

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 preferred

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



- 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

innovate

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

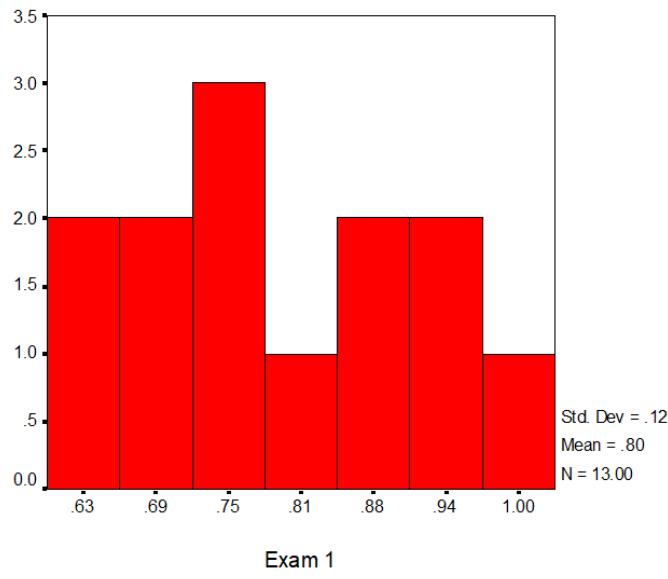
Useful for showing two samples side-by-side



Histograms

innovate achieve lead

■ Univariate histograms

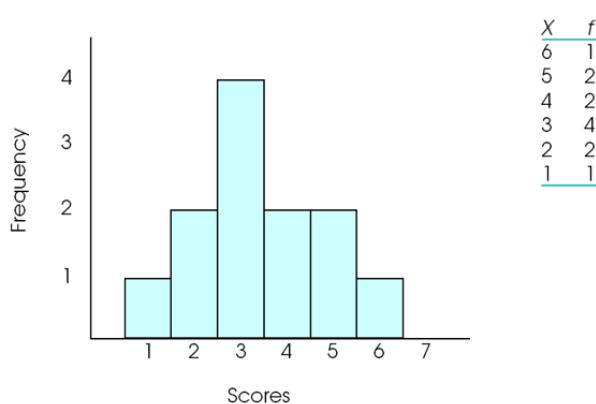


Exam 1

ide 13

Histograms

innovate achieve lead



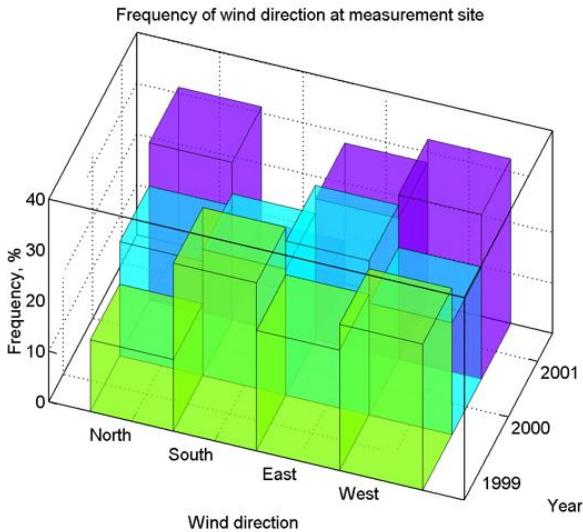
f on y axis (could also plot *p* or %)

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

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

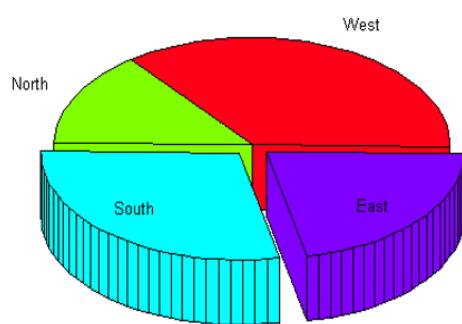
No gaps between bars

Bivariate histogram



Graphing the data – Pie charts

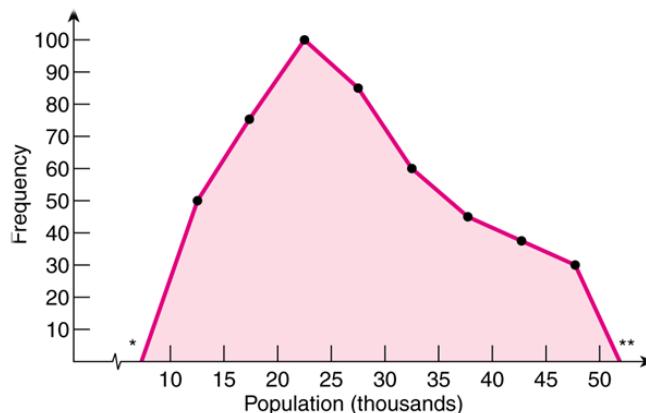
Frequency of wind direction at measurement site



Frequency Polygons

■ Frequency Polygons

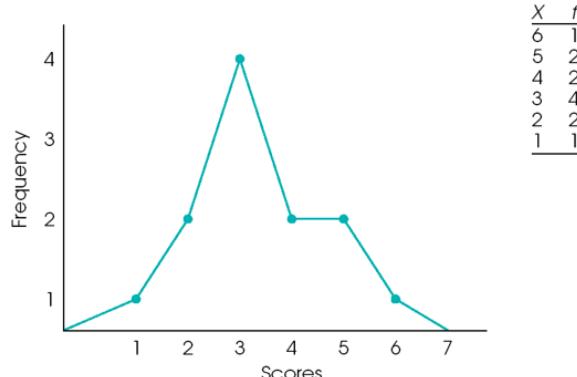
- Depicts information from a frequency table or a grouped frequency table as a **line graph**



* 4 cities had populations of less than 10,000.

** 5 cities had populations of 50,000 or greater.

Frequency Polygon



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.....

innovate achieve lead

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

Visualization

innovate achieve lead

- Summary gives an idea about the data summary(income)

<i>Min</i>	<i>1st QU.</i>	<i>Median</i>	<i>Mean</i>	<i>3rd Qu</i>	<i>Max</i>
- 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

innovate achieve lead

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

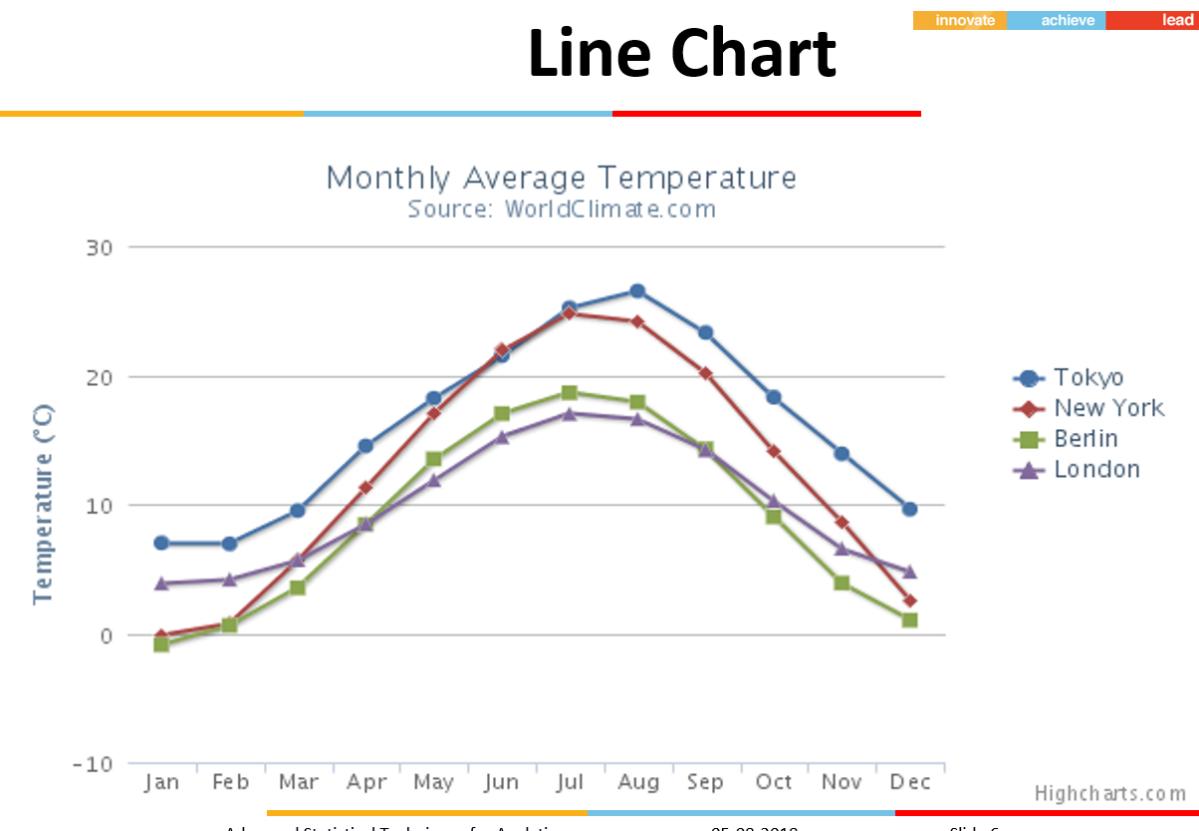
Main challenge in visualization is which one is opt for the outcome we are looking for.

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

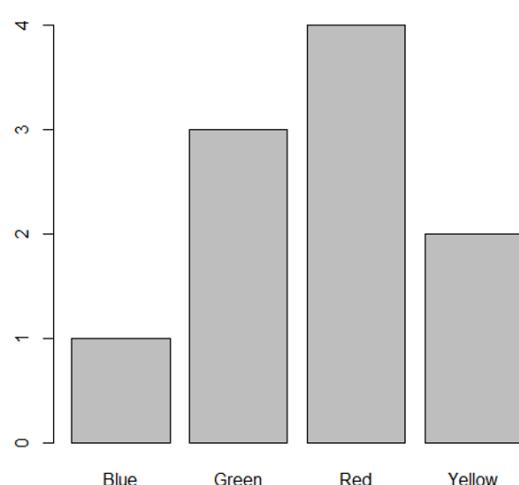
Traffic in Google Maps – visualization – we can plan our journey easily, NSE marker etc.

Resource utilization – we use line charts generally



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

Bar Chart



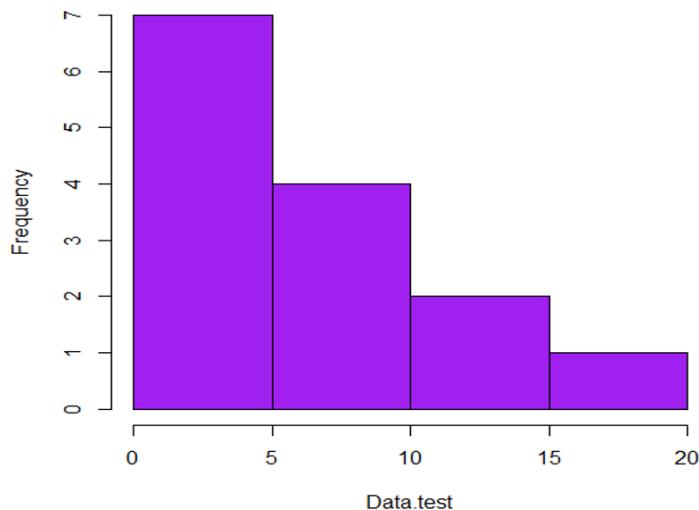
Weather patterns, profits made during quarters – we can observe growth - Grouping

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

Histograms

innovate achieve lead

Histogram of Data.test



Bar chart – they can be categorical values (not numerical values always) – independent
Histograms – numerical – continuous values – in groups → Class intervals – freq. distribution

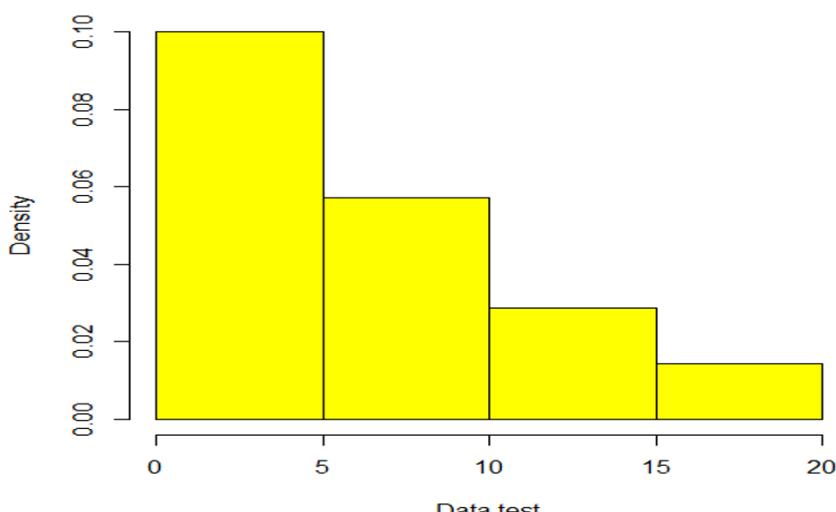
Histogram – Data divided into beans – consecutive ranges/groups

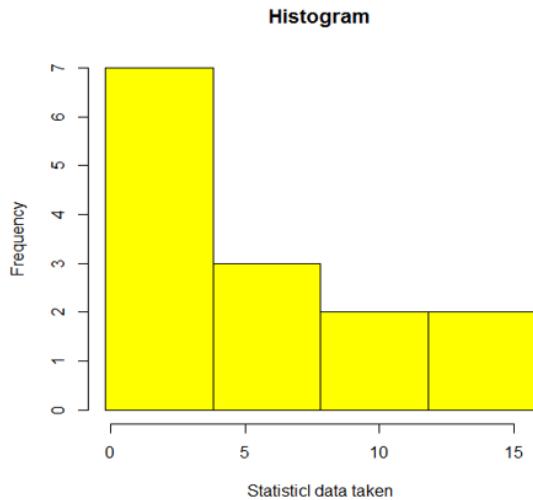
Histogram - how data is distributed in terms of beans or intervals

Histograms

innovate achieve lead

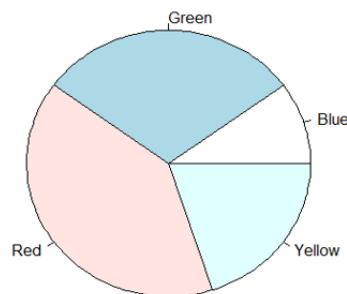
Histogram of Data.test





Pie charts

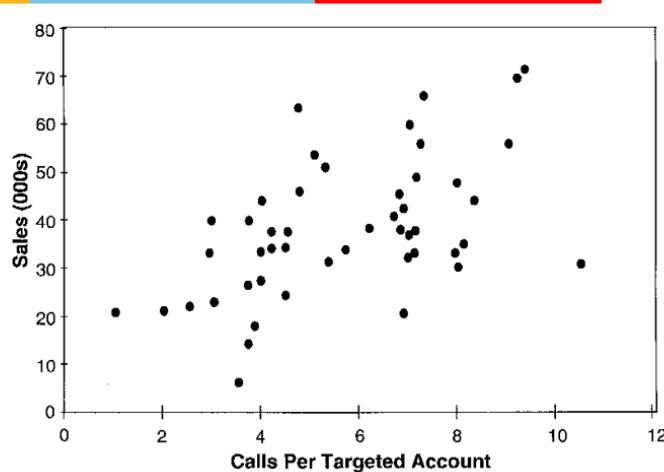
innovate achieve lead



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

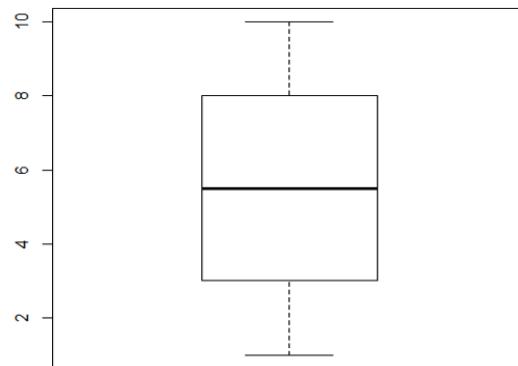
innovate achieve lead



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 $(1/1,000,000) \times (1/1,000,000) = 10^{-12}$. 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

The set with all possible outcomes
for eg:- Tossing of a coin
 $S = \{T, H\}$

Throwing a dice
 $S: \{1, 2, 3, 4, 5, 6\}$

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.

EXP → Tossing of a coin → head event
 Event → getting a head 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...

innovate achieve lead

Probability

$$P(A) = \frac{n}{m} \quad \begin{matrix} \nearrow \text{favourable cases} \\ \searrow \text{total number} \end{matrix}$$

$$P(\bar{A}) = \frac{m-n}{m} : 1 - \frac{n}{m} = 1 - P(A)$$

$$\text{i.e. } P(A) + P(\bar{A}) = 1$$

\downarrow \downarrow
event compl.

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

100% - confirmed (total events) \rightarrow 90% - favorable events

Axioms of Probability

innovate achieve lead

Probability is a number that is assigned to each member of a collection of events from a random experiment that satisfies the following properties:

If S is the sample space and E is any event in a random experiment,

$$(1) \quad P(S) = 1$$

$P(E) = 0 \rightarrow$ impossible event

$$(2) \quad 0 \leq P(E) \leq 1$$

$$P(E) = 1$$

$$(3) \quad \text{For two events } E_1 \text{ and } E_2 \text{ with } E_1 \cap E_2 = \emptyset$$

contain event

$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

Event – getting number less than or equal to 5 – total numbers less than 5 – 0 to 5 \rightarrow 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 – $P(E_1)+P(E_2)$

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

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

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

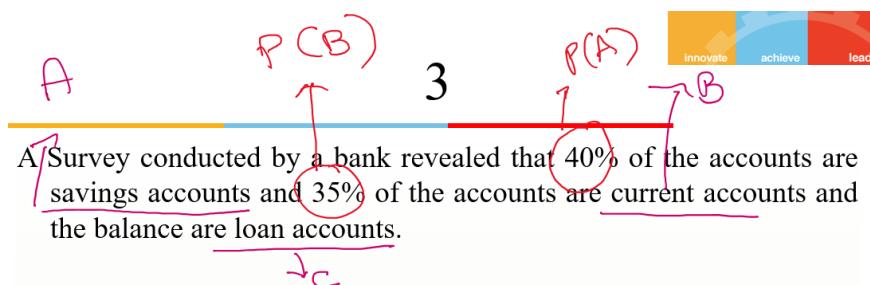
Dependent events (w/o replacement) – probability of second events depends on first event

	Blue	Black	Brown	Total
Software <u>prog</u>	35	25	20	80
Project <u>Mgrs</u>	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? $\frac{80}{100}$

.....42, what
is the probability that he is wearing a blue trouser

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



- What is the probability that an account taken at random is a loan account? $P(C) = 1 - (0.4 + 0.35) = 0.25$
 - What is the probability that an account taken at random is NOT savings account? $P(\bar{A}) = 1 - 0.40 = 0.60$
 - 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? \rightarrow Next session - discu

Lec 3 Descriptive Statistics & Conditional Probability

Today.....

innovate

- 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?

$$\begin{aligned} P(C \cup L) &= P(C) + P(L) \\ &= 0.4 + 0.25 \\ &= 0.65 \end{aligned}$$

Mutually
exclusive events

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \text{ (here 0 for mutually exclusive events)}$$

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.

$A \rightarrow B_P$
 $A \rightarrow \text{Diabetes}$

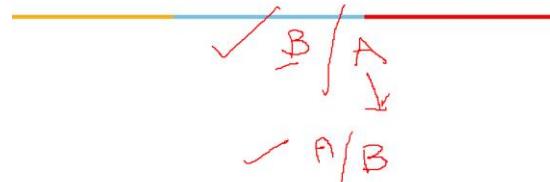
What is the probability that a patient having high BP is also diabetic

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

Conditional Probability

innovate achieve lead

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

$$P(B|A) = P(A \cap B)/P(A) \quad (2-9)$$

for $P(A) > 0$.

Example

innovate achieve lead

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

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

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

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

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

Multiplication Rule

$$P(A \cap B) = P(B|A)P(A) = P(A|B)P(B) \quad (2-10)$$

Total Probability Rule (two events)

For any events A and B ,

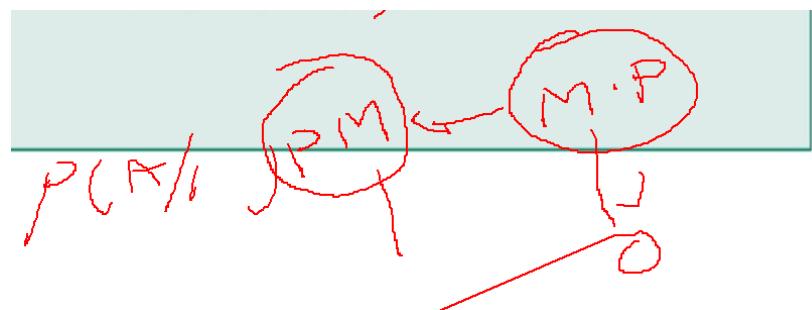
$$P(B) = P(B \cap A) + P(B \cap A') = P(B|A)P(A) + P(B|A')P(A') \quad (2-11)$$

Independence

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.

$$P(B|D) = 0.45$$
$$P(D|B) = 0.45$$

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

$$P(D|B) = ?$$

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

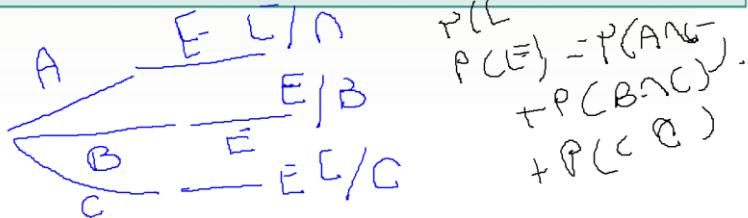
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad \text{for } P(B) > 0 \quad (2-15)$$

Bayes' Theorem

Definition

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

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



Advanced Statistical Techniques for Analytics

04-08-2018

Slide 14

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

For each

$$P(E_i \cap B) = P(B|E_i)P(E_i)$$

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

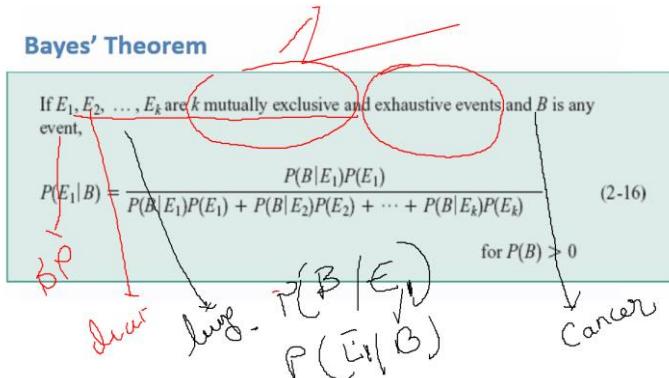
Bayes' Theorem

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

$$P(E_1|B) = \frac{P(B|E_1)P(E_1)}{P(B|E_1)P(E_1) + P(B|E_2)P(E_2) + \dots + P(B|E_k)P(E_k)} \quad (2-16)$$

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 – drunk 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.

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

Applications

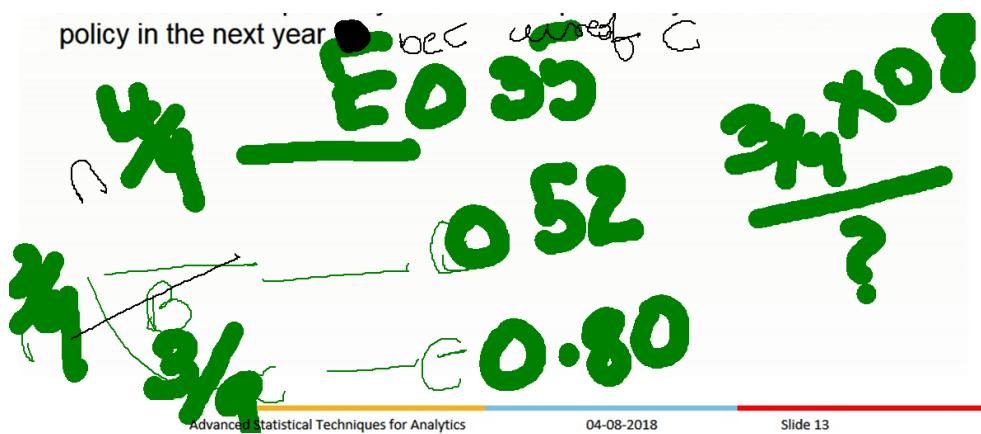
- Diagnostic tests in medicine
- Telecommunication
- Customer service
- Trouble shooting in engineering processes & systems

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.....

innovate achieve

- 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

innovate achieve

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

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

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

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

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

$$\begin{aligned} P(A|E) &= \frac{P(A \cap E)}{P(E)} \\ &= \frac{P(E|A) P(A)}{\sum_i P(E|A_i) P(A_i)} \\ &\downarrow \\ &\text{Baye's theorem}' \end{aligned}$$

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

Baye's theorem

innovate achieve lead

$$P(E) = \sum_{A,B,C} P(E|A)P(A)$$
$$P(A|E) = \frac{P(E|A)P(A)}{\sum P(E|A)P(A)}$$
$$P(B|E) = \frac{P(E|B)P(B)}{\sum P(E|A)P(A)}$$

Expecting some examples

or case studies

or applications from

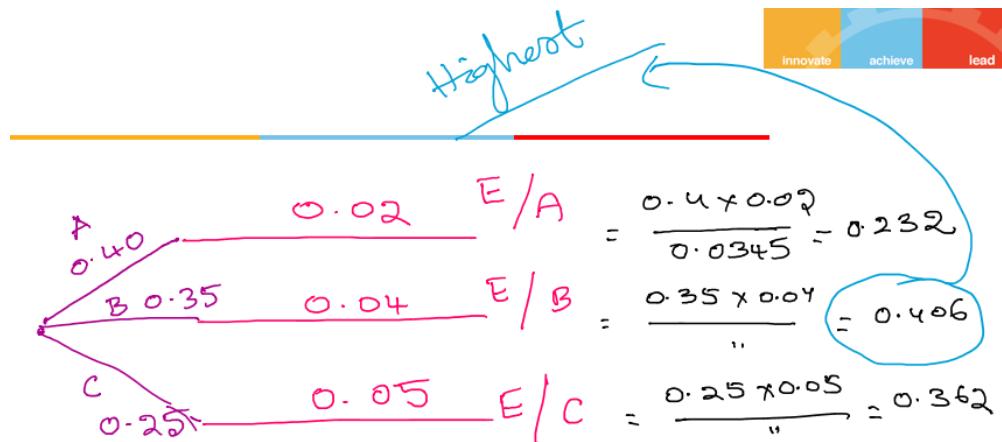
Participants

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

Example :-

innovate achieve lead

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 output 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?



Total probability

$$\begin{aligned}
 &= P(E|A) P(A) + P(E|B) P(B) + P(E|C) P(C) \\
 &= (0.02)(0.4) + (0.04)(0.35) + (0.05)(0.25) \\
 &= 0.0345
 \end{aligned}$$

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 – $X_c = 1$ – 1 Head, $X_c = 2 \rightarrow$ 2 Head – X_c = 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



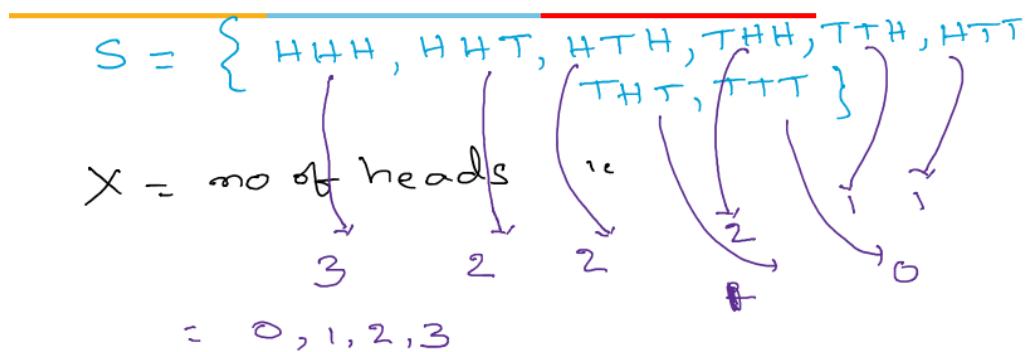
Examples of Random Variables

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



X	0	1	2	3
$P(X)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

$P(x) = ?$

We can have function – $fn(S) \rightarrow x$ – 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

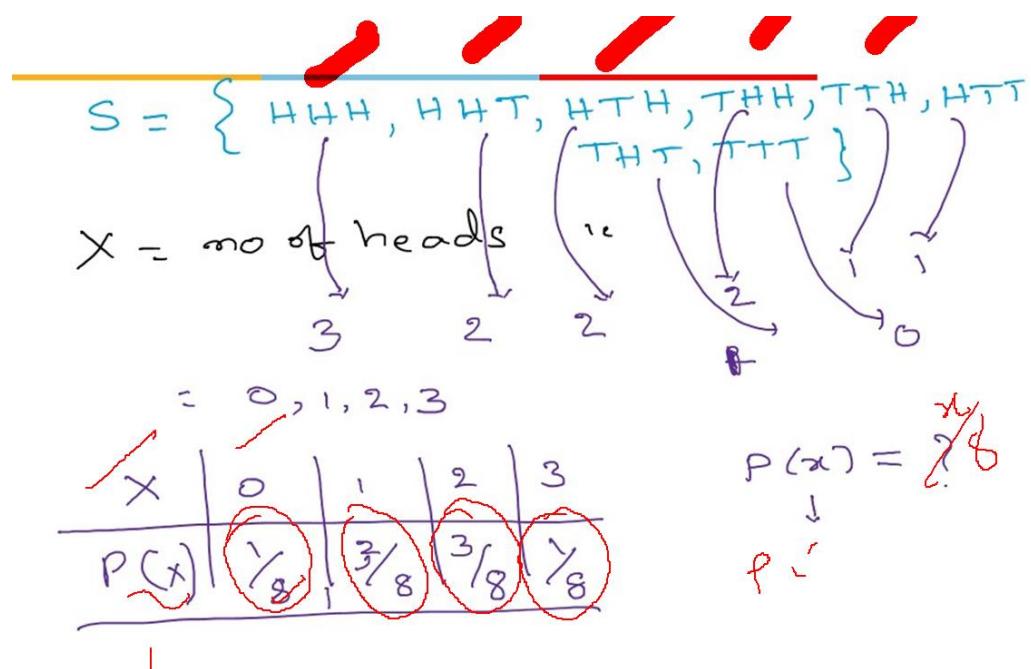
$P(x)$ – Probability distribution function – like frequency distribution – how probabilities are distributed – at $x = 0, x = 1 \dots$

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

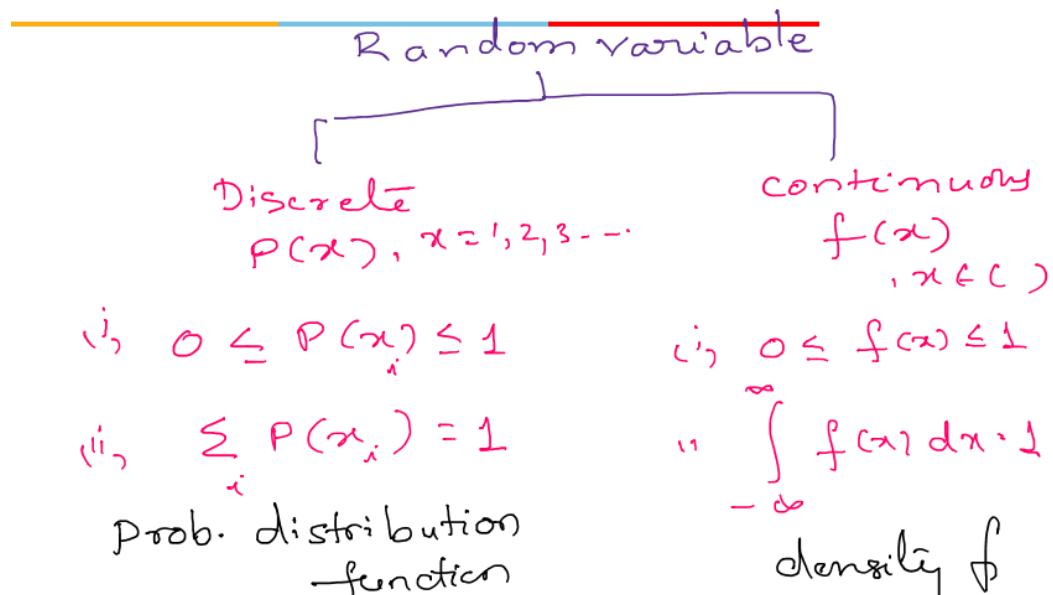
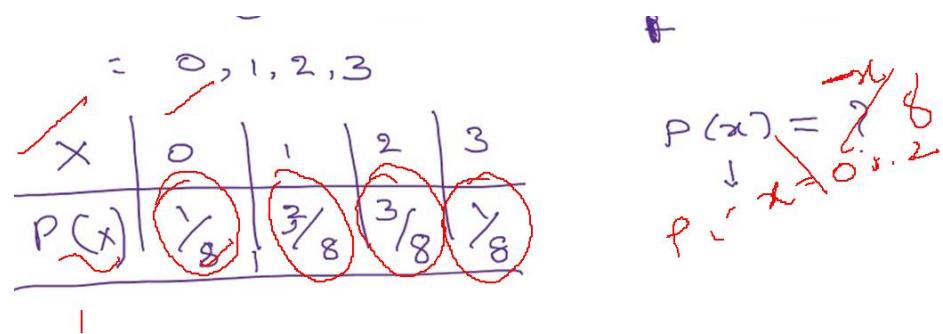
Observation – All values lies between 0 & 1 – all Heads & No Heads – are equally distributed

Frequency distribution - Distribution of marks – students – range - 10-20, range – 20-30 etc.
– class strength = 60 \rightarrow 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

$P(x) = 0 \rightarrow$ impossible event – rarest event – for any value of $x \rightarrow P(x) = 0$

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

Validation

$$\therefore P(x) = \frac{x-3}{2}$$

$$x = 1, 2, 3, 4$$

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

$$P(1) = \frac{1-3}{2}$$

$$< 0 \times$$

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

another

Innovate achieve lead

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

$$\text{Is } \sum P(x) = 1$$

Example:-

Innovate achieve lead

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

i, k value:?

(i), $P(x < 6)$

(ii), $P(x \geq 6)$

(iii), $P(3 < x \leq 6)$

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

i, k value:?

∴ $0+1k+2k+2k+3k+1k^2+2k^2+7k^2+k=1$

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

(i), $P(x < 6)$

$$0+1k+2k+2k+3k+1k^2+2k^2+7k^2+k = 1$$

(ii), $P(x \geq 6)$

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

(iii), $P(3 < x \leq 6)$

$$K = -1/10$$

Example:-

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

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

$$(i), P(x < 6) \rightarrow 0 + 1K + 2K + \dots = 1$$

$$(ii), P(x \geq 6) \rightarrow 10K^2 + 9K - 1 = 0$$

$$(iii), P(3 < x \leq 6) \rightarrow K = -1, 1/10$$

Example:-

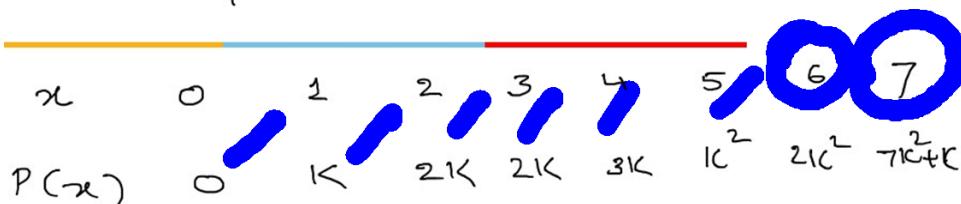
x	0	1	2	3	4	5	6	7
$P(x)$	0	K	$2K$	$2K$	$3K$	$1K^2$	$2K^2$	$7K^2 + K$

i, K value? $K = 1/10$

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

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

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



$$P(x \geq 6) = 1 - P(x < 6)$$

Example:

innovate achieve

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

$$K = ?$$

$$P(1 < x < 2)$$

Example:

innovate achieve lead

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

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

$$\Rightarrow K \left[\frac{x^3}{3} \right]_0^3 = 1$$

$$\Rightarrow K \cdot \frac{27 - 0}{3} = 1 \quad \therefore \quad \boxed{K = \frac{1}{9}}$$

Example:

innovate achieve lead

P

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

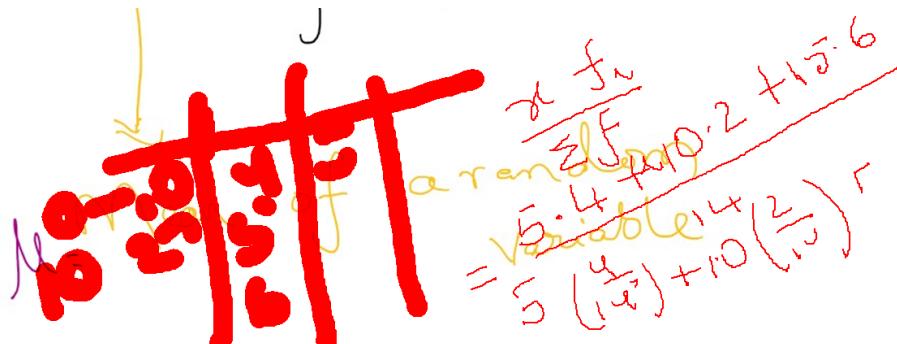
$$\begin{aligned} K = ? &\rightarrow \frac{1}{9} & \int_0^2 Kx^2 dx \\ P(1 < x < 2) &= \frac{1}{9} \left[\frac{x^3}{3} \right]_1^2 \\ &= \frac{1}{9} [8 - 1] = \boxed{\frac{7}{27}} \end{aligned}$$

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



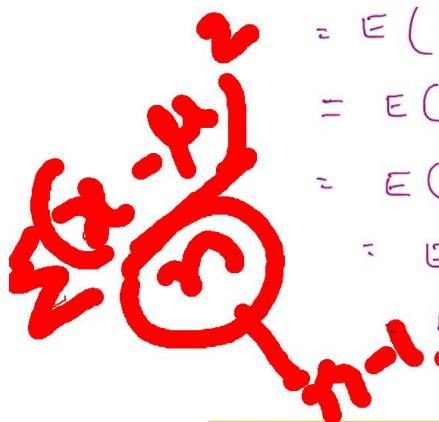
4 / 14, 2/14 – probabilities – 4,2 etc. in corresponding space – 14 – sample space

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



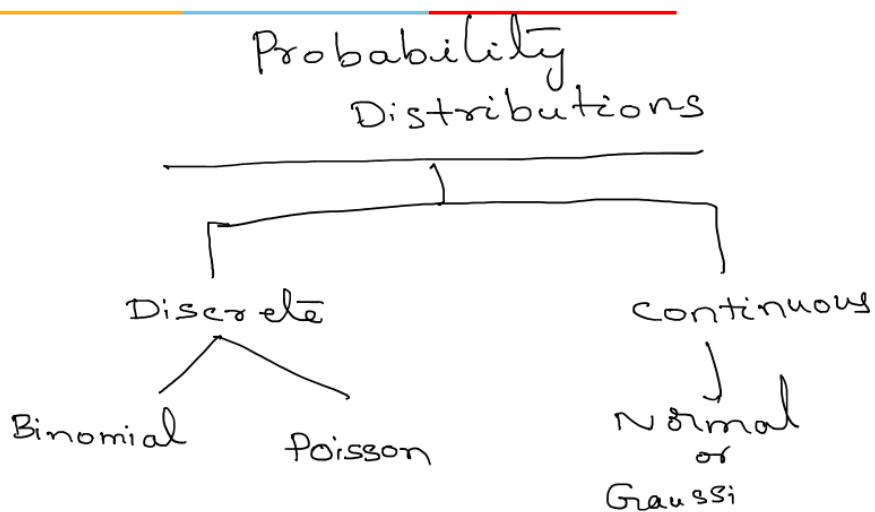
$$\begin{aligned}
 &= E(x^2 + \mu - 2\mu x) \\
 &= E(x^2) + \mu^2 - 2\mu E(x) \\
 &= E(x^2) + \mu^2 - 2\mu \cdot \mu \\
 &= E(x^2) - \mu^2 \\
 &= E(x^2) - [E(x)]^2
 \end{aligned}$$

Sampling - n → (n-1)

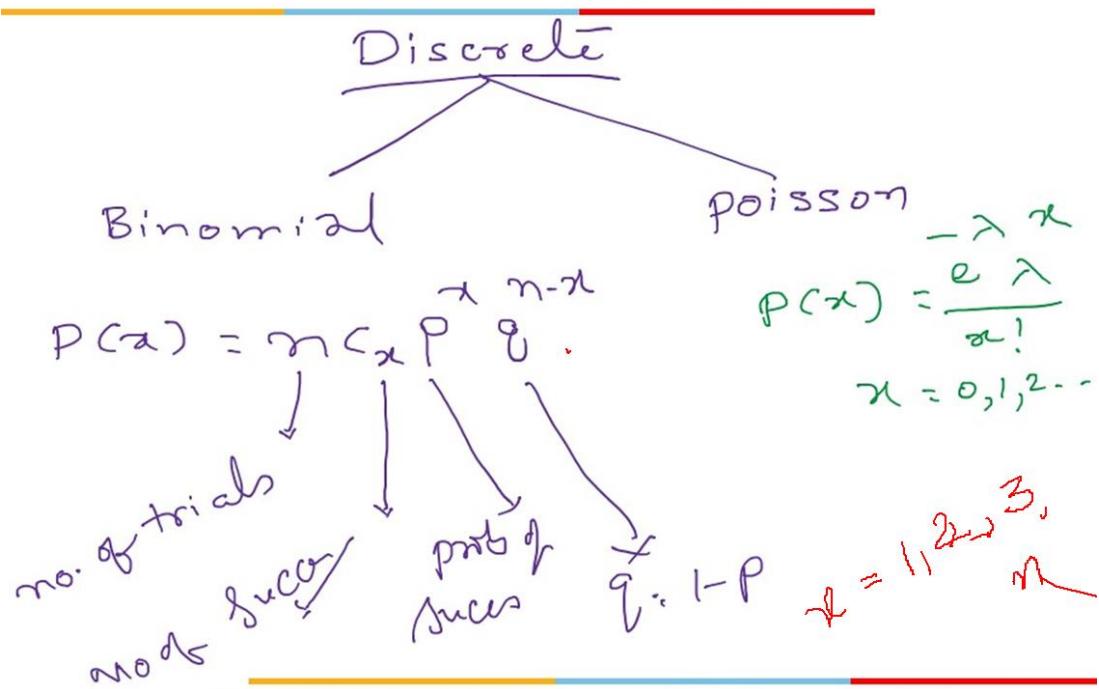
i.e. mean $\mu = E(x) = \sum x p(x)$
 $= \int x f(x) dx$

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

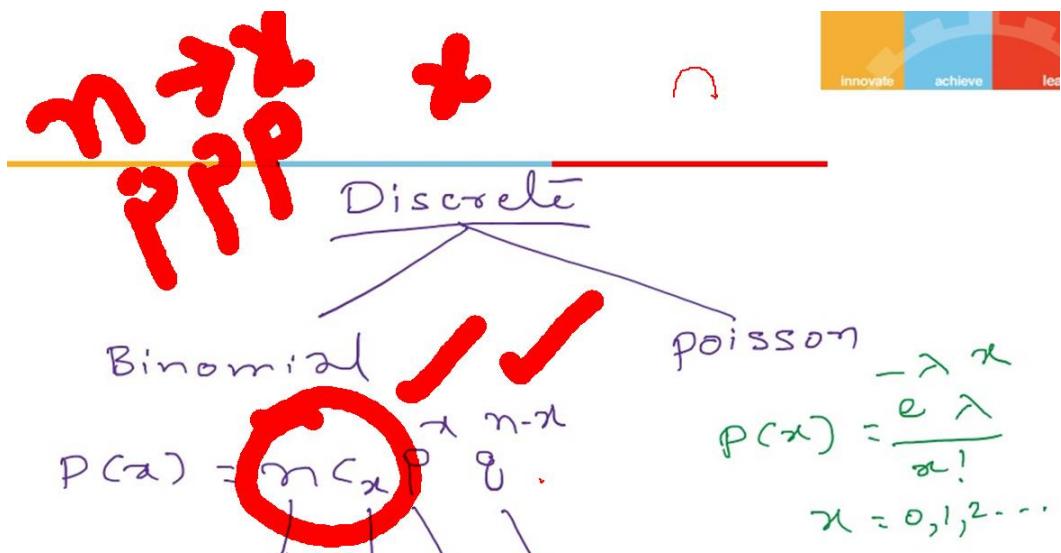
$E(x^2) = \sum x^2 p(x) \rightarrow \text{discrete}$
 $\int x^2 f(x) dx \rightarrow \text{continu.}$



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 → $\frac{1}{2}$, Head/Tail → Trial 2 → $\frac{1}{2}$ - independent & constant



n times – p, (n-x) times – q

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

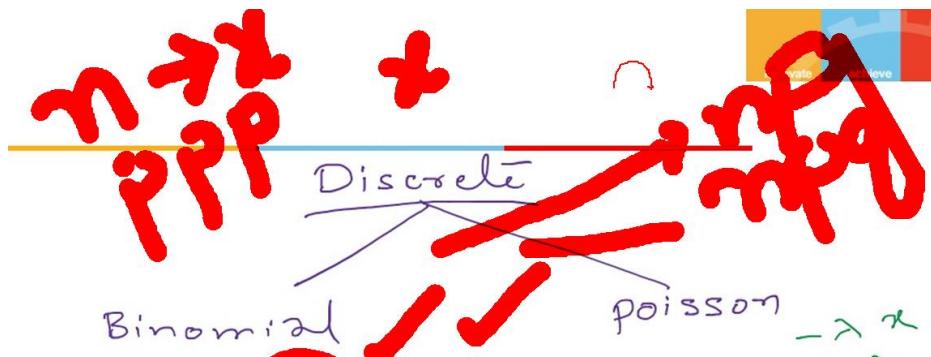
Binomial Expansion – $(p+q)^n = (1)^n = 1$

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 \rightarrow Lambda – Poisson

Poisson – mean = variance = Lambda

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

Lec 5 Descriptive and inferential statistics

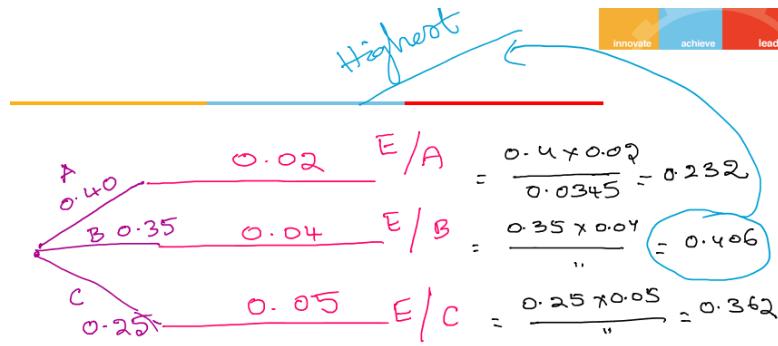


L- 5: Descriptive and inferential statistics

Agenda

innovate achieve lead

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



Total probability

$$\begin{aligned}
 &= P(E|A) P(A) + P(E|B) P(B) + P(E|C) P(C) \\
 &= (0.02)(0.4) + (0.04)(0.35) + (0.05)(0.25) \\
 &= 0.0345
 \end{aligned}$$

Number of products defected – A $\rightarrow 0.4 * 0.232$, 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,

$$\begin{aligned}
 P(B_1|A) &= \frac{P(B_1)P(A|B_1)}{P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + P(B_3)P(A|B_3) + P(B_4)P(A|B_4)} \\
 &= \frac{(0.20)(0.05)}{(0.20)(0.05) + (0.60)(0.10) + (0.15)(0.10) + (0.05)(0.05)} \\
 &= 0.114
 \end{aligned}$$

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?

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

$$P(x=0) = \frac{e^{-5} 5^0}{0!} = \frac{e^{-5}}{1} = e^{-5}$$

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?

$$\lambda = 50,$$

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

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

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 – $n \cdot p$ is almost constant or don't have n value clearly

Normal distribution

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

$$-\infty < x < \infty$$

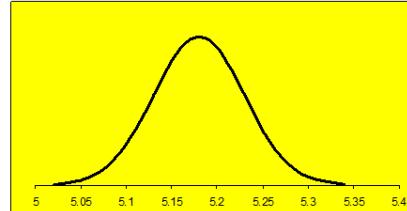
Naïve based algorithm – numerical data, any sample – repeatedly – becomes Normal Distribution, Signal processing – communication – Gaussian curve – important

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

Normal Distribution

innovate achieve lead

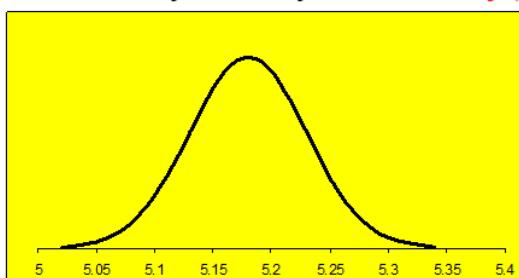
Probability density function - $f(X)$



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

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

Probability density function - $f(X)$

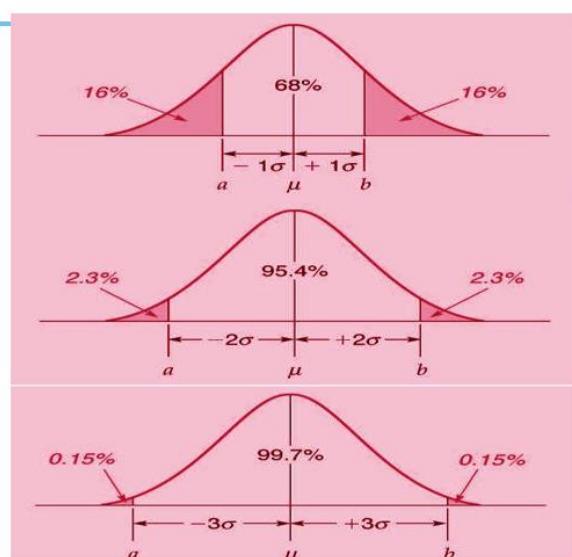


normal curve
or
Gaussian
curve

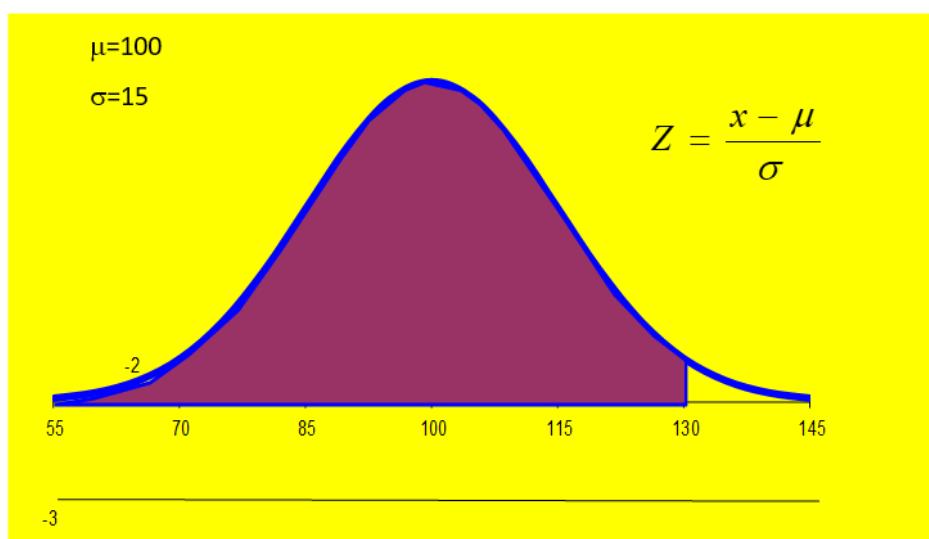
Three Common Areas Under the Curve

innovate achieve lead

Three Normal distributions with different areas



Standard Normal Distribution



How to find



$$P(2 < x < 5)$$

$$\begin{aligned} &= \int_{-2}^5 f(x) dx \\ &= \int_{-2}^5 \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\ &\downarrow \\ &\text{Integration } ???? \end{aligned}$$

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)

$$P(x_1 \leq x \leq x_2) = \int_{x_1}^{x_2} f(x) dx$$

Let $\frac{x-\mu}{\sigma} = z$ i.e. $dx = \sigma dz$

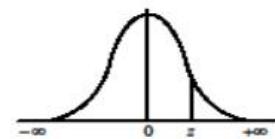
$$= \int_{z_1}^{z_2} \frac{1}{\sigma \sqrt{2\pi}} e^{-z^2/2} \cdot \sigma dz$$

$$= \left(\frac{1}{\sqrt{2\pi}} \int_{z_1}^{z_2} e^{-z^2/2} dz \right) \rightarrow F(z)$$

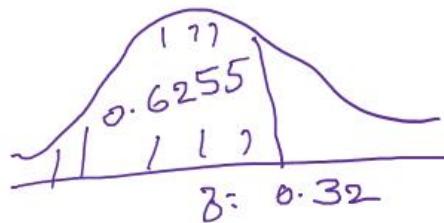
$$= F(z_2) - F(z_1)$$

Probability of x lying between x_1 & x_2 – 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



$z = 0.32 \rightarrow 0.6255 \rightarrow$ limit is -infinity to 0.2 \rightarrow Integration is 0.6255 – area under curve

$0.8413 \rightarrow$ -infinity to 1 \rightarrow area under curve

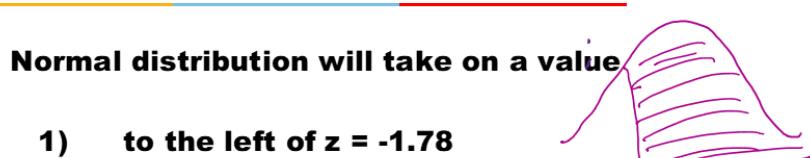
Area under curve – both side – positive & negative – probability – 0.5 – two halves - similar



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 $-0.80 \leq z \leq 1.53$
- 4) to the left of $z = -2.52$ and to the right of $z = 1.83$

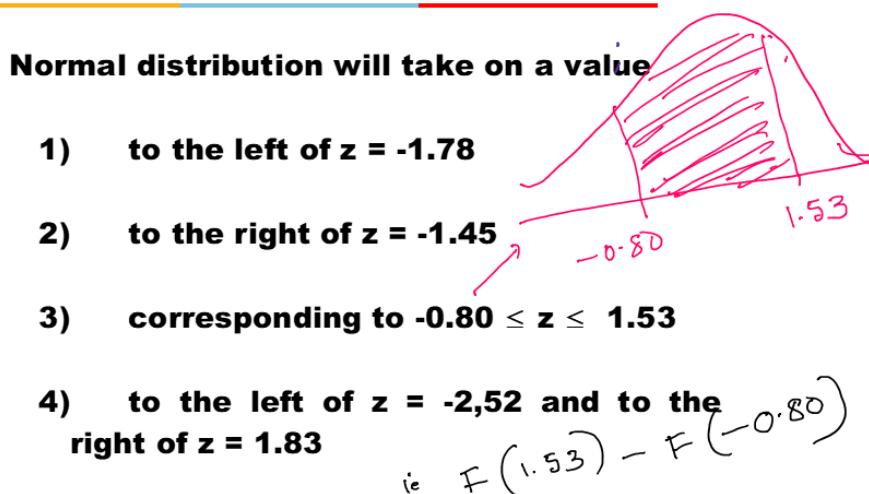
$P(z) < -1.78 \rightarrow$ -1.78 – left side of the curve



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 $-0.80 \leq z \leq 1.53$
- 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 – $(1 - \text{value from log})$



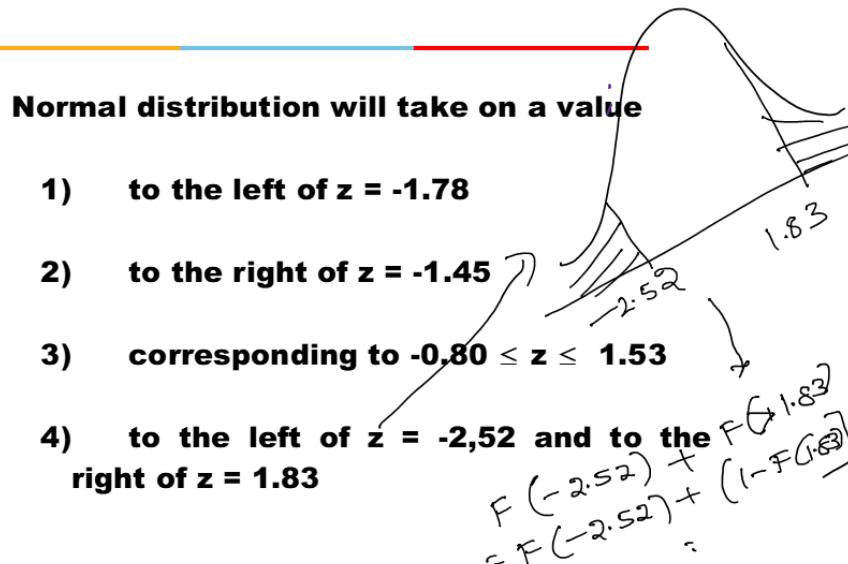
Normal distribution will take on a value

- 1) to the left of $z = -1.78$
- 2) to the right of $z = -1.45$
- 3) corresponding to $-0.80 \leq z \leq 1.53$
- 4) to the left of $z = -2.52$ and to the right of $z = 1.83$
i.e. $F(1.53) - F(-0.80)$

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

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

$$F(z) = 1 - F(-z) \rightarrow F(-1.0) = 1 - F(1.0) = 1 - 0.8413 = 0.1587$$



Calculation of probabilities using a normal distribution

Problem

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

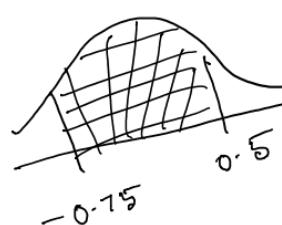
Find 1) $P[5 \leq X \leq 10]$
2) $P[X \geq 5]$

Solution

1) $\mu = 8$

$\sigma = 4$

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



$$\text{When } X=5 \quad Z = \frac{5-8}{4} = -0.75 \quad = F(0.5) - F(-0.75)$$

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

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

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

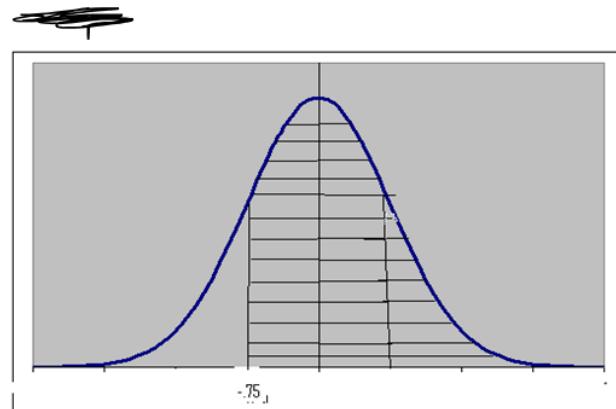
2) $P[X \geq 5] = P[Z \geq -0.75] = 1 - F(-0.75)$



$$= F(0.75)$$

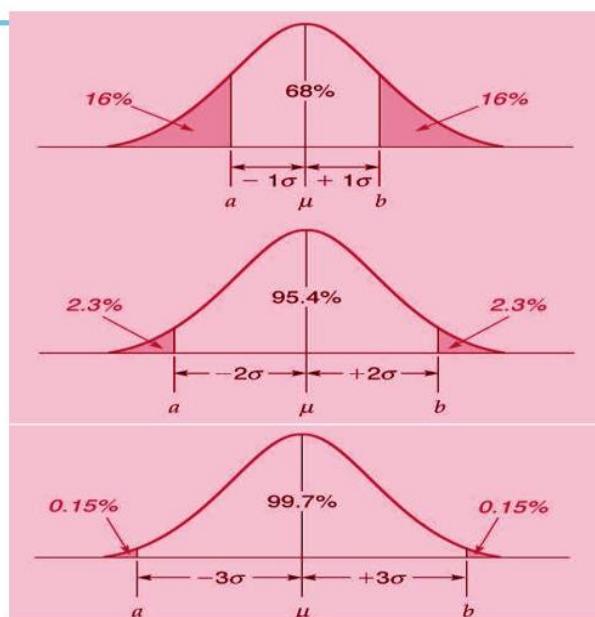
~~0.7734~~

~~= 0.7734~~



Three Common Areas Under the Curve

Three Normal distributions with different areas



Example:

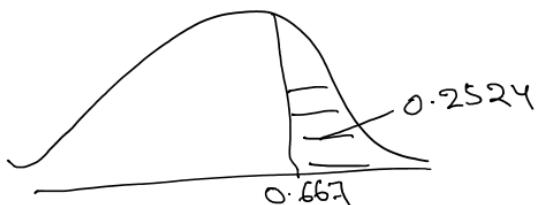
In a test conducted on 1000 candidates, the average score is 42 with a S.D of 24. Assuming normal distribution, find

- no of candidates whose score exceeds 58
- no of candidates whose scores lies b/w 30 and 66

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

$$P(x > 58) \quad z = \frac{x - \mu}{\sigma}$$
$$= \frac{58 - 42}{24}$$

$$P(z > 0.667) \quad = 0.2524$$



1000×0.2524
252 Candidates

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

$$\frac{30 - 42}{24} = -0.5$$

$$\frac{66 - 42}{24} = 1$$

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

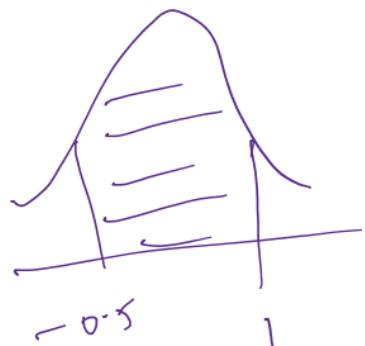
$$= F(1) - F(-0.5)$$

$$= 0.5328$$

$$0.5328 \times 1000$$

$$= 532.8$$

533



Suppose find $P(100 < x < 1000) \rightarrow$ we have to do like this – $P(100) + P(101) + \dots + P(1000)$

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. ($\lambda = 50$).

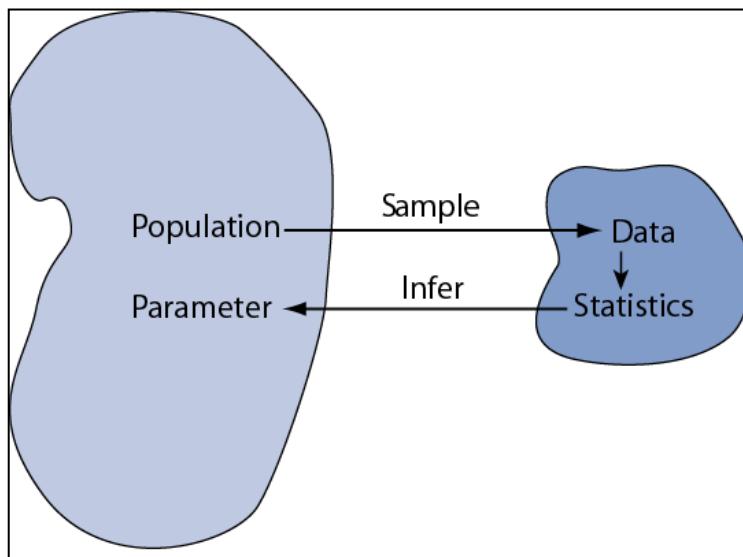
What is the probability that there are between 48 and 50 customers, inclusive?

$$\begin{aligned} X &= 50 \\ P(48 \leq X \leq 50) &= P(48) + P(49) + P(50) \\ &= \frac{e^{-50} 50^{48}}{48!} + \frac{e^{-50} \cdot 50^{49}}{49!} + \frac{e^{-50} \cdot e^{50}}{50!} \end{aligned}$$

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

Estimation

innovate achieve

Point
Estimation

Interval
Estimation

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

Interval Estimation

Sampling → \bar{x} : Mean

mean of the population

$$\left(\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \right)$$

with (1 - α) level of significance

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 $\rightarrow n-1 \rightarrow$ 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 $x + y = 5 \rightarrow$ what are the unique values of x & y that satisfy this equation – when we can have unique selection – if $x = 3$ (known or restricted) \rightarrow 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 $(-\alpha/2, \alpha/2)$ is $1 - \alpha$

In generic terms – **sampling mean** – x & σ/\sqrt{n}

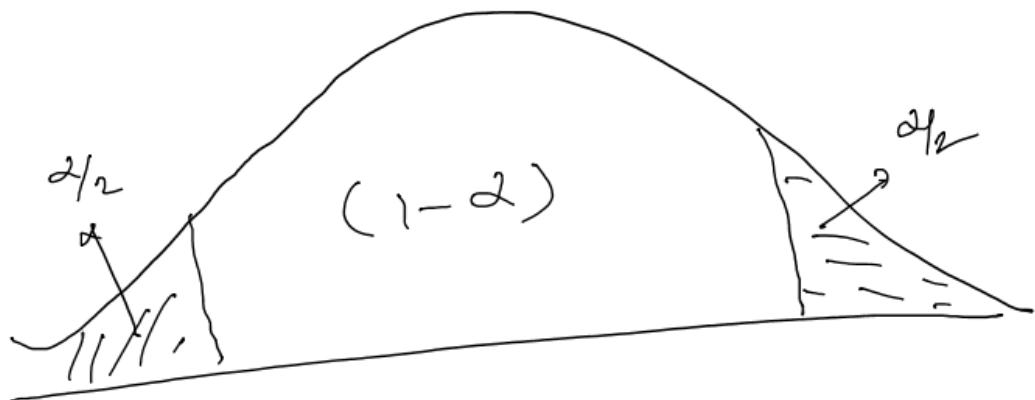
How?

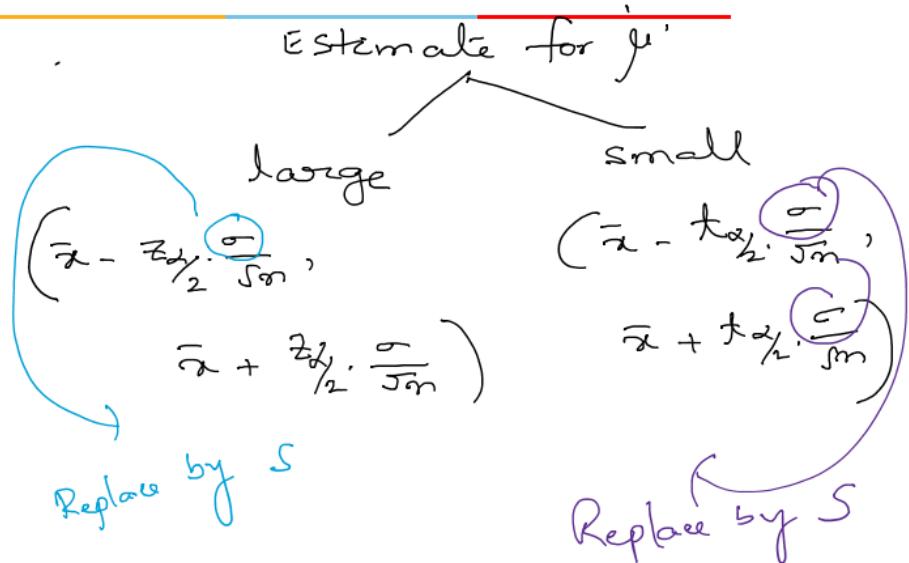
innovate achieve lead

$$P\left(-z_{\alpha/2} < z < z_{\alpha/2}\right) = 1 - \alpha$$

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < \frac{z_{\alpha/2} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}$$

$$P\left(\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu < \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$





$$\text{Confidence} = 95\%, \alpha = 5\%, \alpha/2 = 2.25\%$$

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.

\bar{x} for the mean.

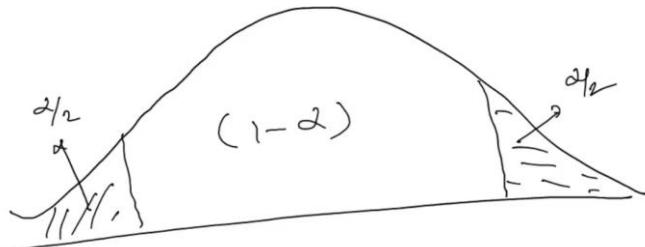
95% is 2σ suppose

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

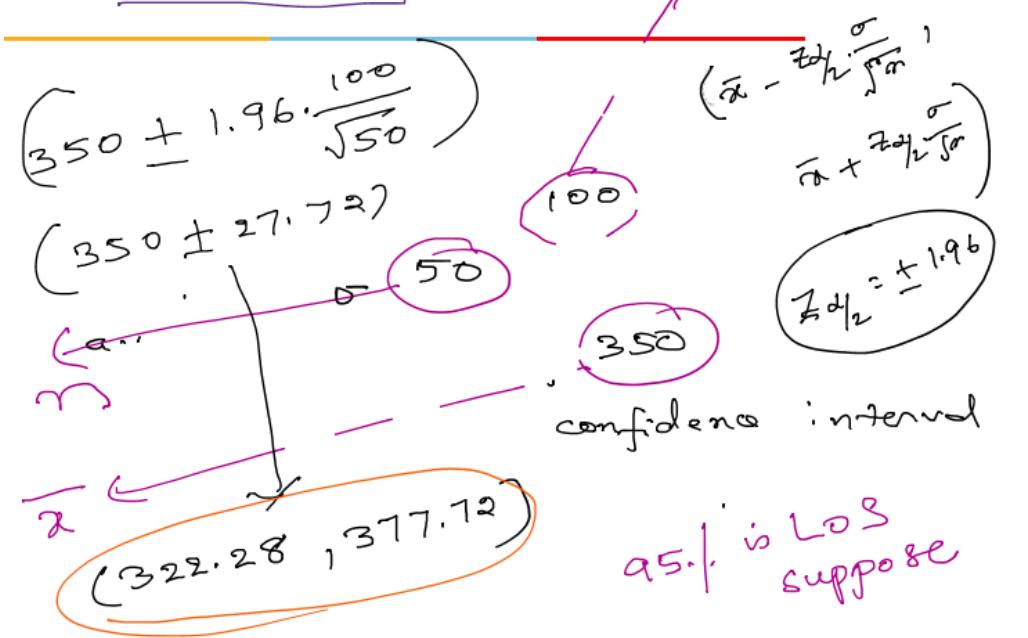
If bike production – sample – mileage 80 kmpl with s.d. → generalize – entire population estimation – almost same - we can proceed further

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

$$\alpha/2 = 2.25\% = 0.025 \text{ -- area under curve}$$



Example:



Sample Size



$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

$$n = \left(\frac{\sigma \cdot z_{\alpha/2}}{E} \right)^2$$

With help of sample - \bar{x} (mean) – we are estimating → μ (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

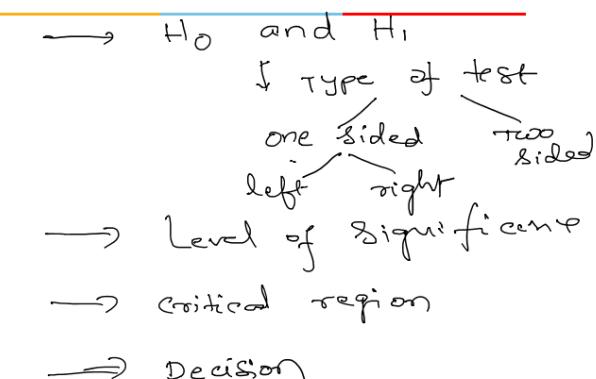
Firm runs clinical trial where some patients receive new drug, and others receive standard treatment

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

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

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:

$$H_0: \mu = 100 \text{ versus}$$

$$H_a: \mu > 100 \text{ (one-sided)}$$

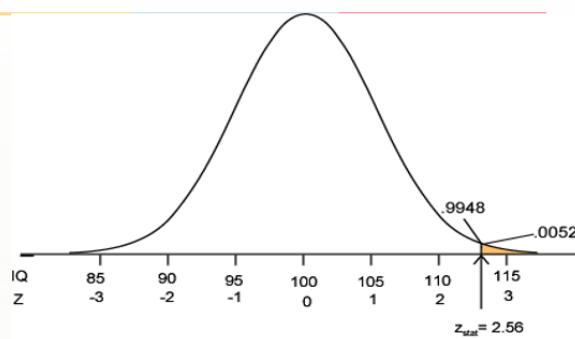
$$H_a: \mu \neq 100 \text{ (two-sided)}$$

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: $P = \Pr(Z \geq 2.56) = 0.0052$



$P = .0052 \Rightarrow$ it is unlikely the sample came from this null distribution \Rightarrow 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

$$\alpha = P(\text{Type I Error}) \quad \beta = P(\text{Type II Error})$$

- Goal: Keep α, β reasonably small

Problem

Innovate achieve

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 27,463 miles 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

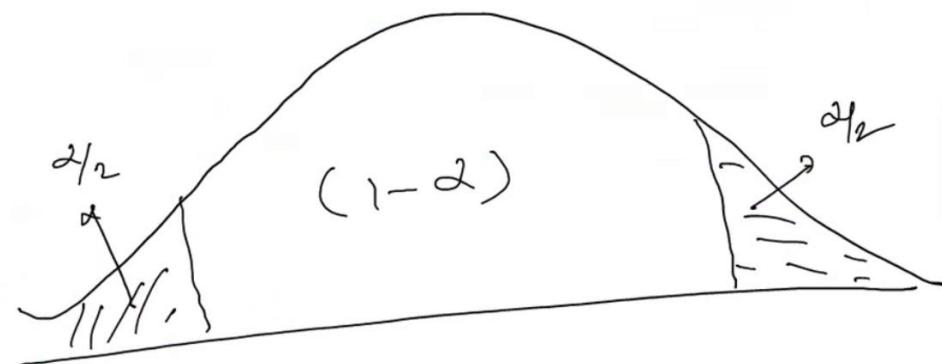
Claim = Assumption = Hypothesis

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

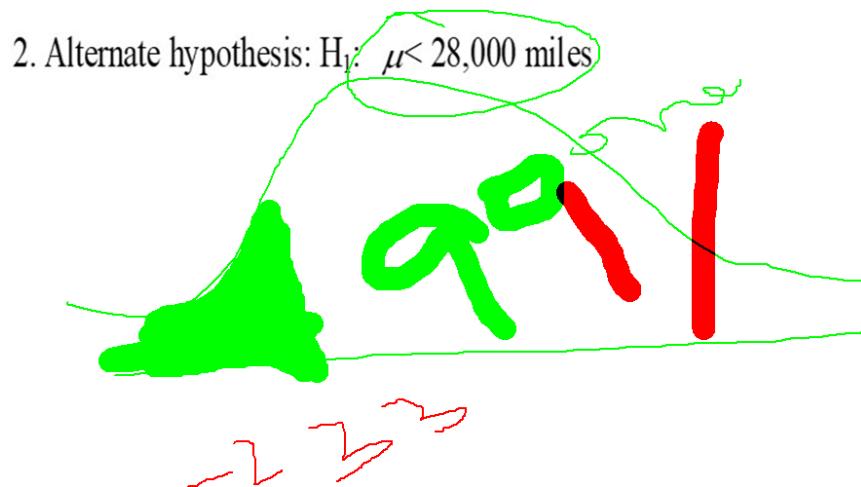
Civil Engineering application – pipe – leakage – 5% of chance – we can fix it,

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

Based on alternative hypothesis \rightarrow here $\mu < 28000$ – lies one side \rightarrow one tailed test, if we have $\mu = 28000$, alternative $\mu \neq 28000 \rightarrow \mu < 28000 \& \mu > 28000$ – two tailed test \rightarrow Critical region – it defines $\alpha/2$ – Critical region or region of acceptance



Two tailed test – right & left test $\rightarrow \alpha/2$, 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 $Z = Z_{\text{cal}} < -Z_\alpha$ we reject null hypothesis

If $Z = Z_{\text{cal}} < -Z_\alpha = -Z_{0.01} = -2.33$ 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 $Z = Z_{\text{cal}} = -2.52 < -2.33$, 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

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

6. Calculation

7. Decision

Problem

innovate achieve lead

It is claimed that a random sample of 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 $\alpha = 0.05$

$$\begin{aligned}\frac{\alpha}{2} &= \frac{0.05}{2} \\ &= 0.025\end{aligned}$$

From table we get

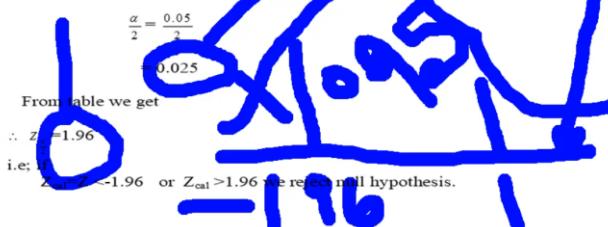
$$\therefore Z_{\frac{\alpha}{2}} = 1.96$$

i.e; if

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

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 $\alpha = 0.05$



6. Computation :

Test statistic

$$\begin{aligned}Z_{cal} = Z &= \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{15200 - 15150}{1200 / \sqrt{49}} \\ &= 0.2916\end{aligned}$$

7. Decision:

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

First check hypothesis for 90% - pass \rightarrow 95% \rightarrow pass \rightarrow 99% - confidence

We can change the claim & confidence \rightarrow tool available – check \rightarrow 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 μ

$\mu = 58,000$
 $\mu \neq 58,000 \rightarrow$ two tailed test
 $n = 6$ → small sample t-test.

Lec 7 Inferential statistics & Predictive Analytics

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 $n \rightarrow \infty$.

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^2 , now $Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ 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 $\rightarrow n > 30$

Examples



- random noise in Comm. Systems
- Errors in Lab measurements
- Errors in regression analysis etc

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

Errors

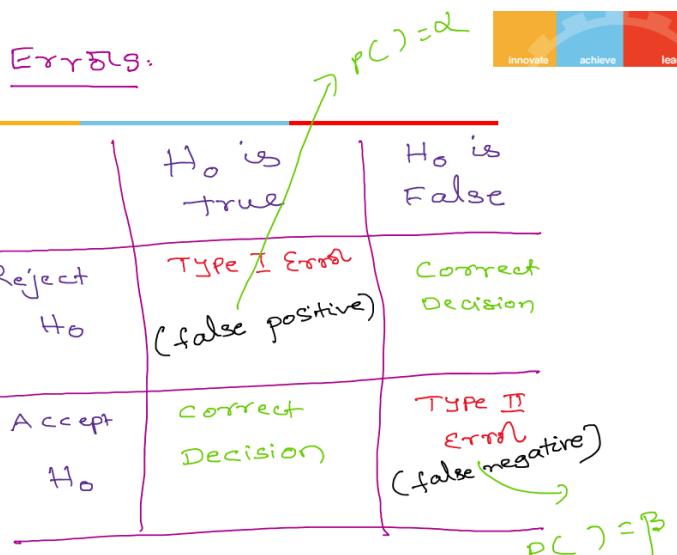


H_0 is true or H_0 is False
decision Accept H_0 or reject H_0

- ① reject H_0 , when it is false ✓
- ② reject H_0 , when it is TRUE } ✗
- ③ Accept H_0 , when it is false } ✗
- ④ Accept H_0 , when it is TRUE ✓

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

2) & 3) – important \rightarrow 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

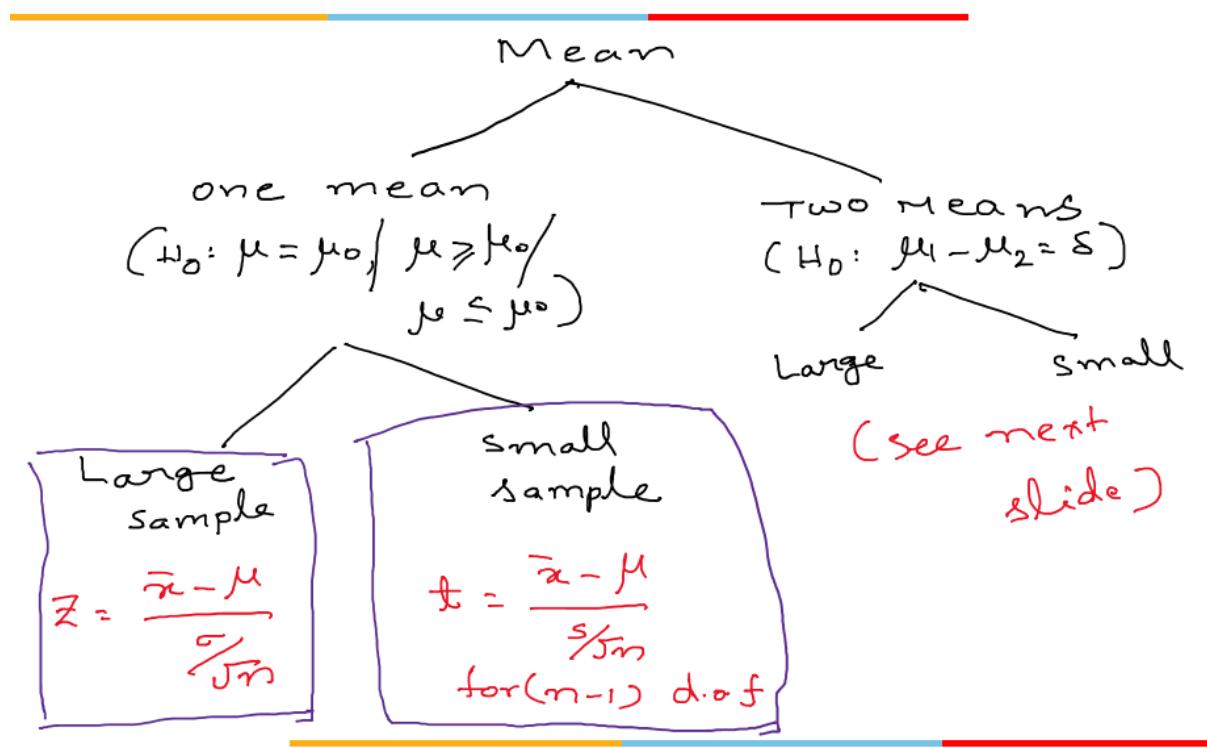
P (Type II error) → Beta – LOC – level of confidence – True Negative

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

Testing of Hypothesis or

Hypothesis testing

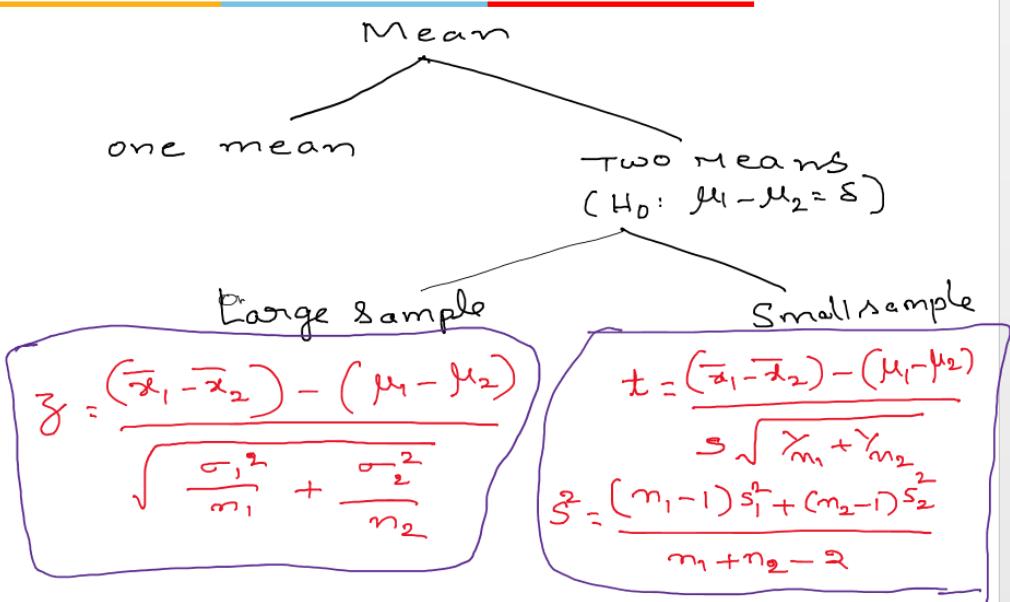
Claim related mean – two means or one mean



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 → $n_1+n_2 - 2$



Testing of Hypothesis

Example - 1

Example:- ①



Can it be concluded that the average life span of Indians is more than 70 yrs.

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 70 yrs.

If a random sample of 100 Indians has average life span of 71.8 years with a S.D of 8.9 years.

$H_0: \mu > 70$

Validation

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

population

Can it be concluded that the average life span of Indians is more than 70 yrs. 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

$$\therefore z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \text{ or } \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$$H_0: \mu > 70$$

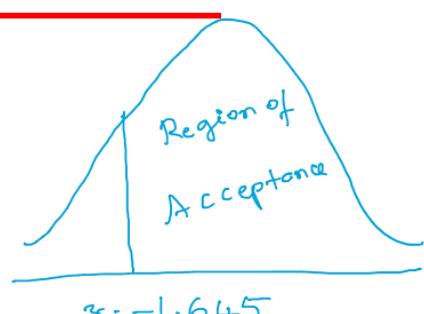
$$H_1: \mu \leq 70$$

$$\alpha = 5.1\% \text{ (Left)}$$

Left tailed test



$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{71.8 - 70}{\frac{8.9}{\sqrt{100}}} \\ = 2.022$$



Lies in the region of acceptance

$\therefore H_0$ 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

$$H_0: \mu > 70$$

$$H_1: \mu \leq 70$$

$$\alpha = 5.1\% \text{ (Left)}$$

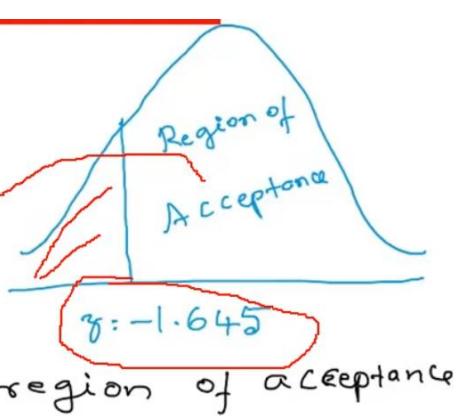
left tailed test

Large Sample – Z distribution → no degree of freedom

$$\alpha = 5\% \text{ (Let)}$$

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

$$= 2.022$$



Lies in the region of acceptance

$\therefore H_0$ is accepted
i.e. Avg life is more than 70 years

Testing of Hypothesis

Example - 2.

Example - 2

innovate achieve lead

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.

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 S.I. Level of significance)

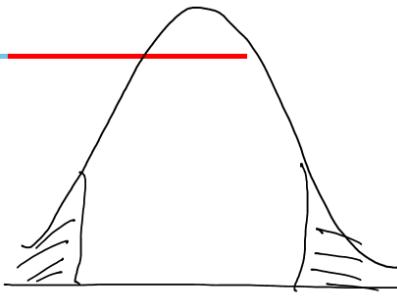
$$\bar{x} \quad s$$

$$H_0: \mu = 10$$

$$H_1: \mu \neq 10$$

$$\alpha = 0.05$$

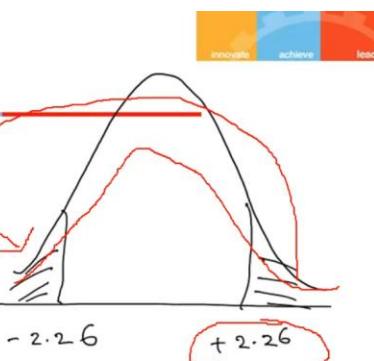
$$\begin{aligned} t &= \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \\ &= \frac{9.52 - 10}{\frac{0.6}{\sqrt{10}}} \\ &= -2.52 \end{aligned}$$



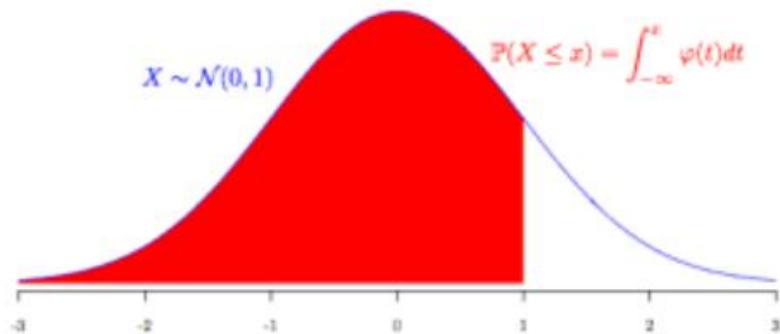
Reject H_0

$$\begin{aligned} H_0: \mu = 10 \\ H_1: \mu \neq 10 \\ \alpha = 0.05 \end{aligned}$$

$$\begin{aligned} t &= \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \\ &= \frac{9.52 - 10}{\frac{0.6}{\sqrt{10}}} \\ &= -2.52 \end{aligned}$$



Reject H_0



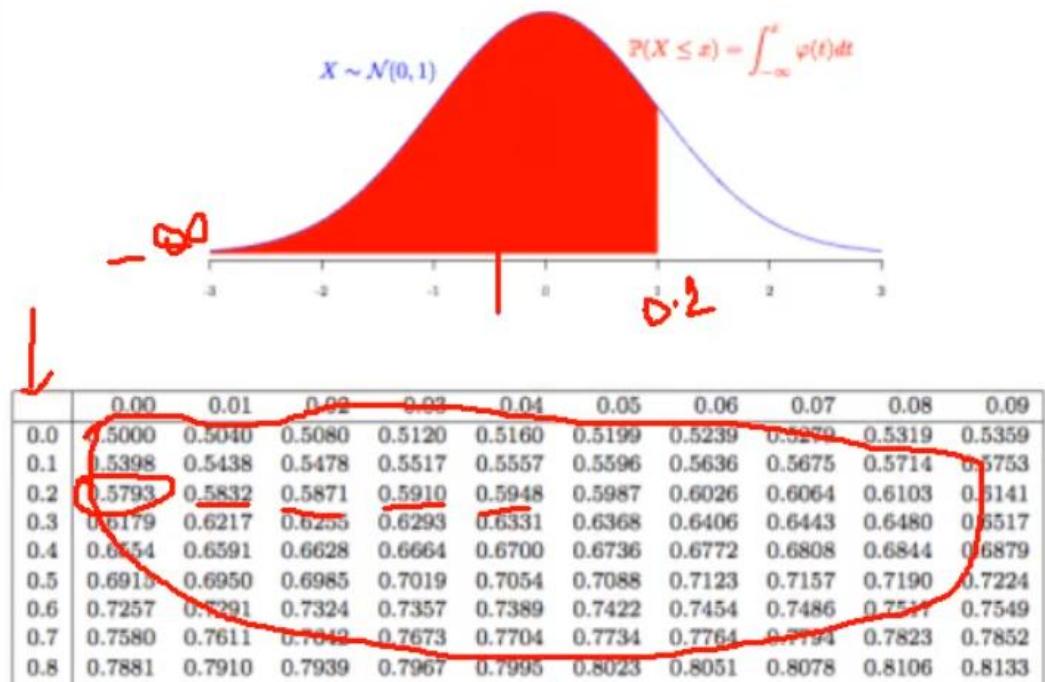
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	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

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{71.8 - 70}{8.9/\sqrt{100}} = 2.022$$

Region of Acceptance

$\gamma = -1.645$



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 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 to mean life

If one sample is large i.e. 35, other one is small i.e. 15 $\rightarrow n_1 + n_2 - 2 = 50 - 2 = 48 \rightarrow z$

Example - 3) Solution?



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

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

$$\text{is } \mu_1 = \mu_2$$

Solution:-



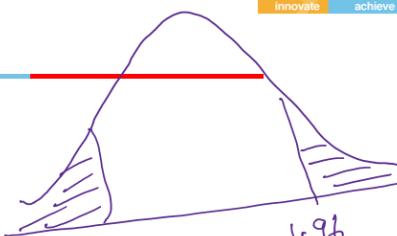
$$H_0: \mu_1 - \mu_2 = 0$$

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

$$\alpha = 0.05$$

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(647 - 638) - 0}{\sqrt{\frac{(27)^2}{40} + \frac{(31)^2}{40}}} = 3.73$$

Reject H_0



Solution:-

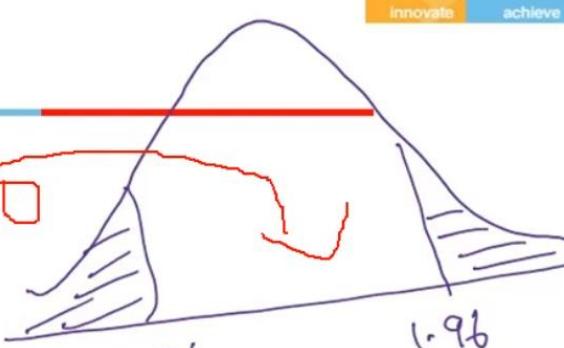


$$H_0: \mu_1 - \mu_2 = 0$$

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

$$\alpha = 0.05$$

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



Example - (3)

earlier
problem
or 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.

$$\text{i.e. } \mu_1 > \mu_2$$

Solution:-

right
sided test



$$H_0: \mu_1 - \mu_2 = 0$$

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

$$\alpha = 0.05$$

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

$$= \frac{\sqrt{\frac{(27)^2}{40} + \frac{(31)^2}{40}}}{\sqrt{\frac{(27)^2}{40} + \frac{(31)^2}{40}}}$$

$$= 3.73$$

$$= \frac{(647 - 638) - 0}{\sqrt{\frac{(27)^2}{40} + \frac{(31)^2}{40}}}$$

1.645

Reject H_0

i.e. Accept alternative hypothesis

Critical region

for some α 's

(Z-distribution)

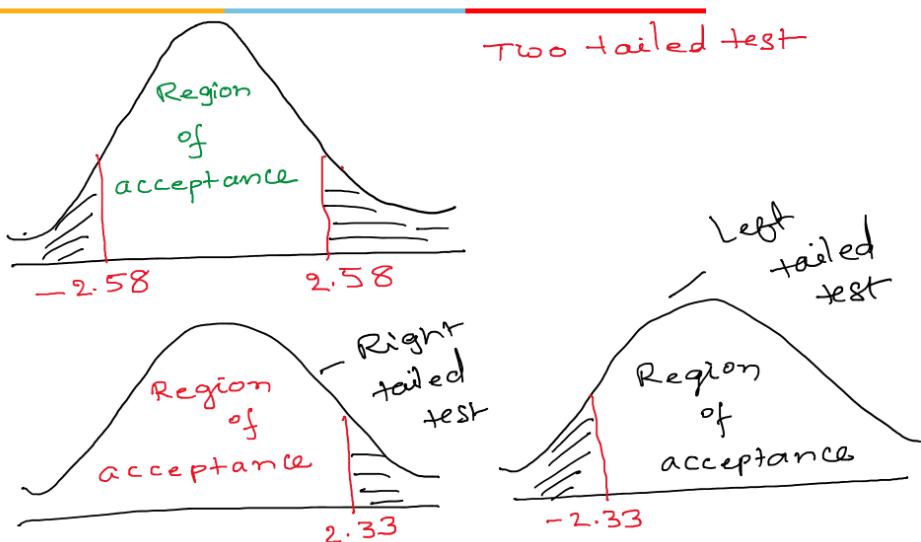
useful table values (Z-distribution)

innovate achieve lead

	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

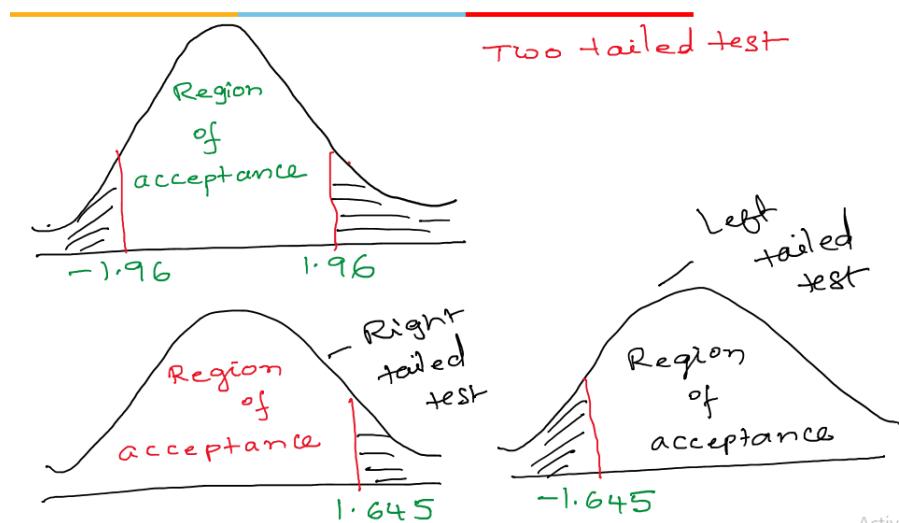
$$\alpha = 1\% \text{ (or } 0.01\text{)}$$

innovate achieve lead

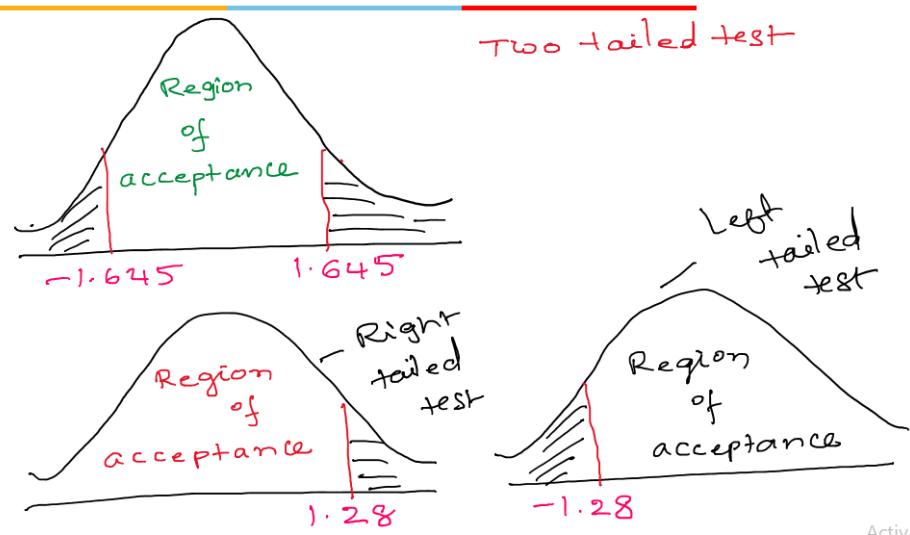


$$\alpha = 5\% \text{ (or } 0.05\text{)}$$

innovate achieve lead



$$\alpha = 10\% \text{ (or } 0.1\text{)}$$



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

Testing of Hypothesis

Example - 5

Example:-

Consider the following data



Travel time	Stress			Total
	High	Moderate	Low	
< 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 hypothesis
is based on proportion

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

$$\frac{x}{n} = \bar{P} \rightsquigarrow p_0$$



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)

$$H_0: p_1 - p_2 = \delta / p_1 - p_2 \geq \delta / p_1 - p_2 \leq \delta$$

$$Z = \frac{(\bar{p}_1 - \bar{p}_2) - \delta}{\sqrt{\bar{p}(1-\bar{p})} \left(\frac{1}{m_1} + \frac{1}{m_2} \right)}$$
 where $\bar{p} = \frac{m_1 p_1 + m_2 p_2}{m_1 + m_2}$

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

Example-4:-



A Company believes that 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 } $\text{Advt (A)} > \text{Advt (B)}$???

Sample A : 18 out of 60

Sample B : 22 out of 100

$$\bar{P}_1 = \frac{18}{60} = 0.3, \quad \bar{P}_2 = \frac{22}{100} = 0.22 \quad \text{Let } \alpha = 5\%$$

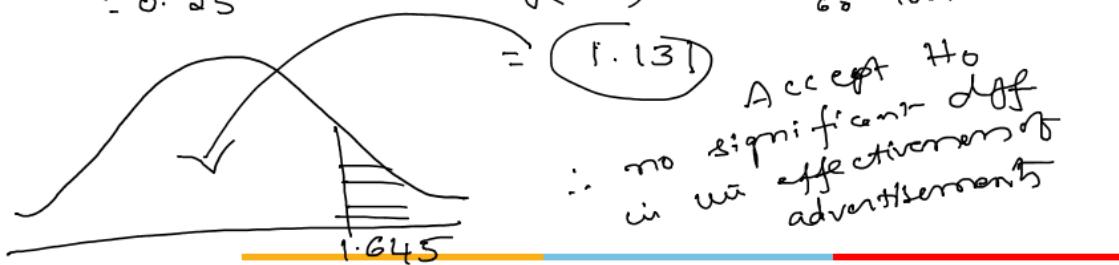
$$H_0: P_1 - P_2 = 0$$

$$H_1: P_1 > P_2$$

$$z = \frac{(\bar{P}_1 - \bar{P}_2) - \delta}{\sqrt{\bar{P}(1-\bar{P})} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

$$\begin{aligned} \bar{P} &= \frac{n_1 \bar{P}_1 + n_2 \bar{P}_2}{n_1 + n_2} \\ &= \frac{(0.3 - 0.22) - 0}{\sqrt{(0.25)(0.75)} \left(\frac{1}{60} + \frac{1}{100} \right)} \end{aligned}$$

$$= 1.13 \quad \text{Accept } H_0 \\ \therefore \text{no significant diff in effectiveness of advertisements}$$



Example:

consider the following data

Travel time	Stress			Total
	High	Moderate	Low	
< 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 $(r-1) + (c-1)$ degrees of freedom



$$\chi^2 = \sum \frac{(O-E)^2}{E} = \sum \frac{(9-12.14)^2}{12.14} \dots = 9.836$$



		Stress			Total
		High	Moderate	Low	
Travel time	< 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 $\alpha = 0.01$

$$\text{d.o.f.} : (r-1) \times (c-1) \\ = (3-1) \times (3-1) = 4$$

$$\chi^2_{0.01, 4} = 13.30$$

$\chi^2_{\text{cal}} = 9.836 < 13.30$

H_0 accepted

Example

innovate achieve lead

A tobacco company claims that there is no relationship between smoking and lung ailments.

	Lung ailment	Non-lung ailment	Total	H_0
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

$\frac{120}{300} \times 100 = 40$

$\frac{120}{300} \times 200 = 80$

$$\chi^2 = \frac{(75-60)^2}{60} + \frac{(105-120)^2}{120} + \dots$$

$$= 14.063$$

From χ^2 -tables

at 0.05 LOS

$$d.f = (r-1) \times (c-1)$$

$$(2-1) \times (2-1)$$

$$= 1$$

	Lung ailment	Non-lung ailment	Total	
Smokers	75/60	105/120	180	= 3.841
Non-smokers	25/40	95/80	120	
	100	200	300	

$$\chi^2 = 14.063 > 3.841$$

Reject $H_0 \rightarrow$

Correlation

Correlation

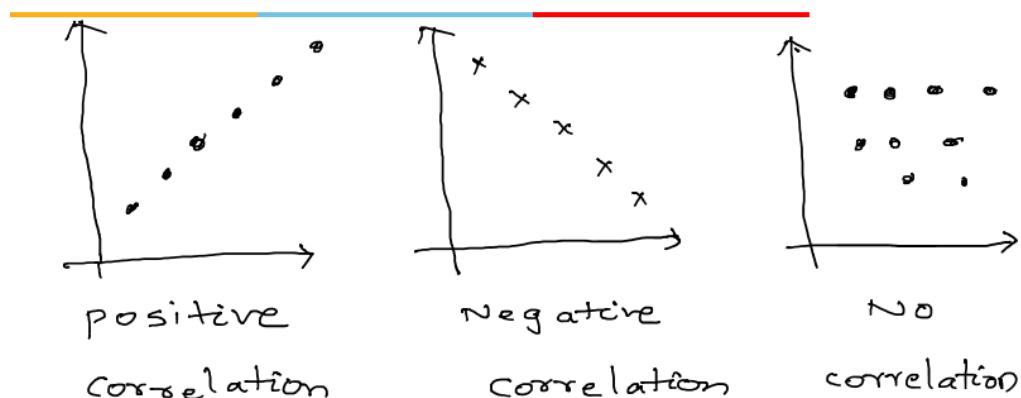


- Sales of a company and Expenditure on advertisement
- Price and Demand of a product
- Inflation and Gold price
- IQ and performance in Entrance.

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

innovate achieve

$r = 1 \Rightarrow$ 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 "

$$r = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} = \frac{\sum dx dy}{\sqrt{\sum dx^2 \sum dy^2}}$$

$$\text{where } dx = x - \bar{x}$$

$$dy = y - \bar{y}$$

$$dx^2 = (x - \bar{x})^2$$

$$dy^2 = (y - \bar{y})^2$$

Example - 1

innovate achieve lead

x	1	2	3	4	5	6	7	8	9
y	10	11	12	14	13	15	16	17	18

$$\bar{x} = \frac{\sum x}{n} = \frac{45}{9} = 5$$

$$\bar{y} = \frac{\sum y}{n} = \frac{126}{9} = 14$$

Whether they are correlated or not \rightarrow how correlated & what is one value – another given

Correlation – two variables – are related or not, relation - Regression – prediction of value

x	d_x	d_x^2	y	d_y	d_y^2	$d_x d_y$
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

Advanced Statistical Techniques for Analytics
Data Science

Slide 56

$r = \frac{\sum d_x d_y}{\sqrt{\sum d_x^2 \sum d_y^2}}$
 $= \frac{59}{\sqrt{60 \times 60}}$
 $= 0.9833$

Coefficient of Determination

innovate achieve lead

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 :-

innovate achieve lead

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

innovate achieve

- Covariance
- Correlation
- Introduction to regression
- Method of least squares
- Simple linear regression

Covariance of X and Y

innovate achieve lead

$$\text{cov}(x, y) =$$

$$= \left[E(x - \mu_x)(y - \mu_y) \right]$$

$$= \sum \sum (x - \mu_x)(y - \mu_y) P(x,y)$$

if discrete

$$= \int \int (x - \mu_x)(y - \mu_y) f(x,y) dx dy$$

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.

Correlation

→ If it is related, how it is related

⇒ Forecasting the sales, given the budget for advertising

Regression

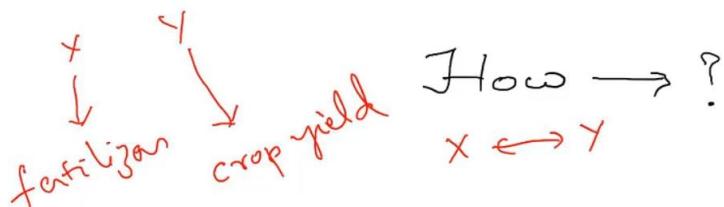
And also

innovate achieve

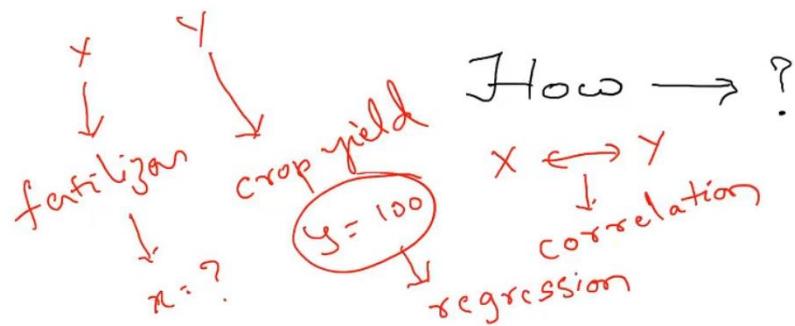
⇒ Farmer has an impression that if he uses more fertilizers, then the crop yield increases.

We need to validate this?

How → ?



Correlation – crop yield & fertilizer → relation between them



Predicting based on related → Future prediction – Regression

Correlation

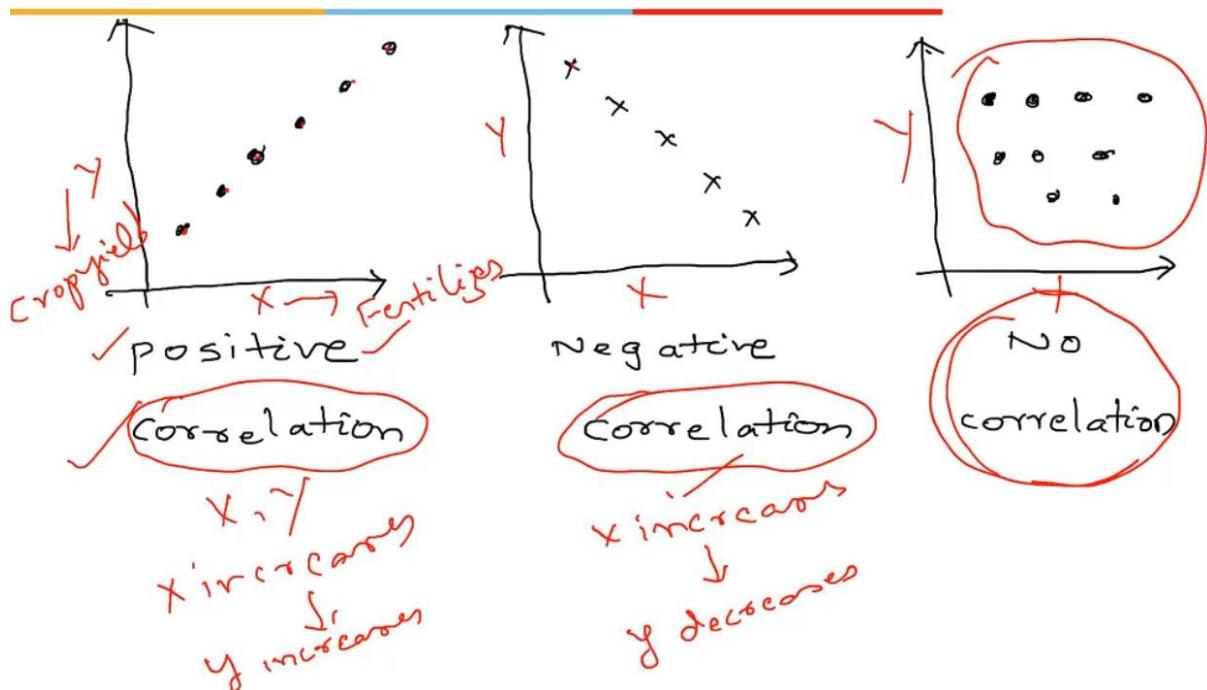
innovate achieve lead

→ Sales of a company and Expenditure on advertisement

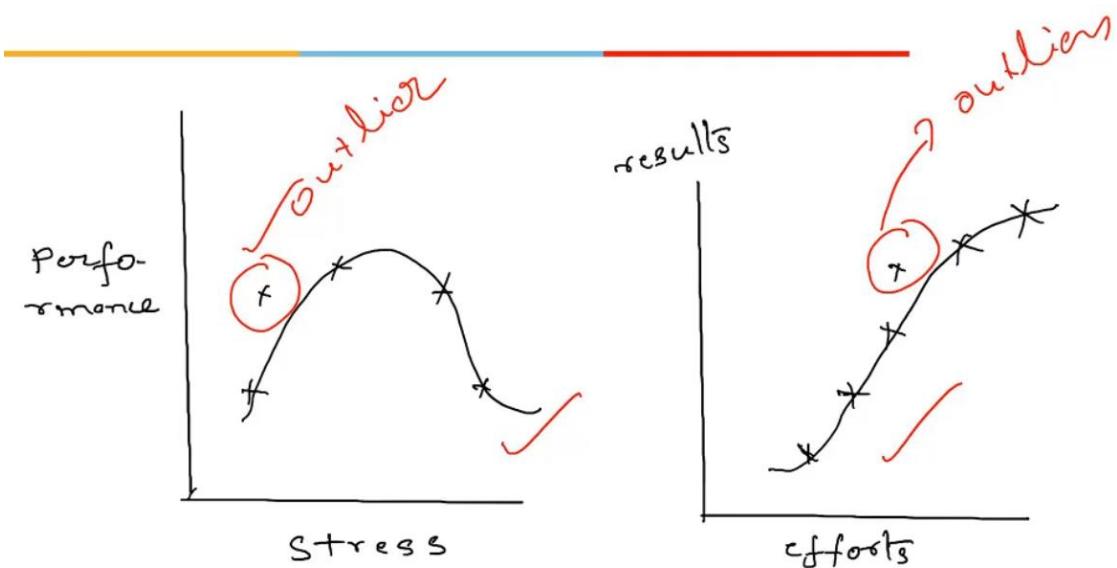
→ Price and Demand of a product

→ 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: ✓



$$r = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} = \frac{\sum x \cdot y}{\sqrt{\sum x^2 \cdot \sum y^2}}$$

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

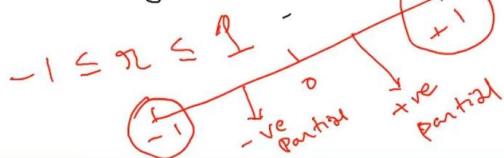
$r = 1 \Rightarrow$ perfect and positive relation

$r = -1 \Rightarrow$ " " negative relation

$r = 0 \Rightarrow$ no relation

$0 < r < 1 \Rightarrow$ partial positive relation

$-1 < r < 0 \Rightarrow$ " " negative "



Example - 1

innovate achieve lead

x	1	2	3	4	5	6	7	8	9
y	10	11	12	14	13	15	16	12	18

$$\bar{x} = \frac{\sum x}{n} = \frac{45}{9} = 5$$

$$\bar{y} = \frac{\sum y}{n} = \frac{126}{9} = 14$$

x	$x - \bar{x}$	$(x - \bar{x})^2$	y	$y - \bar{y}$	$(y - \bar{y})^2$	xy	$\sum xy$	$\sum (x - \bar{x})^2$	$\sum (y - \bar{y})^2$
1	-4	16	10	-4	16	16	59	60	60
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			

$r = \frac{\sum xy}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$
 $= \frac{59}{\sqrt{60 \times 60}} = 0.9833$

Coefficient of determination - r^2 - At which extent - x expands y

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.

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

Regression :-



x	1	2	3	4	5
y	1	4	9	16	25

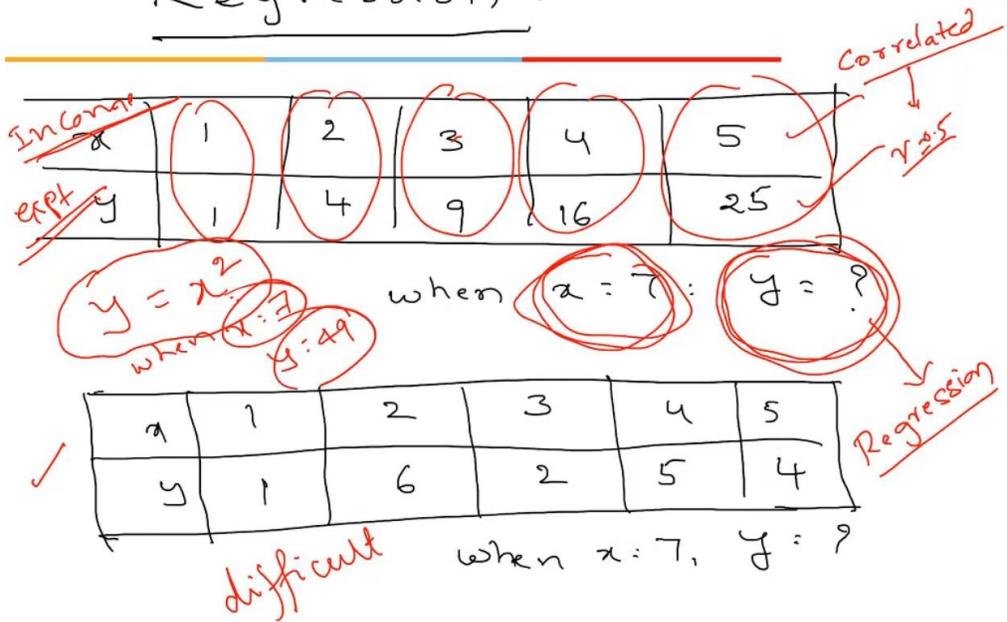
when $x = 7$: $y = ?$

x	1	2	3	4	5
y	1	6	2	5	4

when $x = 7$, $y = ?$

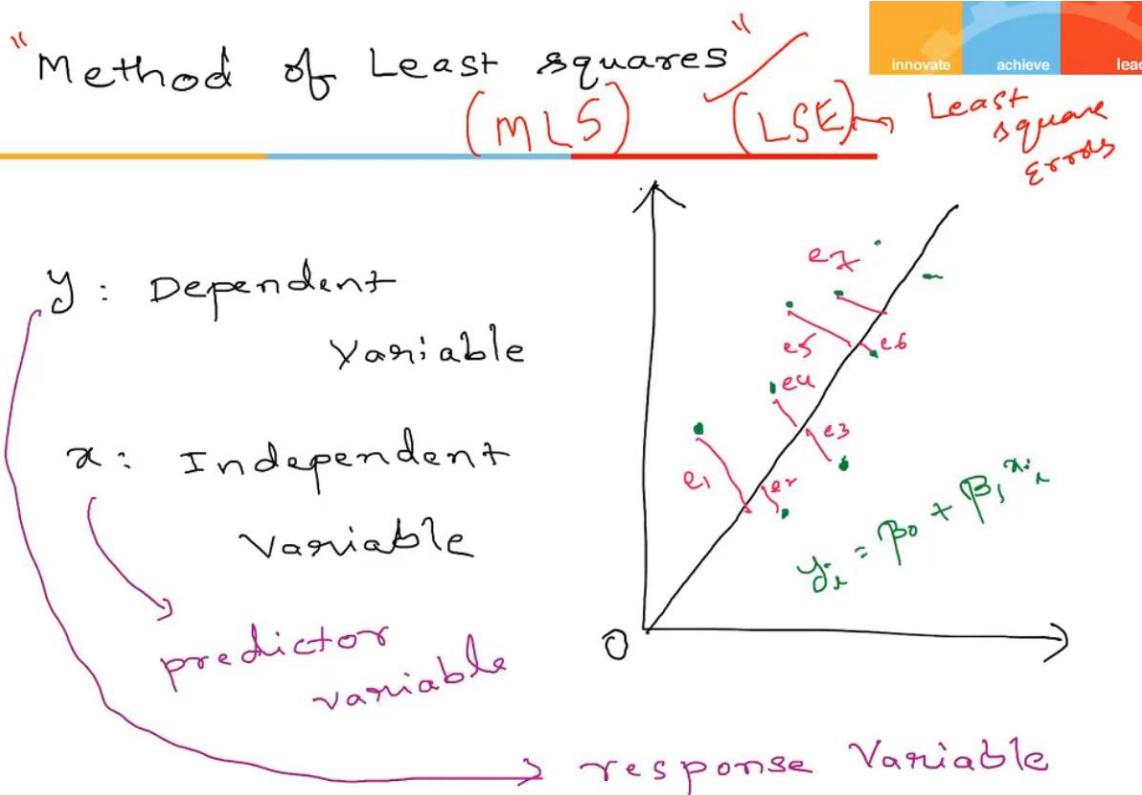
Regression :-

innovate achieve lead

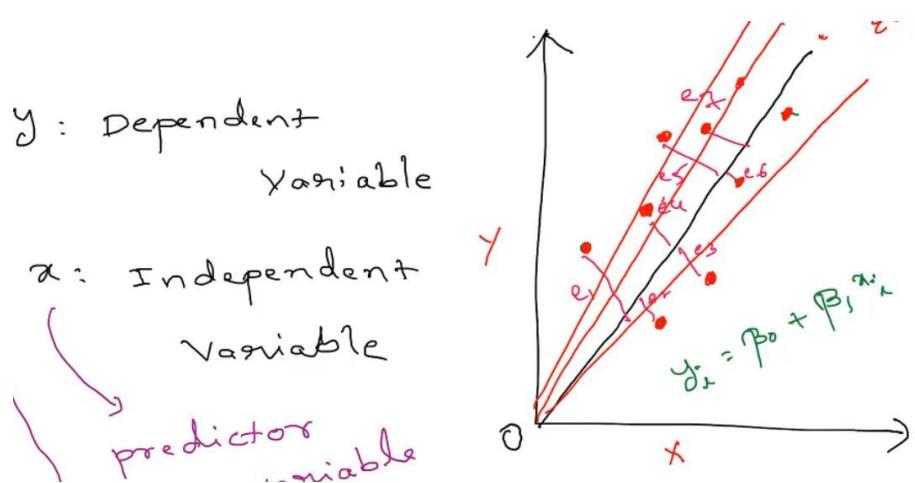


First find relation – having relation or not – Correlation \rightarrow Forecasting \rightarrow regression – exact relation is required to predict

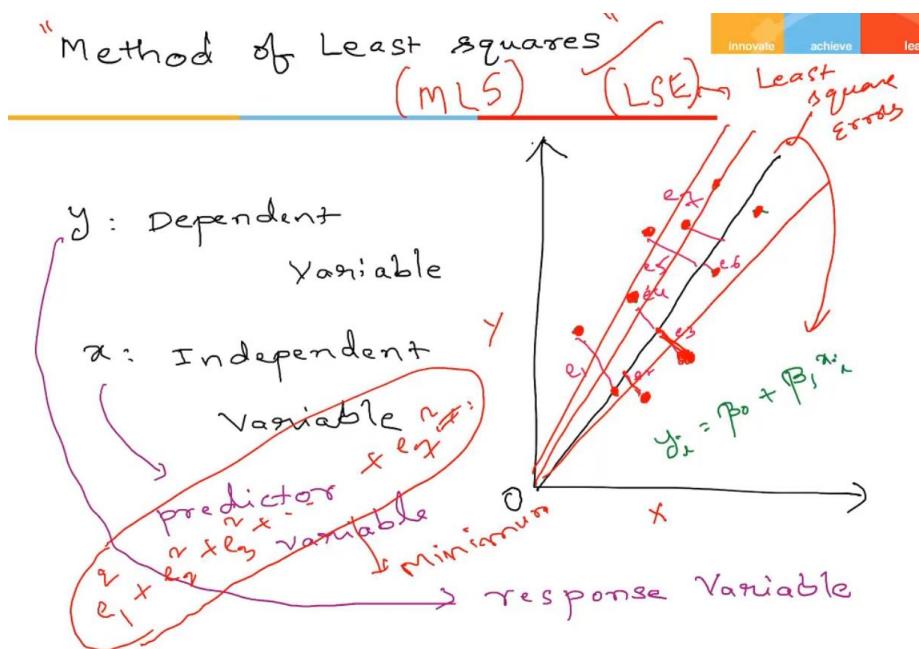
To establish a relation – what exact relation – Regression, To find out what relation and whether they are related or not \rightarrow Correlation



Try to draw straight line – so most points are covered



Which line is optimal? – Error approximation is minimal – LSE



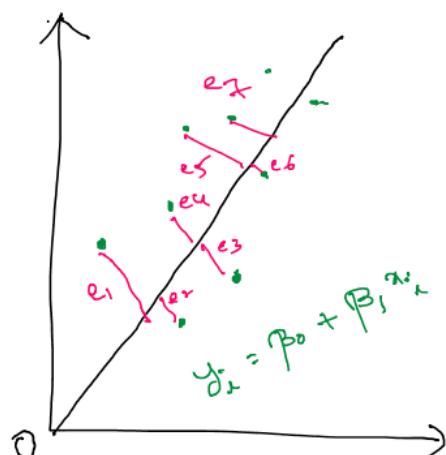
Method of Least squares

innovate achieve lead

$$S(\beta_0, \beta_1)$$

$$= \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

β_0 and β_1 are chosen to minimize the sum of squared errors.



Method of Least Squares

innovate achieve lead

$$S(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

β_0 and β_1 are chosen to minimize the sum of squared errors.

$$y_i = (\beta_0 + \beta_1 x_i)$$

Method of Least Squares

innovate achieve lead

$$S(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{\partial S}{\partial \beta_0} = 0 \Rightarrow \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) (-1)$$

$$\Rightarrow \sum_{i=1}^n y_i = n \beta_0 + \beta_1 \sum_{i=1}^n x_i$$

$$\frac{\partial S}{\partial \beta_1} = 0 \Rightarrow \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) (2)(-x_i)$$

$$\Rightarrow \sum_{i=1}^n x_i y_i = \beta_0 \sum x_i + \beta_1 \sum x_i^2$$

on solving these, we get β_0 & β_1 which minimizes error.

Method of Least Squares

innovate achieve lead

$$S(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{\partial S}{\partial \beta_0} = 0 \Rightarrow \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) (-1) = 0$$

$$\Rightarrow \sum_{i=1}^n y_i = n \beta_0 + \beta_1 \sum_{i=1}^n x_i$$

$$\frac{\partial S}{\partial \beta_1} = 0 \Rightarrow \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) (2)(-x_i)$$

$$\Rightarrow \sum_{i=1}^n x_i y_i = \beta_0 \sum x_i + \beta_1 \sum x_i^2$$

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.

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.

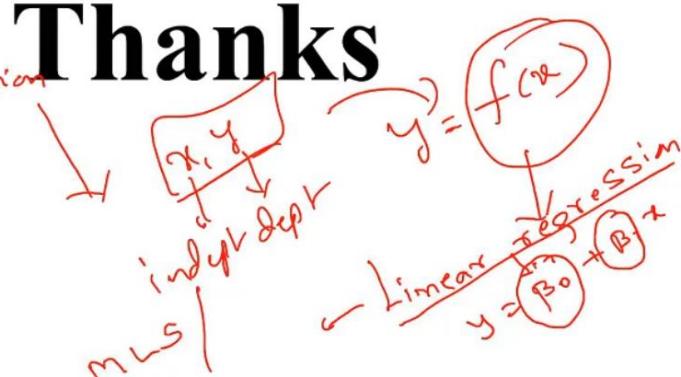
$$\begin{aligned} \beta_0, \beta_1 &= \frac{\sum (\bar{x} - \bar{y})}{\sum (\bar{x} - \bar{y})^2} \\ 2 &= \frac{\sum (\bar{x} - \bar{y})}{\sum (\bar{x} - \bar{y})^2} \\ 6 &= \frac{\sum (\bar{x} - \bar{y})}{\sum (\bar{x} - \bar{y})^2} \end{aligned}$$

Find $\beta_0, \beta_1 \rightarrow$ so most of the points are nearer \rightarrow optimal case

1) Covariance $\rightarrow \text{cov}(xy)$

2) Correlation $\rightarrow \rho_{x,y}$

3) Regression



Matrix Approach:



$$\text{Let } y = \beta_0 + \beta_1 x_1 + \dots + \beta_{p-1} x_{p-1}$$

Observations $y_i = 1, 2, \dots, n \rightarrow$ by a vector γ

Unknowns $\beta_0, \beta_1, \dots, \beta_{p-1} \rightarrow \dots \beta$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \dots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{n,p-1} \end{bmatrix}$$

$$\therefore \hat{\gamma} = X \beta \quad n \times p \quad p \times 1$$

Find β to minimize

$$S(\beta) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \dots)^2$$

$$= \|\gamma - X\beta\|^2 = \|\gamma - \hat{\gamma}\|^2$$

Diff S wrt to each β we get linear eqns

$$X^T X \hat{\beta} = X^T \gamma \rightarrow \text{normal eqns}$$

If $X^T X$ is non-singular, the soln is

$$\hat{\beta} = (X^T X)^{-1} X^T \gamma$$

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:-

	size	No of rooms	No of floors	Age of home	price Lakh
1	2000	5	2	45	4000
1	1400	3	1	40	2000
1	1600	3	2	30	3000
1	800	2	1	35	2000

$x_1 \quad x_2 \quad x_3 \quad x_4 \quad y$

Linear regression (multiple regression)



example:-

$$X = \begin{matrix} & \overbrace{\quad\quad\quad}^T \\ \left[\begin{array}{ccccc} , & 2000 & 5 & 2 & 45 \\ , & 1400 & 3 & 1 & 40 \\ | & 1600 & 3 & 2 & 30 \\ | & 800 & 2 & 1 & 35 \end{array} \right] & \left[\begin{array}{c} 4000 \\ 2000 \\ 3000 \\ 2000 \end{array} \right] \end{matrix}$$

$$B = (X^T X)^{-1} X^T y$$

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	$y = \beta_0 + \beta_1 x$
1	0.5	0.5	1	$\Sigma y = n\beta_0 + \beta_1 \Sigma x$
2	1	2	4	$\Sigma xy = \beta_0 \Sigma x_1 + \beta_1 \Sigma x^2$
4	2	8	16	$3.5 = 4\beta_0 + \beta_1 \quad (1)$
0	0	0	0	$10.5 = -\beta_0 + \beta_1 \quad (2)$
$\Sigma x = 7$		$\Sigma y = 3.5$	$\Sigma xy = 10.5$	$\Sigma x^2 = 21$
				<u>on solving these</u>
				$\beta_0 = 0$
				$\beta_1 = 0.5$
				i.e. $y = 0 + (0.5)x$
				$\boxed{\text{when } x=5, \quad y = (0.5)5 \\ = 0.25}$

Lec 9 Predictive Analytics & Revision

L- 9: Predictive Analytics & Revision

Agenda

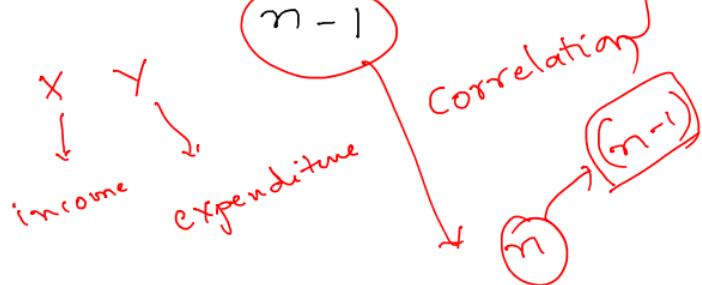
- Review of last session
- Introduction to regression
- Method of least squares
- Simple linear regression

Covariance of x and y

$$\begin{aligned} \text{cov}(x, y) &= E(x) - \sum x P(x) \\ &= \left[E(x - \mu_x)(y - \mu_y) \right] \quad \text{Joint p.d.f.} \\ &= \sum \sum (x - \mu_x)(y - \mu_y) P(x, y) \quad \text{if discrete} \\ &= \iint (x - \mu_x)(y - \mu_y) f(x, y) dxdy \quad \text{if continuous} \end{aligned}$$

$$\text{cov}(x, y)$$

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



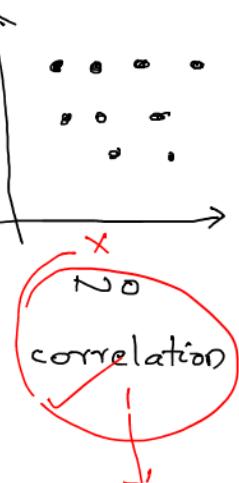
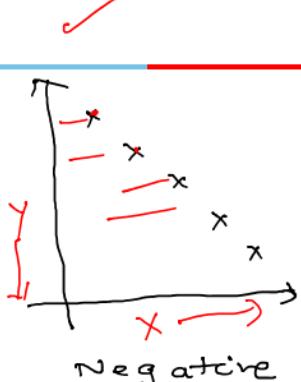
And also

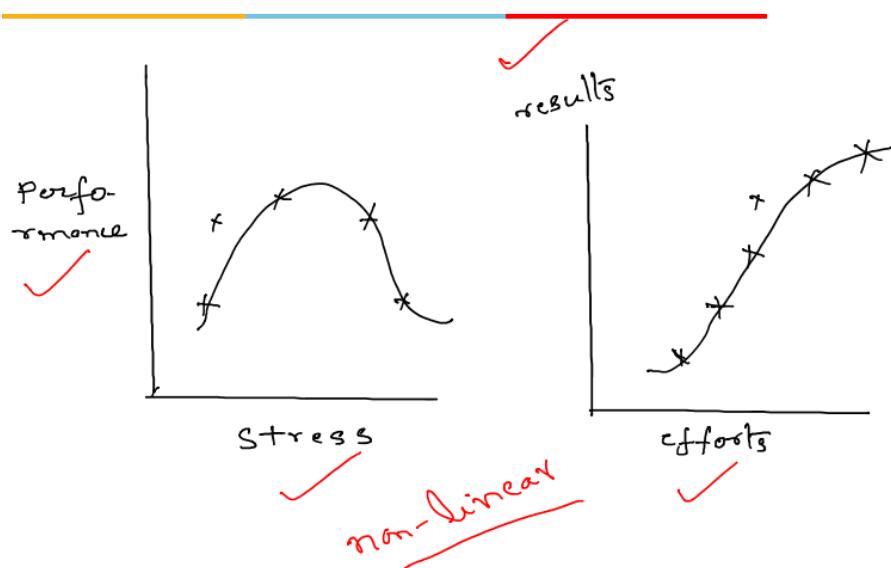


⇒ Farmer has an impression that if he uses more fertilizers, then crop yield increases.

We need to validate this?

How → ?





Coefficient of correlation:

$$r = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} = \frac{\sum xy}{\sqrt{\sum x^2 \cdot \sum y^2}}$$

$$\text{where } x = x - \bar{x}$$

$$y = y - \bar{y}$$

$$x^2 = (x - \bar{x})^2$$

$$y^2 = (y - \bar{y})^2$$

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

innovate achieve lead

x	1	2	3	4	5	6	7	8	9
y	10	11	12	14	13	15	16	12	18

$$\bar{x} = \frac{\sum x}{n} = \frac{45}{9} = 5$$

$$\bar{y} = \frac{\sum y}{n} = \frac{126}{9} = 14$$

x	$x - \bar{x}$	$(x - \bar{x})^2$	y	$y - \bar{y}$	$(y - \bar{y})^2$	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

$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$
 $= \frac{59}{\sqrt{60 \times 60}}$
 $= 0.9833$

$r = 0.9833$
 → Corrected
 → Corrected

x	$x =$	y	$y = y - 14$	y^2	xy
1	-4	10	-4	16	16
2	-3	11	-3	9	9
3	-2	12	-2	4	4
4	-1	14	0	0	0
5	0	13	-1	0	0
6	1	15	1	1	1
7	2	16	2	4	4
8	3	17	3	9	9
9	4	18	4	16	16

60 60 59

cov(x,y) = $\frac{\sum xy}{n-1}$ = $\frac{59}{8} = 7.375$

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.

~~$r \neq r^2$~~

$r = 0.9833 \rightarrow$ Coeff correlation

$\text{cov}(x, y) = 7.375 \rightarrow$ Covariance Interpretation

$r^2 = 0.81 \rightarrow$ Coeff of determination

$-1 \leq r \leq 1$

94.90233116

farmer:

x : fertilizer
 y : crop yield

Correlation r

Regression

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
X	100	105	110	115	120	125	130	135	140	145	150	155
Y	100	105	110	115	120	125	130	135	140	145	150	155

Regression :-

X	1	2	3	4	5
Y	1	4	9	16	25

$y = x^2$ when $x = ?$
 $y = ?$ when $x = 7$: $y = ?$

X	1	2	3	4	5
Y	1	6	2	5	4

$y = ?$ when $x = 7$, $y = ?$

Correlation

- Measuring strength or degree of the relationship between two variables
- no estimation
- both variables are independent

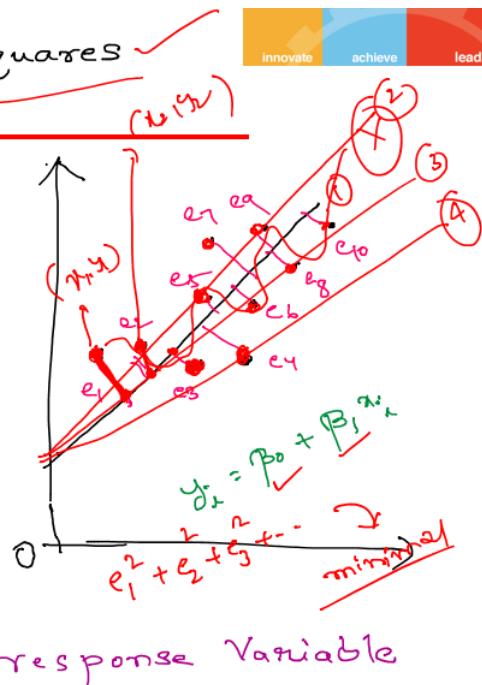
Regression

- having an algebraic equation between two variables
- estimation
- one is dep't variable and other indept variables

Method of Least squares



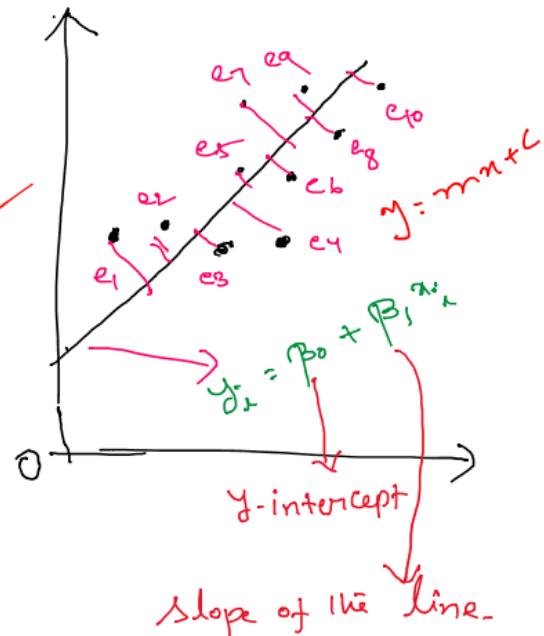
y : Dependent variable
 x : Independent variable
 predictor variable
 response Variable



"minimizing the error"

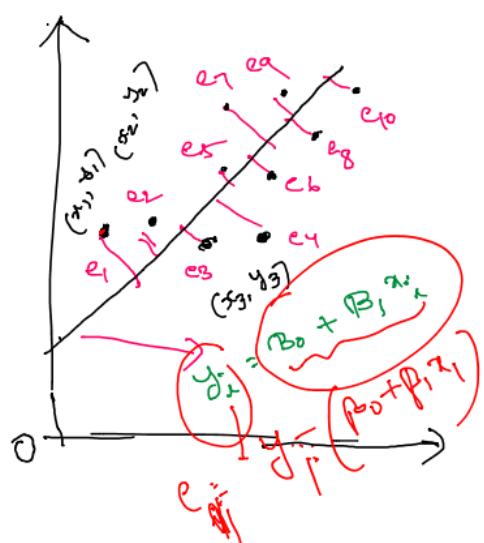
i minimize

$$e_1^2 + e_2^2 + e_3^2 + \dots + e_n^2$$



$$S(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

we need to choose β_0 and β_1 which minimizes the error.



Method of Least squares

Mat
innovate
lead

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.

Linear regression

$$y = \beta_0 + \beta_1 x$$

$$y = \beta_0 + \beta_1 x$$

$$\sum x_i y_i = \beta_0 (\sum x_i) + \beta_1 (\sum x_i^2)$$

Normal equations.

$$y = \frac{P_0 + P_1 t}{P_0}$$

Regression Coefficients



$$y = a + b_x x \quad \boxed{\text{regression line of } y \text{ on } x}$$

b_{yx} : Regression coeff of y on x

$$x = c + b_y y \quad \boxed{\text{regression line of } x \text{ on } y}$$

b_{xy} : regression coeff of ~~y on x~~
 ~~x on y~~

Correlation coefficient

$$r = \sqrt{b_{yx} \times b_{xy}}$$

$r = +0.9$ (positive geometric mean)

$r = -0.9$ (negative)

$r^2 = 0.81$ (Determination)

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	

$$y = a + b x$$

$$\sum y = a n + b \sum x$$

$$\sum xy = a \sum x + b \sum x^2$$

Example :-

Sales	Advt.	x^2	xy	$\Sigma y = m\beta_0 + \beta_1 \Sigma x$
1	1	1	1	$\Sigma y = m\beta_0 + \beta_1 \Sigma x$
2	3	9	6	$\Sigma xy = \beta_0 \Sigma n + \beta_1 \Sigma x^2$
2	4	16	8	$\Rightarrow 40 = 8\beta_0 + 56\beta_1$
4	6	36	24	$373 = 56\beta_0 + 524\beta_1$
6	8	64	48	on solving
8	9	81	72	$\beta_0 = 0.072$
8	11	121	88	$\beta_1 = 0.704$
9	14	196	126	$\therefore y = (0.072) + (0.704)x$
$\Sigma 40$	$\Sigma 56$	$\Sigma 524$	$\Sigma 373$	

$$\therefore y = (0.072) + (0.704)x$$

when $x = 0.075$, then

$$y = (0.072) + (0.704)(0.075)$$

$$= 0.1248 \approx 12.48\%$$

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	$y = \beta_0 + \beta_1 x$
1	0.5	0.5	1	$\sum y = n\beta_0 + \beta_1 \sum x$
2	1	2	4	$\sum xy = \beta_0 \sum x_1 + \beta_1 \sum x^2$
4	2	8	16	$3.5 = 4\beta_0 + \beta_1 (7)$
0	0	0	0	$10.5 = 7\beta_0 + \beta_1 (21)$
$\sum x = 7$		$\sum y = 3.5$	$\sum x^2 = 10.5$	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)

innovate achieve lead

Example:-

x_0	size	No of rooms	No of floors	Age of home	price Lakh
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

Price
y
increment my pur
20 years
1 floor
2 rooms
1200 sqft

$y =$

Multiple Linear Regression



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

$$\sum y = \beta_0 n + \beta_1 \sum x_1 + \beta_2 \sum x_2 + \beta_3 \sum x_3 + \beta_4 \sum x_4$$

$$\sum xy = \beta_0 \sum x_1 + \beta_1 \sum x_1^2 + \beta_2 \sum x_1 x_2 + \beta_3 \sum x_1 x_3 + \beta_4 \sum x_1 x_4$$

$$\sum x_2 y = \beta_0 \sum x_2 + \beta_1 \sum x_1 x_2 + \beta_2 \sum x_2^2 + \beta_3 \sum x_2 x_3 + \beta_4 \sum x_2 x_4$$

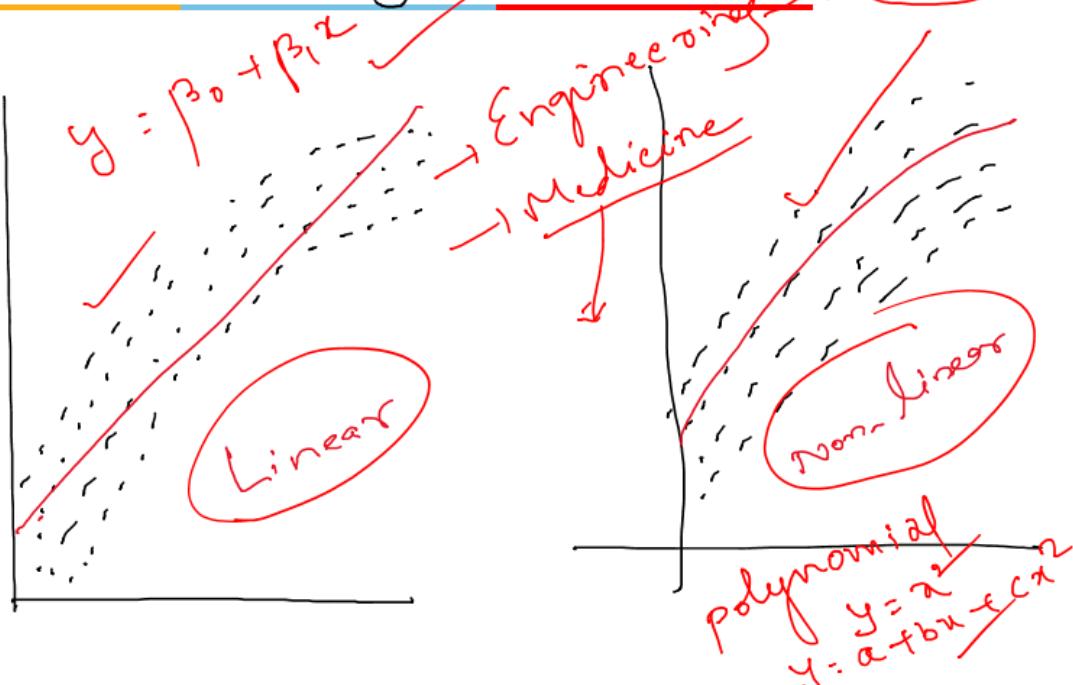
$$\sum x_3 y = \beta_0 \sum x_3 + \beta_1 \sum x_1 x_3 + \beta_2 \sum x_2 x_3 + \beta_3 \sum x_3^2 + \beta_4 \sum x_3 x_4$$

$$\sum x_4 y = \beta_0 \sum x_4 + \beta_1 \sum x_1 x_4 + \beta_2 \sum x_2 x_4 + \beta_3 \sum x_3 x_4 + \beta_4 \sum x_4^2$$

separately $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$

Other regressions

just a look



Suppose $y = a e^{bx}$ exponential curve

$$\log y = \log a + b \log x$$

i.e. $y = A + bX$ linear eqn type

$$\sum y = A \sum x + b \sum x^2 \rightarrow 1$$

$$\sum xy = A \sum x^2 + b \sum x^3 \rightarrow 2$$

$A = ?$ $b = ?$

Hence, we get $y = ae^{bx}$

Suppose $y = ax^b$ non linear Power Curve

$$\log y = \log a + b \log x$$

i.e. $y = A + bX$

$$\sum y = A \sum x + b \sum x^2$$

$$\sum xy = A \sum x^2 + b \sum x^3$$

$A = ?$ $b = ?$

$y = ax^b$

Matrix Approach:

$$\text{Let } y = \beta_0 + \beta_1 x_1 + \dots + \beta_{p-1} x_{p-1}$$

Observations $y_i = 1, 2, \dots, n \rightarrow$ by a vector γ

Unknowns $\beta_0, \beta_1, \dots, \beta_{p-1} \rightarrow \dots \beta$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1(p-1)} \\ 1 & x_{21} & x_{22} & \dots & x_{2(p-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{n(p-1)} \end{bmatrix}$$

$$\hat{Y} = X\beta$$

Diagram illustrating the matrix equation $\hat{Y} = X\beta$. A red bracket on the left indicates the columns of X are $n \times p$, and a red bracket on the right indicates the columns of β are $p \times 1$. The resulting vector \hat{Y} is $n \times 1$. Below the diagram, a red circle contains the equation $\hat{Y} = BX$, where B is a red bracketed matrix. To the right, another red circle contains $\hat{Y} = Bx$, where x is a red bracketed vector. A red bracket on the far right indicates the columns of B are $n \times p$.

Find β to minimize

$$S(\beta) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \dots)^2$$

$$= \|y - X\beta\|^2 = \|y - \hat{Y}\|^2$$

Diff S wrt to each β we get linear eqns

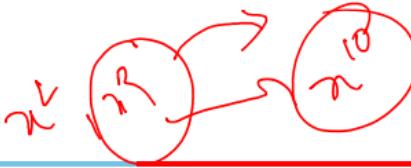
$$X^T \hat{Y} = X^T y \rightarrow \text{normal eqns}$$

If $X^T X$ is non singular, the soln is

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

$$X^T X = B^T B$$

$$X = B^{-1} B$$



Computationally, it is sometimes unwise even to form the normal equations because the multiplications involved in forming $\mathbf{x}^T \mathbf{x}$ can introduce undesirable round-off error.

→ If $\mathbf{x}^T \mathbf{x}$ is non-invertible ... ?

✓ Redundant features

✓ Too many features

Scaling

Revision

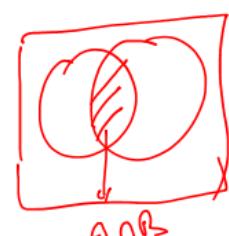
→ probability → $P(A \cup B)$
 $P(A \cap B)$



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

→ Conditional probability:

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



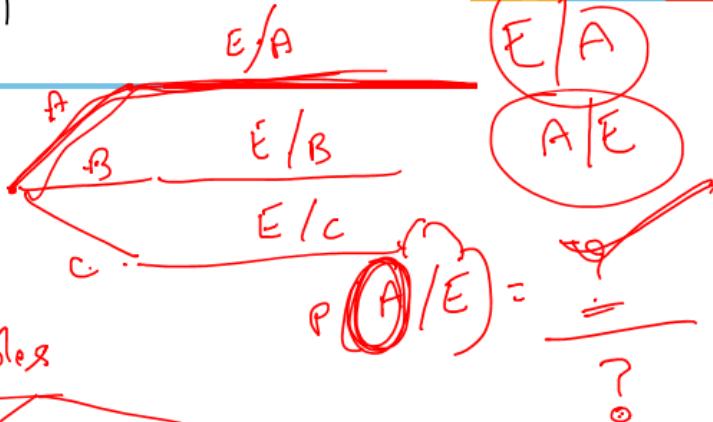
$$P(B | A) = \frac{P(A \cap B)}{P(A)}$$

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

Revision

innovate achieve lead

→ Bayes' Theorem :-



→ Random variables

Discrete

$$\begin{aligned} & P(x) \\ \text{i, } & 0 \leq P(x) \leq 1 \\ (\text{ii, }) & \sum P(x) = 1 \\ \text{Mean} = & E(x) = \sum x P(x) \end{aligned}$$

continuous

$$\begin{aligned} & f(x) \\ \text{i, } & 0 \leq f(x) \leq 1 \\ (\text{ii, }) & \int f(x) dx = 1 \\ & \text{Mean: } E(x) = \int x f(x) dx \\ & \text{variance: } E(x^2) - \mu^2 \\ & : E(x^2) - [E(x)]^2 \end{aligned}$$

→ Bimomial dist $P(x) = \frac{n!}{x!(n-x)!} p^x q^{n-x}$, $x=0, 1, 2, \dots, n$
 poisson dist $P(x) = \frac{e^{-\lambda} \lambda^x}{x!}$, $x=0, 1, 2, \dots, \infty$

→ Normal distribution :- $P(30 \leq x \leq 50)$

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

$P(\zeta_1 \leq \zeta \leq \zeta_2) = F(\zeta_2) - F(\zeta_1)$

Mean → one ζ
 two ζ
 proportion ζ → χ^2 -distribution

→ Testing of Hypothesis

Thanks

9490233116