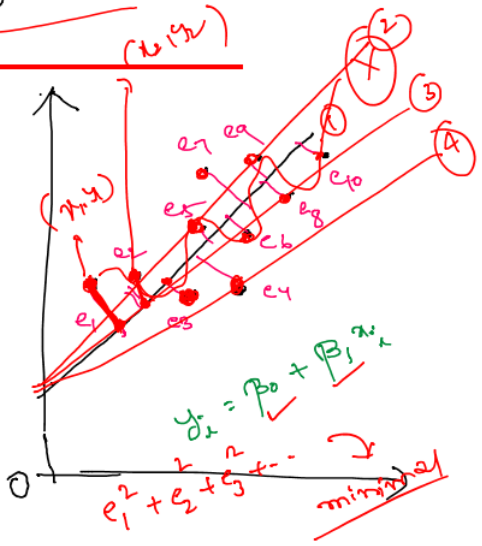

r = 0.9833
Cov(x,y) = 7.375
r² = 0.81
Coeff correlation
Covariance
Interpretation
coeff of determinate
-1 ≤ r ≤ 1
9490233116


farmer:
x: fertilizer
y: crop yield
Correlation
r
Regression


Regression :-
correlation
r
x | 1 | 2 | 3 | 4 | 5
y | 1 | 4 | 9 | 16 | 25
y = x² when x = 7
y = 49
when x = 7 : y = ?
x | 1 | 2 | 3 | 4 | 5
y | 1 | 6 | 2 | 5 | 4
when x = 7, y = ?

Correlation	Regression
→ Measuring strength or degree of the relationship between two variables	→ having an algebraic equation between two variables
→ no estimation	→ estimation
→ both variables are independent	→ one is dept variable and other indept variables

Method of Least squares



y : Dependent Variable

x : Independent Variable

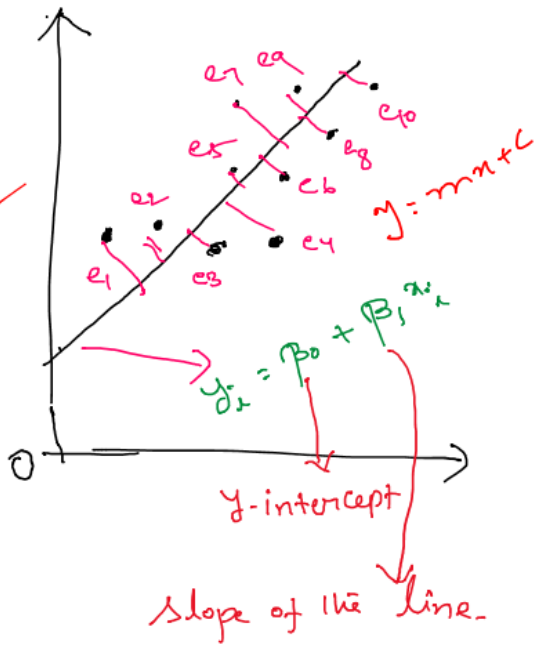
→ predictor variable

→ response Variable

" minimizing the error "

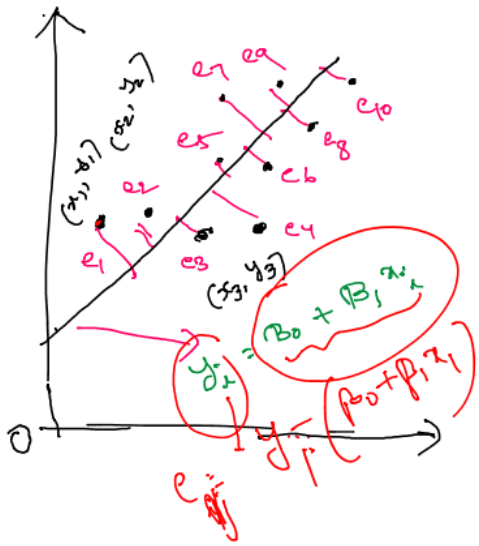
" minimize

$e_1^2 + e_2^2 + e_3^2 + \cdots e_{10}^2$



$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

we need to choose β_0 and β_1 which minimizes the error.



Method of Least squares → Max / Min → first derivative = 0

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{\partial S}{\partial \beta_0} = 0 \Rightarrow 2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)(-1)$$

$$\Rightarrow \sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i$$

$$\frac{\partial S}{\partial \beta_1} = 0 \Rightarrow 2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)(2)(-x_i)$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i = \beta_0 \sum x_i + \beta_1 \sum x_i^2$$

on solving these, we get β_0 & β_1 which minimizes error.

Linear regression

$$y = \beta_0 + \beta_1 x$$

$$\sum y = \beta_0 n + \beta_1 \sum x$$

$$\sum xy = \beta_0 \sum x + \beta_1 \sum x^2$$

Normal equations. β_0, β_1

$y = \beta_0 + \beta_1 x$

Regression Coefficients

$y = a + bx$ — regression line of y on x

b_{yx} : Regression coeff of y on x

$x = c + dy$ — regression line of x on y

b_{xy} : regression coeff of x on y

Correlation coefficient

$$r = \sqrt{b_{yx} \times b_{xy}}$$

b_{yx}, b_{xy} → Positive → Geometric mean

r = $+0.9$ or -0.9

Negative

r^2 : Determination : 0.81

Example :-

company	Advt Expt	Sales Revenue
A	1	1
B	3	2
C	4	2
D	6	4
E	8	6
F	9	8
G	11	8
H	14	9

$y = a + bx$

$\Sigma y = an + b\Sigma x$

$\Sigma xy = a\Sigma x + b\Sigma x^2$

Example :-

Sales Revenue y	Advt. expt. x	x^2	xy
1	1	1	1
2	3	9	6
2	4	16	8
4	6	36	24
6	8	64	48
8	9	81	72
8	11	121	88
9	14	196	126
Σ 40	Σ 56	Σ 524	Σ 373

$y = \beta_0 + \beta_1 x$

$\Sigma y = n\beta_0 + \beta_1 \Sigma x$

$\Sigma xy = \beta_0 \Sigma x + \beta_1 \Sigma x^2$

$\Rightarrow 40 = 8\beta_0 + 56\beta_1$

$373 = 56\beta_0 + 524\beta_1$

on solving

$\beta_0 = 0.072$

$\beta_1 = 0.704$

$\therefore y = (0.072) + (0.704)x$

$\therefore y = (0.072) + (0.704)x$

when $x = 0.075$, then

$y = (0.072) + (0.704)(0.075)$

$= 0.1248 \simeq 12.48\%$

Example :

Consider the following data

x	1	2	4	0
y	0.5	1	2	0

Fit a linear regression line

Estimate y when $x = 5$.

x	y	xy	x^2
1	0.5	0.5	1
2	1	2	4
4	2	8	16
0	0	0	0
$\Sigma = 7$	$\Sigma 3.5$	$\Sigma 10.5$	$\Sigma 21$

$y = \beta_0 + \beta_1 x$

$\Sigma y = n\beta_0 + \beta_1 \Sigma x$

$\Sigma xy = \beta_0 \Sigma x_1 + \beta_1 \Sigma x^2$

$3.5 = 4\beta_0 + \beta_1 (7)$

$10.5 = 7\beta_0 + \beta_1 (21)$

on solving these

$\beta_0 = 0$

$\beta_1 = 0.5$

i.e $y = 0 + (0.5)x$

When $x = 5$, $y = (0.5)5$

$= 0.25$

Linear regression

(Multiple regression)

Example:-

No ↓	Size	No of rooms	No of floors	Age of home	price Lakhs
1	2000	5	2	45	4000
1	1400	3	1	40	2000
1	1600	3	2	30	3000
1	800	2	1	35	2000
	x_1	x_2	x_3	x_4	y

House → price

$y \rightarrow x$

increment my put

$y =$

20 years
1 floor
2 rooms
1200 sqft

Multiple Linear regression

$y = \beta_0 + \beta_1 x$ → $\Sigma y = \beta_0 n + \beta_1 \Sigma x$, $\Sigma xy = \beta_0 \Sigma x + \beta_1 \Sigma x^2$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

$$\Sigma y = \beta_0 n + \beta_1 \Sigma x_1 + \beta_2 \Sigma x_2 + \beta_3 \Sigma x_3 + \beta_4 \Sigma x_4$$

$$\Sigma x_1 y = \beta_0 \Sigma x_1 + \beta_1 \Sigma x_1^2 + \beta_2 \Sigma x_1 x_2 + \beta_3 \Sigma x_1 x_3 + \beta_4 \Sigma x_1 x_4$$

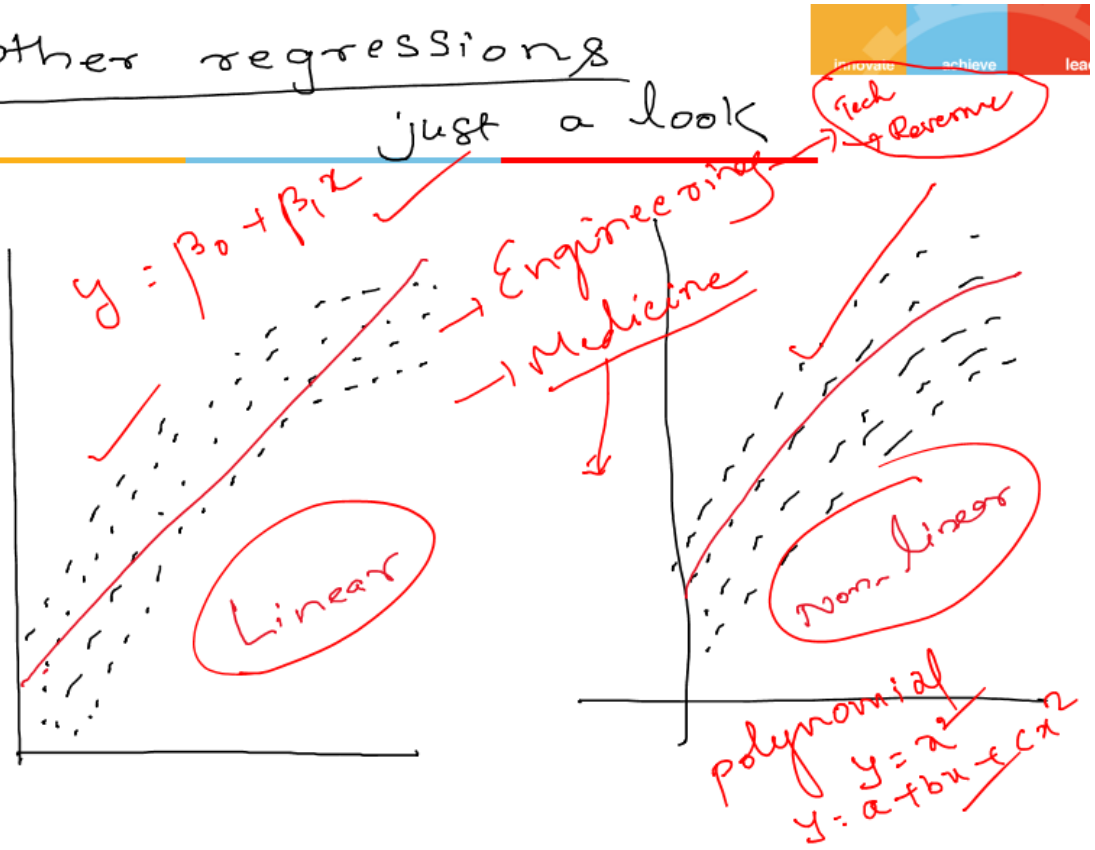
$$\Sigma x_2 y = \beta_0 \Sigma x_2 + \beta_1 \Sigma x_1 x_2 + \beta_2 \Sigma x_2^2 + \beta_3 \Sigma x_2 x_3 + \beta_4 \Sigma x_2 x_4$$

$$\Sigma x_3 y = \beta_0 \Sigma x_3 + \beta_1 \Sigma x_1 x_3 + \beta_2 \Sigma x_2 x_3 + \beta_3 \Sigma x_3^2 + \beta_4 \Sigma x_3 x_4$$

$$\Sigma x_4 y = \beta_0 \Sigma x_4 + \beta_1 \Sigma x_1 x_4 + \beta_2 \Sigma x_2 x_4 + \beta_3 \Sigma x_3 x_4 + \beta_4 \Sigma x_4^2$$

solve $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$

Other regressions just a look



Suppose $y = ae^{bx}$ → Exponential curve → Non-Linear

$\log y = \log a + b \log x$

$Y \quad A \quad X$

i.e. $Y = A + bX$ → Linear eqn

$\Sigma Y = An + b \Sigma X \rightarrow (1)$

$\Sigma XY = A \Sigma X + b \Sigma X^2 \rightarrow (2)$

$A = ?$ $\quad b = ?$ $\Rightarrow a$

Hence, we get $y = ae^{bx}$

Suppose $y = ax^b$ → Non Linear → Power Curve

$\log y = (\log a) + b \log x$

$Y \quad A \quad X$

i.e. $Y = A + bX$

$\Sigma Y = An + b \Sigma X$

$\Sigma XY = A \Sigma X + b \Sigma X^2$

$A \rightarrow a$, $b \rightarrow b$

$y = ax^b$

Matrix Approach:

Let $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}$

observations $y_i = 1, 2, \ldots n$ → by a vector Y

unknowns $\beta_0, \beta_1 \cdots \beta_{p-1}$ → " " β

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,p-1} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,p-1} \end{bmatrix}$$

$$\hat{Y} = X \beta$$

$n \times 1 \quad n \times p \quad p \times 1$

$Y = BX$, A, $X = B$

$\begin{bmatrix} 2 & 3 & 4 \\ 3 & 5 & 6 \\ 4 & 6 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 5 \\ 4 \\ \cdots \end{bmatrix}$

$X = A^{-1}B$

Find β to minimize

$$S(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_1 - \beta_2 x_2 \cdots)^2$$

$$= \| Y - X\beta \|^2 = \| Y - \hat{Y} \|^2$$

Diff S wrt to each β we get Linear eqns

$$X^T X \hat{\beta} = X^T Y \rightarrow \text{normal eqns}$$

If $X^T X$ is non singular, the soln is

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

$AX = B$

$X = A^{-1}B$

Computationally, it is sometimes unwise even to form the normal equations because multiplications involved in forming $X^T X$ can introduce undesirable round-off error.

→ If $X^T X$ is non-invertible ... ?

✓ Redundant features

✓ too many features → scaling

Revision

→ probability → $P(A \cup B)$

$P(A \cap B)$

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

→ Conditional probability:

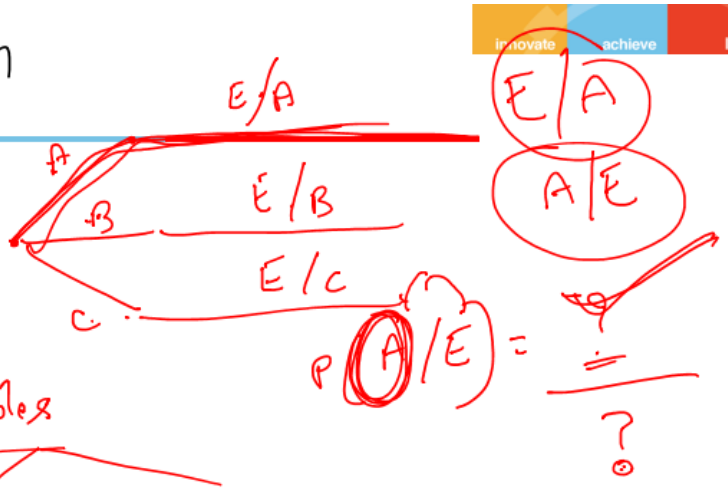
$P(A/B) = \frac{P(A \cap B)}{P(B)}$

$P(B/A) = \frac{P(A \cap B)}{P(A)}$ i.e. $P(A \cap B) = P(A/B)P(B)$

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

Revision

→ Bayes theorem :-



→ Random variables

Discrete
$P(x)$
i) $0 \le P(x) \le 1$
ii) $\sum P(x) = 1$
Mean $= E(X) = \sum x P(x)$

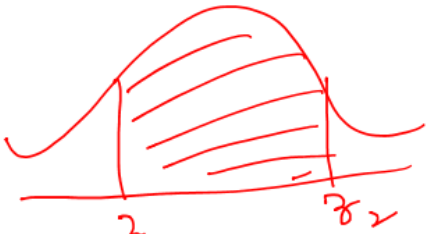
Continuous
$f(x)$
i) $0 \le f(x) \le 1$
ii) $\int f(x)\,dx = 1$
$= \int x f(x)\,dx$

Variance : $E(x-\mu)^2$
$= E(X^2) - \mu^2$
$= E(X^2) - [E(X)]^2$

→ Binomial dist $P(x) = {}^nC_x\, p^x q^{n-x}$, $x = 0,1,2 \cdots n$

poisson dist $P(x) = \frac{e^{-\lambda}\lambda^x}{x!}$, $x = 0,1,2 \cdots \infty$

→ Normal distribution :-



$P(30 \le x \le 50)$

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

$P(z_1 \le z \le z_2)$

$= F(z_2) - F(z_1)$

→ Testing of Hypothesis → Mean → one → z, t
→ two → z, t
→ Proportion → χ^2 - distribution

Thanks

9490233116

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
Work Integrated Learning Programmes Division

Cluster Programme - M. Tech in Data Science and Engineering

II Semester, 2018- 19

Mid semester Examination (**Regular**)

Course No	:	DSECL ZC413
Course Title	:	Introduction to Statistical Methods
Nature of Exam.	:	Closed Book
Weightage	:	30 Marks
Duration	:	90 minutes
Date	:	4th August, 2019.(10 to 11.30 AM)

Number of questions: 5
Number of Pages: 3

NOTE: Assume and answer wherever more data/information is required or missing. State the assumption clearly.

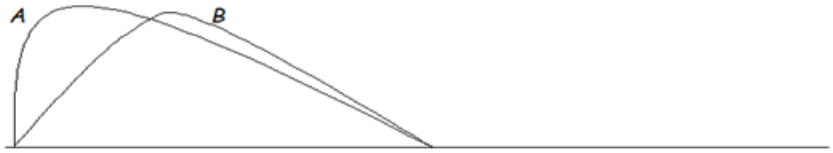
Q1*).* ***To be answered only in Pages 3 to 5*** **[3 + 3 = 6M]**

a) The following Fig refer to two distributions A & B. Based on the Fig answer the following questions, if possible.

(i) Which one of the two has a higher mean?

(ii) Which of the two is more likely to produce smaller values than larger values?

(iii) Which of the two has its mean exceeding its median?



b) The probability of Mr. Rakesh passing examination is 1/2 and Mr Ramesh passing examination is 1/3. What is the probability that at least one person will pass the examination?

Q2). ***To be answered only in Pages 6 to 8*** **[3 + 3 = 6M]**

a) The probability that doctor A will diagnose a disease X correctly is 0.6. The probability that a patient will die by his treatment after correct diagnosis is 0.4 and the probability of death by wrong diagnosis is 0.7. A patient of doctor A, who had disease X died. What is the probability that his disease was diagnosed correctly?

b) Consider the following

"A company wants to find reasons for the dissatisfaction among employees because of which they are leaving the company. With this view they took the feedback among the employees and identified three reasons i.e. working conditions, pay hike, commuting. The probabilities of dissatisfaction with these three factors are 0.6, 0.4 and 0.2

respectively. And the probabilities that they are leaving the organization with these reasons are 0.3, 0.5and 0.8 respectively."

As a data scientist, use an appropriate statistical model / method to model this case and suggest the company the probable reasons on priority so that they can focus on it to retain the employees.

.

Q3). ***To be answered only in Pages 9 to 13*** **[3 + 3 = 6M]**

a) (i) Validate the following probability distribution. Justify your answer.

x	-1	0	1	2	3
P(X = x)	0.2	0.2	0.2	0.2	0.2

(ii) Is it possible to find the mean of the random variable aX + b, if the mean of X is μ and variance is σ2.If possible find it?

b) Let X be a continuous random variable with probability density function

$$f(x) = \begin{cases} kx, & 0 \le x \le 1 \\ k, & 1 \le x \le 2 \\ -kx + 3k, & 2 \le x \le 3 \end{cases}$$

Find (i) value of k ii) Compute P(X ≤ 1.5).

Q4). ***To be answered only in Pages 14 to 17*** **[3 + 3 = 6M]**

a) The incidence of occupational disease in an industry is such that the workers have a 20% chance of suffering from it. What is the probability that out of six workers chosen at random. Four or more will suffer from the disease.

b) Marks obtained in a course follows normal distribution with mean 75 and standard deviation 10.If 300 students appeared at the examination of the course, then the number of students scoring

(i) Less than 70 marks

(ii) More than 90 marks

Q5). ***To be answered only in Pages 18 to 22***

a) "A manufacturer claims that mean life of their product is at least 60 months with a standard deviation of 3 months. But a sample of this product with 45 items is having mean 55 months with a standard deviation of 3 months. Now we want to validate the claim". **[1 + 1 M]**

(i) Mention the test statistic to use in the testing the claim and justify it.

(ii) Is it a one tailed or two tailed test? Justify the answer.

b) *"A company claims that the advertisement A is as effective as advertisement B. In a random sample of 60 customers who saw advertisement A, 18 tried the product. In a random sample of 100 customers, who saw the advertisement B, 22 tried the product. Does this support the claim of the company?"* **[1+1 M]**

(i)Which test statistic is useful in this validation? Justify the selection.

(ii)Is it a one tailed or two tailed test? Justify it.

c) An independent-measures t statistic is used to evaluate the mean difference between two treatments with n = 8 in one treatment and n = 12 in the other. What is the degrees of freedom value for the t statistic? **[1 M]**

d) Discuss the importance of Central Limit theorem. **[1 M]**

XXXXXX

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
Work Integrated Learning Programmes Division

Cluster Programme - M. Tech in Data Science and Engineering

II Semester, 2018- 19

Mid semester Examination (**MAKEUP**)

Course No : DSECL ZC413
Course Title : Introduction to Statistical Methods
Nature of Exam. : Closed Book
Weightage : 30 Marks
Duration : 90 minutes
Date : 18th August, 2019.(10 to 11.30 AM)

Number of questions: 5
Number of Pages: 3

NOTE: Assume and answer wherever more data/information is required or missing. State the assumption clearly.

Q1***).*** ***To be answered only in Pages 3 to 5*** **[3 + 3 = 6M]**

a) What is the criteria to decide whether the data is skewed or not? Use this to find whether the given data is skewed or not. ***12,18,9,6,15,13,12,8,20,18,22,18,9.***

b) A piece of equipment will function only when the three components A, B and C are working. The probability of A failing during one year is 0.15, that of B failing is 0.05 and that of C failing is 0.10. What is the probability that the equipment will fail before the end of the year?

Q2). ***To be answered only in Pages 6 to 8*** **[3 + 3 = 6M]**

a)

Event	A	B	C	A and B	A and C	B and C	A and B and C
Probability	0.14	0.23	0.37	0.08	0.09	0.13	0.05

Find (i) P(A / B U C) (ii) P(A / at least one of A,B,C) (iii) P(AUB /C)

b) In a certain town patients visit three doctors A, B and C in the ratio 6:5:7. The chances that these doctors will refer the case to a specialist are 0.45, 0.60 and 0.55 respectively.

(i) What is the probability that a patient selected at random visits a specialist?

(ii) If the patient has visited a specialist what is the probability that the reference was made by A.

Q3). ***To be answered only in Pages 9 to 13*** **[3 + 3 = 6M]**

a) The monthly demand for the items follows the following probability distribution.

x	1	2	3	4	5	6
P(X = x)	0.10	0.15	0.20	0.20	0.18	0.10

Then find the expected demand for the items. And find the variance of X.

b) A college Professor never finishes his lecture before the end of the hour and always finishes his lecture within 2 minutes after the hour. Let X = the time that elapses between the end of the hour and the end of the lecture and the pdf of X is

$f(x) = kx^2, \quad 0 \leq x \leq 2$. Find

(i). Value of k.

(ii). What is the probability that the lecture continues beyond the hour for between 60 and 90 sec?

(iii). What is the probability that the lecture continues for at least 90 sec beyond the end of the hour?

Q4). ***To be answered only in Pages 14 to 17*** **[3 + 3 = 6M]**

a) In a town, 10 accidents took place in a span of 40 days. Assuming that the number of accidents per day follows the Poisson distribution, find the probability that there

(i) will be three or more accidents in a day.

(ii) will be less than 3 accidents in a day.

(iii) will be exactly 3 accidents in a day.

b) Suppose the force acting on a column that helps to support a building is a normally distributed random variable X with mean value 15.0 kips and standard deviation 1.25 kips. Then find the following probabilities

$(i) P(X \leq 15)$ $(ii) P(X \leq 17.5)$ $(iii) P(|X - 15| \leq 3)$

Q5). ***To be answered only in Pages 18 to 22*** **[3 + 3 = 6M]**

a) *"The target thickness for silicon wafers used in a certain type of integrated circuit is 245* μm. *A sample of 50 wafers is obtained and thickness of each one is determined, resulting in a sample mean thickness of 246.18* μm *and a sample standard deviation of 3.6* μm. *Does this data suggest that the true average wafer thickness is something other than the target value?* **[1+1+1 M]**

(i) Mention Null hypothesis and suitable alternate hypothesis. Justify the selection

(ii) Is it a one tailed or two tailed test? Justify the answer.

(iii) Which test statistic is useful in this validation? Justify the selection.

b) *Consider the following table related to hypothesis testing and decisions related to it. (H0 is Null Hypothesis).* **[3M]**

Decision	H0 is true	H0 is false
Accept H0	***A***	***B***
Reject H0	***C***	***D***

Choose the correct statement from the following in place of ***A, B, C*** and ***D***

(i) Correct decision with confidence (1 – α).

(ii) Correct decision with confidence α.
(iii) Correct decision with confidence (1 – β).
(iv) Correct decision with confidence β.
(v) Type I error (α).
(vi) Type I error (1 - α).
(vii) Type II error (β).
(viii) Type II error (1 - β).

STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361

XXXXXX