2018

SATURDAY
DAY (076-289)
11th Week

17

Appointment	Notes	Work to do

DATA ANALYTICS

18 SUNDAY

Data Analyst

APR MAY JUN

S M T W T F S S M T W T F S S M T W T F S S M T W T F S S M T

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 •

APR '18

Appointment | Notes | Work to do

-: Statistics :-

Beyond the Data : Navigating the World of Statistics.

- Introduction to Statistics.
- Descriptive Statistics
- Probability.
- Inferential Statistics

Appointment | Notes | Work to do

>> Introduction of Statistics :-

- What is Statistics ?
- Types of statistics - Descriptive and Inferential
- Types of Data
- Population and Sample
- Sampling Techniques
- Statistical Data Analysis Steps.

Descriptive Statistics :-

>> Measure of central tendency - Mean, Median and Mode

>> Measure of dispersion - Range, Variance, Standard Deviation, Percentiles and Quantiles

>> Graphical Representations - Boxplots, Histogram, Scatterplots.

>> Outliers and understanding their impact

>> Correlation and Covariance

23 FRIDAY
DAY (082-283)
12th Week

March 2018

Appointment | Notes | Work to do

Probability :-

>> Basic Probability concepts : sample space, event etc.

>> Types of Events :-

(i) Disjoint or Non-disjoint event

(ii) Independent or Dependent event.

>> Complement of Probability

>> Conditional Probability

>> Bayes Theorem

>> Probability Distributions

>> Random Variables and its Types
(Discrete & Continuous)

>> PMF and PDF

> PMF : probability mass Function

>> PDF : " density "

Mar '18
1 F 2 S 3 S 4 M 5 T 6 W 7 T 8 F 9 S 10 S 11 M 12 T 13 W 14 T 15 F 16 S 17 S 18 M 19 T 20 W 21 T 22 F 23 S 24 S 25 M 26 T 27 W 28 T 29 F 30 S 31

SATURDAY
DAY (083-282)
12th Week
24

Appointment | Notes | Work to do

Discrete Distributions

- Binomial (binomial)
- Bernoulli (bernoulli)

Continous Distributions

- Uniform
- Normal

Standard Normal Distribution

Standardization

Normalization

Empirical Rule

25 SUNDAY

S 1 M 2 T 3 W 4 T 5 F 6 S 7 S 8 M 9 T 10 W 11 T 12 F 13 S 14 S 15 M 16 T 17 W 18 T 19 F 20 S 21 S 22 M 23 T 24 W 25 T 26 F 27 S 28 S 29 M 30 T Apr '18

APR MAY JUN

Appointment | Notes | Work to do

Inferential Statistics :

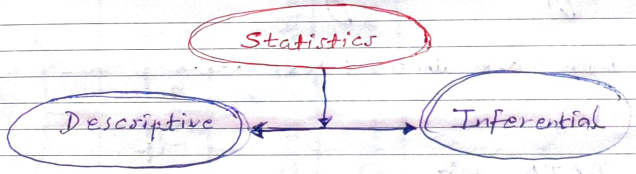
- Relationship with Descriptive Statistics
- Point and Interval Estimation
- Confidence Interval (Z | T distribution)
- Hypothesis testing :-
  - >> Types of Hypothesis : Null and Alternate
  - >> Level of significance and p value
  - >> Types of Errors
  - >> One tailed or two tailed test.
- Types of test in statistics (z test, t test, ANOVA, Chi square etc ...)

Appointment | Notes | Work to do

Statistics :-

The study and manipulation of data, including ways to gather, review, analyze, and draw conclusion from data.

- * collecting
- * analyzing
- * interpreting
- * drawing conclusion ...



- Types :-

(i) Descriptive Statistics :-

- > Collection, Analyzing and Intrelation of data.
- >> Understanding main features of data.
- >>> Organizing and Summarizing information from data.
  - graphs
  - tables
  - average
  - variations etc ...

Appointment | Notes | Work to do

(ii.) Inferential Statistics :-

> drawing conclusion from data. (representing complete population).

>> sample ...

>>> statical method. (statistical methods) ...

- Confidence Interval
- Estimation
- Hypothesis Testing ...

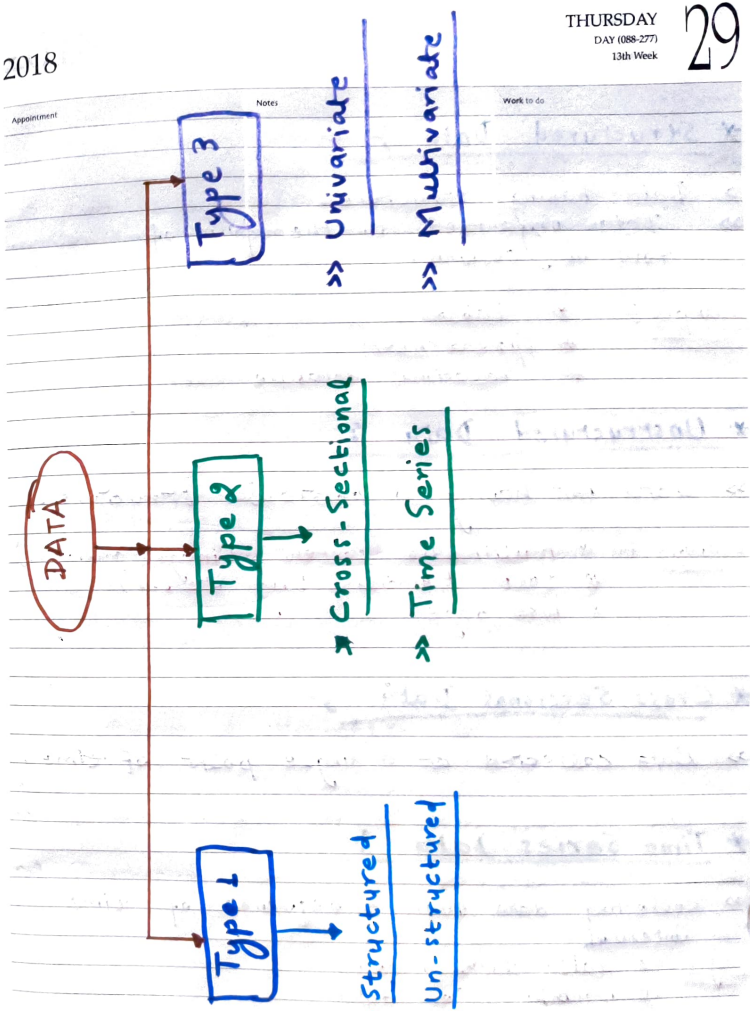
>> Both are interrelated ....

DATA :-

> collection of facts, observations or measurements used for analysis.

>> It can be ;

- Numerical
- Categorical
- Combination of both ...



Appointment | Notes | Work to do

* Structured Data :-

» Data having structure.
»» Data organized in the form of row & column.

example :- * Tables
* Spreadsheet
* relational database ...

* Unstructured Data :-

»» data not having a particular structure.

example :- * multimedia content (image, audio ...)
* text (emails, blog etc ...)
* web page

* Cross Sectional Data :-

»» Data collected at a single point of time.

* Time Series Data :-

»» collecting data over a sequence of time intervals.

eg :- * daily stocks prices
* monthly sales data

* Univariate Data :-

»» Single variable (single data collection).

* Multivariate Data :-

»» Having two or more than two variables in data set.

Variables :-

»» Variables may also called a data item.

examples :- * age
* sex
* country of birth etc ...

- types of Variables :-

(i) Nominal : gender, colours
(ii) Ordinal : education levels, customer rating
(iii) Numerical : integer, (age, price)
(iv) Categorical : cars, product categories
(v) Interval : temperature, IQ scores
(vi) Ratio : height, weight

01 SUNDAY
DAY (091-274)
13th Week

April

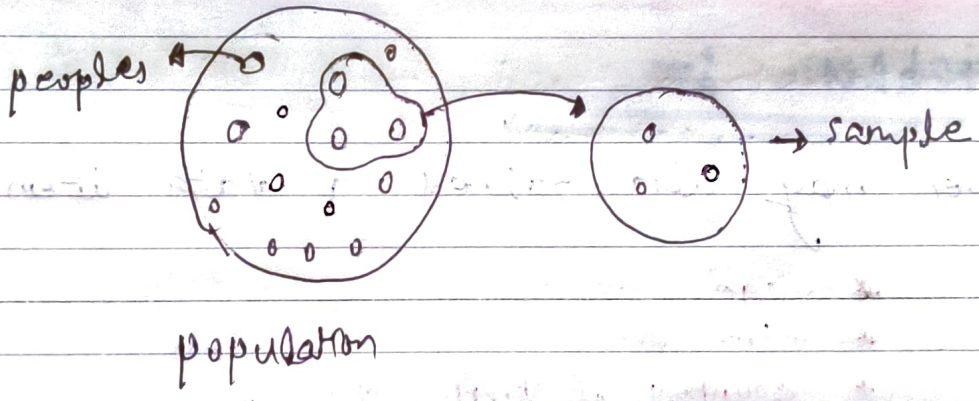
Appointment | Notes | Work to do

Population and Samples :-

>> population is the entire group of individuals.

example : people in india ;
all users on social-media ;

>> Sample is a subset of population.



* Why Sample :-

>> To reduce the cost of data collection.

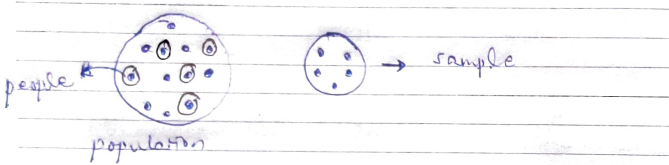
>> When a full census data cannot be taken.

Appointment | Notes | Work to do

Sampling Techniques :-

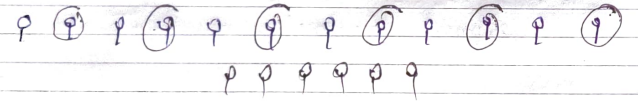
- Nature of population.
- Research objective.
- Available resource.

(i) Random :-



(ii) Stratified :- based on characteristics. (gender).

(iii) Systematic :- system follow. (Kth element) (number wise).



(iv) Clustered :-

>> population divided in cluster.

>> randomly select complete cluster.

Appointment | Notes | Work to do

Statistical Analysis :-

>> The procen of collecting large volumes of data and then using statistics and other data analysis techniques to identify :

(i) Trends
(ii) Patterns
(iii) Insights...

* Steps follow while performing :-

(i) Define the problem or research question ?
(ii) Data Collection ?
(iii) Data cleaning ...
(iv) Exploratory Data Analysis.
(v) Data Transformation ?
(vi) Hypothesis Formulation
(vii) Statistical Testing.
(viii) Interpretation of Results ...
(ix) Draw Conclusions ?
(x) Document the Analysis Procen / Report Making ...

Appointment Notes Work to do

Measure of Central Tendency :-

>> The numerical values that are used to represent mid-value or central value of a large collection of numerical data.

>> Part of Descriptive statistics.
>> Describe centre of dataset.

→ Three measures :-

(i) Mean
(ii) Median
(iii) Mode

(i) MEAN : used for the arithmetic mean of data.
→ also known as "Average" (avg).

→ $\bar{x} = \frac{\sum x_i}{N}$

mean = sum of all observations ÷ total no. of observation

e.g:- If there are 5 observation, which are 27, 11, 17, 19, and 21 then the mean ($\bar{x}$) is given by ?

⇒ $\bar{x} = (27+11+17+19+21) \div 5$
$= 95 \div 5$
$= 19$

Appointment Notes Work to do

(ii) MEDIAN :
>> middle value of the dataset when it is ordered.
>> Numerical variables.
>> Less influenced by outliers.
>> Even numbers, median is the average of two middle value.

example :- $x = \{3, 1, 7, 5, 9\}$

$x = \{1, 3, (5), 7, 9\}$
↓
median

e.g:- $x = \{1, 3, 5, 7, 9, 11\}$

$x = \frac{5+7}{2}$

$= \frac{12}{2}$

$= 6$ → median

(iii) MODE :-
>> most frequent value (repeated value).
>> for categorical variables.
>> Numerical variables, unique count be less

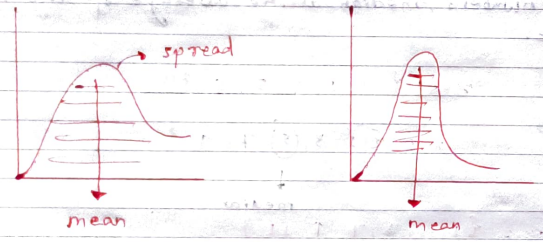
e.g:- $x = \{(2), 4, 3, 4, (2), 6, 7, (2)\}$

$= 2$ → mode

Appointment | Notes | Work to do

Measure of Dispersion :-

>> Degree of variation.
>> Known as (Spread, Variables).



- Range
- Interquartile Range
- Quartile Deviation
- Percentiles
- Variance
- Standard Deviation

(i) **Range :-** difference b/w maximum and minimum value of data.

range = Maximum - Minimum

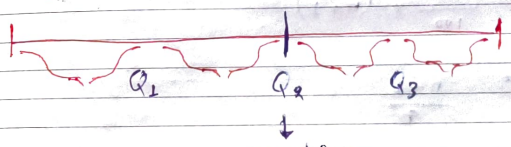
- sensitive to extreme values.

$x = \{1, 2, 3, 4\}$
$= 4 - 1$
$\Rightarrow 3 \rightarrow$ range

Appointment | Notes | Work to do

(ii.) **Quartile Deviation :-** half of the difference between the third quartile and the first quartile in a given data set.

>> divide four equal part with three quartile in data set.



$\Rightarrow Q_1 = \frac{n+1}{4} \quad \Rightarrow Q_2 = \frac{n+1}{2} \quad \Rightarrow Q_3 = \frac{3(n+1)}{4}$

(i) $x = \{4, 6, 7, 8, 10, 23, 34\}$ (a_1, a_2, a_3)

$\Rightarrow Q_1 = \frac{n+1}{4} \Rightarrow \frac{7+1}{4} \Rightarrow \frac{8}{4} \Rightarrow 2$

$\Rightarrow Q_1 = 6$

$\Rightarrow Q_2 = \frac{n+1}{2} \Rightarrow \frac{7+1}{2} \Rightarrow \frac{8}{2} \Rightarrow 4$

$\rightarrow Q_2 = 8$

08 SUNDAY

$\Rightarrow Q_3 = \frac{3(n+1)}{4} \Rightarrow \frac{3(7+1)}{4} \Rightarrow \frac{3 \times 8}{4} \Rightarrow \frac{24}{4} \Rightarrow 6$

$Q_3 = 23$

Appointment Notes Work to do

iii) Percentiles :-

>> Divide data into 100 equal parts.

>> $P_{25} = \frac{p}{100} \times (n+1)$

$= \frac{25}{100} \times (n+1)$

$= \frac{n+1}{4}$

iv.) Interquartile Range :-

>> The difference between upper (Q_3) and lower (Q_1) quartile.

>> $Q_3 - Q_1$

>> Less sensitive to extreme values.

v.) Variance :- the average of the square deviation from the mean of the given data set.

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

x_i = individual data points

μ = mean

Appointment Notes Work to do

vi.) Standard Deviation :-

>> Square root of arithmetic average of the square of the deviations measured from the mean.

>> unit is same to the original data set.

$\therefore$ std = $\sqrt{\text{variance}}$

Frequency :-

>> Number of times, a value of data occurs.

>> Categorical variables.

>>> Frequency Distribution Table :-

Example :- {1, 3, 3, 2, 4, 1, 2, 4, 1, 2, 3, 5, 4, 1, 2, 1, 3, 4, 1}

=>

Data Value	Frequency
1	6
2	5
3	4
4	3
5	1

Appointment | Notes | Work to do

* Relative Frequency :

- percentage or portion of the data value present in complete dataset.
- R.F = $\frac{\text{Frequency}}{\text{total no. of observation}} \times 100$

Data Value	Frequency	Relative Frequency
1	6	6/20 = 3.33
2	5	5/20 = 4
3	4	4/20 = 5
4	3	3/20 = 6.66
5	1	1/20 = 20
	20	

* Cumulative Frequency :

>> way to show the running total frequency as moved through the categories.

>> statistical method that can be used to organize and examine data.

e.g : Customer complaints at a Company

	C.F
> Shipping Delay : 15 complaints	15
> Billing Issues : 10 complaints	15+10 = 25
> Returns : 12 complaints	25+12 = 37

37 - complaints

Appointment | Notes | Work to do

Graphical Representations :-

>> A way to representing any data in picturized form.

>> help reader to understand the large set of data very easily as it gives us various data patterns in visualized form.

→ Types :-

(i.) Histogram : F 
(ii.) Boxplots
(iii.) Scatter plots
(iv.) Line Graphs
(v.) Bar Graphs
(vi.) Line Plot
(vii.) Pie Chart

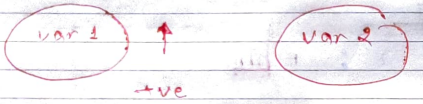
Appointment | Notes | Work to do

Covariance :-

» statistical method (measure)

» Describe that how much two variables change together.

» Explain direction (not the strength of the relationship)...



* Correlation :-

» describes the strength between two variables.

» Linear relation.

» Direction of the relationship. $(-1 < \rho < 1)$ (-ve / +ve)

* 0.8 → strong (+ve)
* 0.2 → (+ve) ↓ no-relation
* $0.5 < \rho < 0.5$ → no-relation

⇒ Limitations :

- outliers
- it does not measure non-linear relationships.

Appointment | Notes | Work to do

⇒ X and Y

› pearson's correlation coefficient.

$$\rho(x,y) = \frac{cov(x,y)}{std(x) \cdot std(y)} = \frac{cov(x,y)}{\sqrt{var(x)}\sqrt{var(y)}}$$

∴ correlation

⇒ X and Y

» $cov(x,y) = \sum_{i=1}^{N} \frac{(x_i - \mu_x)(y_i - \mu_y)}{N}$

∴ covariance

* Causation :-

» Relationship between cause & effect.

» Direct connection in which one variable influences the others.

15 SUNDAY

Appointment | Notes | Work to do

Probability :-

» measure of likehood of a particular event occurring.

» measure between 0 and 1. (0 ⟷ 1).

* 0 : impossibility
* 1 : certainty

» 0.5 (50%), 0.6 (60%) : varying degree of likehood.

(Basic Concept) :-

- Sample Space
- Random Experiment
- Events
- Probability Function

(i.) Sample Space :-

» set of all posible outcomes of random experiment.

example : Tossing a coin.

→ S = { Tail, Head }

Appointment | Notes | Work to do

(ii.) Event :-

» subset of sample space.

eg:- A = Getting an even number.
A = { 2, 4, 6 }

B = Getting prime numbers.
B = { 2, 3, 5 }

(iii.) Probability Function :-

» assigns probability to each event.

eg:- P(getting head) = 1/2 = No. of favourable outcomes / Total no. of outcomes

P(getting even No.) = 3/6

⇒ 1/2 ⇒ 0.5 ⇒ 50%

(iv.) Complement of an event :-

» it consists of all outcomes i.e not in A.

» taking of probability of an event it should not include probability.

eg:- A = { 2, 4, 6 }
A' = { 3, 5, 6 }

∴ A = 0.5
A' = 1 - P(A)

Appointment | Notes | Work to do

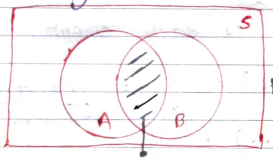
Event :-

» subset of sample space.

Types :-

(i) Joint Event :-

- non - disjoint events.
- non - mutually exclusive.



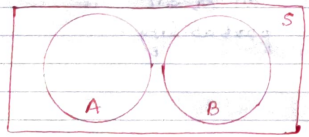
Common outcomes

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

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

(ii) Disjoint Event :-

- mutually exclusive.
- do not occur at sametime.



$P(A \text{ and } B) = 0$

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

Appointment | Notes | Work to do

(iii.) Dependent Event :-

- there event affect each other.

$\star \quad A \rightarrow B$

eg:- A = drawing a red card
B = drawing a black card.

$A = \frac{26}{52}$

$B = \frac{25}{51}$

$\therefore P(A \text{ and } B) = P(A) * P(B|A)$

$\because P(B|A) =$ P of B given A.

(iv.) Independent Events :-

- not affecting each other.

eg:- A = tossing a coin — {H, T}
B = " " 2nd coin — {H, T}

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

MAY JUN

Appointment | Notes | Work to do

Conditional Probability :-

>> probability of an event based on the occurance of previous event.

>> previous event has already been occurred.

∴ P(A|B) = probability of A given B already occurr.

⇒ $P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$

Bayes Theorem :-

>> Reverend Thomas Bayes.

>> update the probability of an event based on new evidence / information.

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

- P(A|B) = update probability
- P(A) = initial belief
- P(B) = marginal probability
- P(B|A) = evidence that supports the initial belief.

Appointment | Notes | Work to do

→ Use Case :-

* Medical diagnosis
* Spam Classification (email spam)
* Recommendation System
* Fraud Detection.

example: Suppose there are 1000 people from which 100 are healthy and 100 are sick and they diagnosed using a test that is 99% effective.

	Diagnosed Healthy	Diagnosed Sick	
Healthy	891	9	→ error
Sick	1	99	

→ P(S|DS')

⇒ P(S) = 100/1000 = 0.1

→ P(S') = 900/1000 = 0.9

22 SUNDAY

→ P(P|S) = 0.99

→ P(P|S') = 0.01

→ P(P) = 0.108

∴ $P(S|P) = \frac{P(P|S) \times P(S)}{P(P)}$

$= \frac{0.99 \times 0.1}{0.108}$ ⇒ 0.9167 / 91.67 %

Appointment | Notes | Work to do

Probability Distribution :-

» mathematical distribution function that describes the probability of different possible values of a variable.

» Random Variables :
- Discrete : Countable / Finite numbers {0,1}
- Continuous : uncountable numbers {0,1}

• PMF (Discrete Random Variables) :-
(probability mass function)

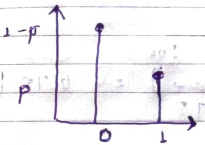
• PDF (Continuous Random Variables) :-
(probability density function)

Appointment | Notes | Work to do

Bernoulli Distributions :-

» Discreet distributions
» outcomes : {0, 1}

∴ PMF = $\begin{cases} p & \text{if } x = 1 \\ q = 1-p & \text{if } x = 0 \end{cases}$



$\Rightarrow f(x,p) = p^x (1-p)^{1-x}$

mean = p
variance = pq | p(1-p)

» multiple trials

Binomial Distributions :-

» single trials
» (n = 1)
» outcomes : {0, 1}

∴ PMF = $f(x : n, p) \Rightarrow {}^nC_x \, p^x (1-p)^{n-x}$

$\Rightarrow \frac{n!}{(n-x)!\,x!} \cdot p^x (1-p)^{n-x}$

* mean = n.p
* variance = npq $\Rightarrow$ np(1-p)

Appointment | Notes | Work to do

Example :- Customer conversion rate

Imagine you're an e-commerce business owner, and you want to analyze the conversion rate of visitors to your website. You're interested in understanding the likelihood of specific no. of purchases (successes) out of fixed no. of website visits during a given period.

n (no. of website visits) = 100
p (probability of making purchase) = 0.75 (75% rate)
x (no. of purchases) = 75

Question :- What is probability of exactly 75 visitors making a purchase out of website visits given a 75% conversion rate.?

$$\Rightarrow \text{PMF} = \frac{100!}{(100-75)! \times 75!} \times 0.75 \times (1-0.75)^{25}$$

$$= 0.0917$$
$$= 9.17\ \%$$

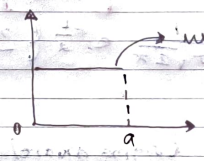
Appointment | Notes | Work to do

→ Uniform Distribution :-

» Continuous distribution.

» Characterized by a constant probability density (CPD) over the specify interval.

» Equal chance to occurs.



∴ PDF = $f(x) = \frac{1}{b-a}$

$\Rightarrow a \leq x \leq b$

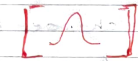
- mean = $\frac{a+b}{2}$
- variance = $\frac{(b-a)^2}{12}$

Appointment Notes Work to do

Normal Distributions :-

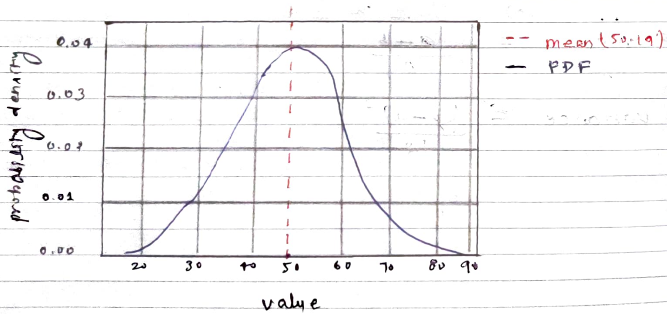
>> Gaussian Distribution

>> it appears as a "bell curve" when graphed.
>> describe a symmetrical plot of data around its mean value.

→ 

>> PDF ⇒ $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

* mean : μ
* variance : σ^2 (sigma square)
* std. deviation : σ



⇒ mean = 50
⇒ std = 10

Appointment Notes Work to do

Standardization :-

>> process of converting normal distribution into standard normal distribution.

>> mean = 0
standard deviation = 1

>> $z = \frac{x - \mu}{\sigma}$

→ Normalization :-

>> Rescales a dataset so that each value falls between 0 & 1.

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

example :- Salary : minimum = 3 Lakhs
maximum = 50 "
x = 4 Lakhs

⇒ $\frac{4-3}{50-3}$ ⇒ $\frac{1}{47}$ ⇒ 0 ⟷ 1

30 MONDAY
DAY (120-245)
18th Week

April

Appointment | Notes | Work to do

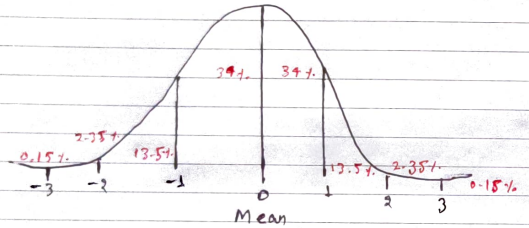
Normalization	Standarization
Minimum and maximum value for variable are used for scaling.	Mean and standard deviation of variables are used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scale value between 0 and 1.	It is not bounded to a certain range.
It was affected by outliers.	It is less effected by outliers.
It is useful when we don't know about the distribution.	It is useful when the feature distribution is normal or gaussian.
Maps the data to a specific range, which may leads to loss of information about the original distribution.	Preserves the shape of the original data distribution but shifts and scales it.

Appointment | Notes | Work to do

Empirical Rule :-

>> also known as 68 - 95 - 99.7

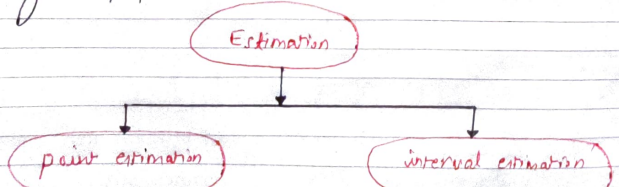
>> describe how data is distributed in a normal distribution.



Inferential Statistics :-

>> involves the use of a sample (1) to estimate some characteristics in a large population.

>> (2) to test a research hypothesis about a given population.



Appointment | Notes | Work to do

- **Estimation :-** a process of drawing conclusion for population based on sample data.

>> population parameter
>> Sample statistics

(i) **Point Estimation :-**

>> provide a single value as your best guess for unknown population parameter.

>> **Properties :-**

- Consistent : larger the sample size : more accurate estimation.
- Unbiased

>> **Drawbacks :-**

(i) problem of reliability
(ii) not enough evidence

Appointment | Notes | Work to do

(ii) Interval Estimation :-

>> provide intervals, in which population parameters fall.

>> more accurate than point estimation.

Confidence Interval :-

>> an interval of values i.e computer from sample data i.e likely to contain the true population parameter.

>> 95% — 99% (confidence coefficient)

$$\therefore \quad CI = \bar{x} \pm z \frac{s}{\sqrt{n}}$$

$\bar{x}$: sample mean
z : confidence level value
s : sample standard deviation
n : sample size

Appointment | Notes | Work to do

Student's T Distribution :-

>> used while assumptions about a mean when we don't know the standard deviation.

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

Hypothesis Testing :-

>> statistical method used to evaluate a claim / hypothesis about population parameter.

>> proving True / False.

>> comparing sample data to a hypothesis about a population parameter to determine whether there is enough evidence to reject or correct the hypothesis.

>> without hypothesis test, you risk drawing wrong conclusion or making bad decisions.

→ Types :-

(i) Null
(ii) Alternate

Appointment | Notes | Work to do

(i) **Null Hypothesis :-** (H_0)

» baseline
» default assumption
» " position

» Accept | Reject

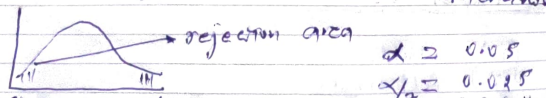
(ii) **Alternate Hypothesis :-** (H_1)

» Logical opposite of null hypothesis.
» there is significant difference or effect.

Example :- Sanitizer company claims that their product kills 99% germs ?

06 SUNDAY → H_0 : $\mu = 99\%$.
H_1 : $\mu < 99\%$

→ score | level
∴ level of significance (α) : predetermined threshold



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

(ERROR) :

Appointment | Notes | Work to do

→ **Error in Hypothesis Testing :-**

» Error occurs when conclusions are drawn that are incorrect.

* **Types of Errors :-**

(i) Type I error : (false +ve)

(ii) Type II error : (false -ve)

- Type I errors occur when the null hypothesis is rejected when it is true.

- Type II error occur when null hypothesis is accepted when it is false.

	Reject H_0	Accept H_0
H_0 True	Type I error	correct
H_0 False	correct	Type II error

Appointment | Notes | Work to do

Example : You are conducting a medical test to detect a rare disease.

H_0 : patient is healthy

H_1 : patient has disease

⇒ Type I error : patient has a disease when he does not or he is healthy.

⇒ Type II error : patient is healthy when he suffer from that disease.

Appointment | Notes | Work to do

TEST :-

» a statistical procedure used to determine if there's enough evidence in a sample data to draw conclusion about a population, by comparing the observed data to what's excepted under a null hypothesis.

* Types of Tests :-

(i) One Tailed : (unidirectional)

» critical region for rejecting H_0 is located in only one tail of the distribution.

(ii) Two Tailed : (bi-directional)

» critical region for rejecting H_0 is divided between both tails of the distribution.

Appointment | Notes | Work to do

Test in Statistics :-

(i) Z-Test :-

>> Compare sample mean with population mean.

>> population standard deviation is known.

>> useful for large sample size ($n > 30$)

* One sample Z-test :-

- when we have single sample to compare with population mean.

$$\left[\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}\right]$$

* Two sample Z-test :-

- two independent samples.
- Compare their mean

$$\left(\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right)$$

Appointment | Notes | Work to do

(ii) T-Test :-

>> sample size small ($n < 30$)

>> Compare mean of two groups or mean of a sample to a unknown value.

>> when σ (sigma) (μ (mu)) is unknown.

* Independent T-test :-

- Compare mean of two independent groups to determine if there is any significant difference between them.

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

* Paired T-test :-

- Comparing means of two related group such as before and after measurement on the same subjects.

$$\Rightarrow \frac{\bar{x}_{diff}}{\frac{s_{diff}}{\sqrt{n}}}$$

Appointment | Notes | Work to do

(iii.) Anova Test :- (Analysis of Variance)

>> used to determine if there's a statistically significant difference between the means of three or more groups, generalizing the t-test beyond two groups.

>> F-value (F-test)

>> Compare more than two groups.

* One-way Anova :-

• used when you have one independent variable (factor) with multiple levels.

* Two-way Anova :-

• used when you have two or more independent variables.

13 SUNDAY

Appointment | Notes | Work to do

(iv.) Chi-square Test :-

>> compare two categorical variables.

>> Test the association or independence between categories.

* Chi-square Test for independence :-

• used to determined if two categorical vanables are related.

* Chi-square Test for Goodness-of-fit :-

• checks if a sample distribution matches an expected distribution.