Statistics.

** What is statistics and its types :-

* Defination :- statistics is the science of collecting organizing and analyzing ~~data~~ and interpreting data to support decision making and understanding.

* Data :- Facts or pieces of information.

* Types of statistics :-

Descriptive statistics (understanding data)

Def^n:- It consist of organizing and summarizing of data.

(a) Measure of central tendency (mean, mode, median)

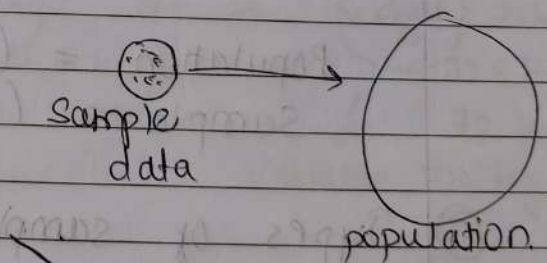
(b) Measure of disperssion (Variance & standard deviation)

(c) Shape
Different type of distribution of data eg. skewness, kurtosis.

(d) Histogram, probality density function & probability mass function. (Visuals)
↙ frequency of the element.

Inferential statistics (Drawing conclusions of data)
It consist of using data; you have measured to form conclusion.

(a) Z-test
(b) T-Test
(c) Anova test
(d) Chisquare test
} Hypothesis testing



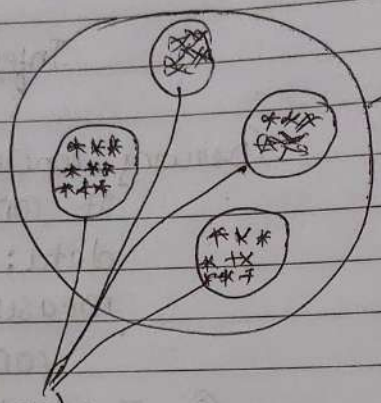
4W for descriptive statistics.
what → summarization
why → find pattern & insights
when → before decision making
where → everywhere.

* Types of Data :-

* Sampling Technique :-

(a) population - The group that we are interested in studying.

(b) Sample - A subset of the population.



The goal of sampling is to create a sample that is representing of the entire population

Population = (N)
Sample = (n)

(c) Types of sampling :-

1) Simple random sampling :- When performing simple random sampling every member of the population (N) has an equal chance of being selected for your sample (n)

2) Stratified sampling
↓
Layering (non overlapping groups)

population

male female

They does not overlap.

3) Systematic sampling :- Mall → outside

n^{th} person will choose for survey.

↓

4^{th} person

↓

survey.

4) Convenience sampling :-

Voluntery Response sampling

Survey

Data Science

↓↓

Data science knowledge.

* What are variables & it's types :-

Defn - Variable is a property that can take on many values.

eg. (age) = 12, 13, 40, 100 ↓ variable

(height) = 172 cm, 172.5 cm, 180 cm. ↓ Variable.

(weight) = 72 kg, 72.5 kg, 73 kg, 100 kg ↓ Variable

Variable is only in singular mode

Ages = {15, 20, 25, 30}

plural = that's why it's not variable.

* Types of Variables :-

Quantitative variable

Discrete variable
eg. whole number
& no decimal value especially in output.

eg
1) No. of bank account
2) No. of childeren.

Continueous variable.
eg. 1) height = 175.50 cm
2) weight = 72.3 kg

Qualitative or categorical variable

properties

Classification
eg. 1) Grender < male / female
2) Type of flower
Rose, lotus, lily

Nominal (No order matter)
Ordinal (order matters) eg. size, feedback

(*) Measure of Central Tendency :-

① Mean
② Median
③ Mode

Defn

central tendency refers to the measure used to determine the "center" of the distribution of data.

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

① Mean (Average)

Population Data (N)

μ = population mean

$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$

Sample data (n)

$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

Σ = Summation
x_i = particular dat point
i = row number.

$$\mu = \frac{1+2+2+3+4+5}{6}$$

$$\mu = \frac{17}{6}$$

$$\mu = 2.83$$

If will focus on first mean i.e. 2.83 & after adding one outlier it turned to 16.71. But when we worked on median even though we added two outlier it's 3.5. This is how median deal with outlier.

(2) Median ∵ $x = \{1, 2, 2, 3, 4, 5, 100\}$
help to deal with outlier ↳ outlier

↓
A datapoint who doesn't follow the pattern of data or are extreme points.

$$\mu = \frac{1+2+2+3+4+5+100}{7}$$

$$\mu = \frac{117}{7}$$

mean = μ = 16.71

(*) Sort all the numbers
(*) find the central element. < odd length / even length

$\{1, 2, 2, \boxed{3}, 4, 5, 100\}$ → odd length → $\left(\frac{n+1}{2}\right)^{th}$
↓ Median

$\{1, 2, 2, \boxed{3, 4}, 5, 100, 101\}$ → even length. → $\frac{\left(\frac{n}{2}\right)^{th} + \left(\left(\frac{n}{2}\right)+1\right)^{th}}{2}$ ↑ for even length

$$\frac{3+4}{2} = 3.5$$ Median

③ Mode :- Most frequent element. We mostly use of categorical feature.
{1, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 5}

$$\text{Mode} = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h$$

Question :- find mode of following data:-

Age (in years)	No. of patients (f)
5-15	6
15-25	11
25-35	21 f_0
35-45	23 f_1
45-55	14 f_2
55-65	5

lower limit (l)

difference bet^n class interval (or range) $h = 15 - 5 = 10$

$$\text{Mode} = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h$$

$l = 35$
$f_0 = 21$
$f_2 = 14$
$h = 10$
$f_1 = 23$

$$= 35 + \left(\frac{23 - 21}{2(23) - 21 - 14}\right) \times 10$$

$$= 36.8$$

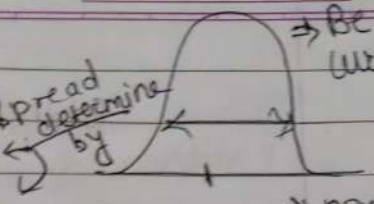
M T W T F S S

Page No.: YOUVA

Date:

★ Measure of Dispersion :- → Spread.

1) Variance
2) Standard deviation



★ Mean :-

Sample mean (n)

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Population mean (N)

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

★ Variance :-

Sample variance

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

Bassel correction or also known as degree of freedom

Population variance

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

(population variance)

e.g. $X = \{1, 2, 2, 3, 4, 5\}$

x	$\bar{x}$	$(x - \bar{x})$	$(x - \bar{x})^2$
1	2.83	-1.83	3.34
2	2.83	-0.83	0.6889
2	2.83	-0.83	0.6889
3	2.83	0.17	0.03
4	2.83	1.17	1.37
5	2.83	2.17	4.71
$\bar{x} = 2.83$			10.84

$$S^2 = \frac{10.84}{5} = 2.168$$

↓ sample variance

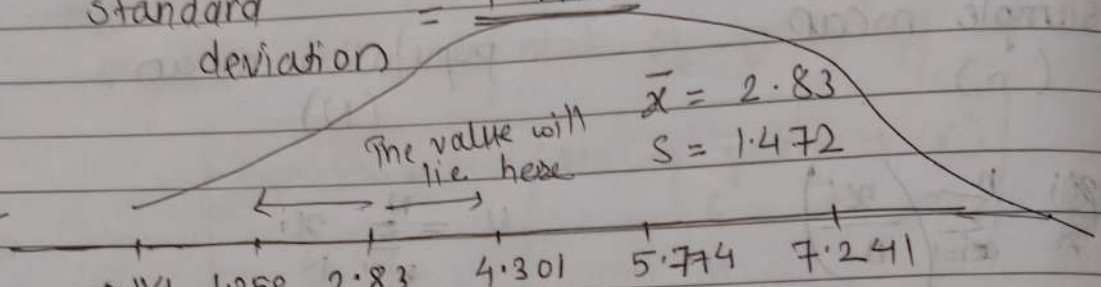
⇒ {Spread of the data}

Standard deviation =

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

$= \sqrt{2.168}$

Sample standard deviation = 1.472



2.83 + 1.472 = 4.301

+ 1.472 = 5.774

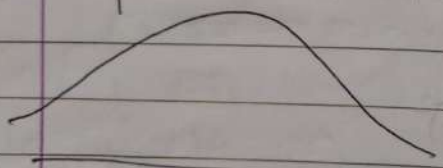
+ 1.472 = 7.246

Left direction will substract

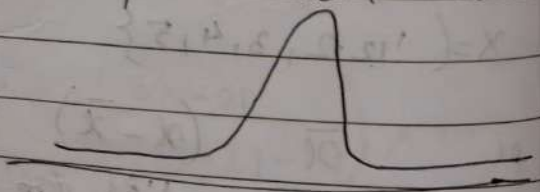
Right direction will add.

Variance decide the spread.

Variance → Big number
S.d. → Big number
Spread will increase



Variance → small number
S.d. → small number
spread will decrease



** Percentile & Quartiles :-

(*) Difference between percentage & percentile :-

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

% of the numbers that are even ?

% of even = $\frac{2}{5}$ = 0.4 = 40% of the numbers are even.

Percentiles :- eg. {CAT, GATE, SAT}

Defn :- A percentile is a value below which a certain percentage of observations lie.

95 percentile means that the person has got better marks then 95% of the entire student.

e.g.

(**) Dataset :- {2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12}

n = 20

(1) What is the percentile ranking of 10 ?

percentile rank of 10 = $\frac{\#\text{ of values below } 10}{n} \times 100$

= $\frac{16}{20} \times 100$

= 80% (percentile)

(2) What is the percentile ranking of 11 ?

Date:

③ What value exist at percentile rank of 25 ?

$$\text{value} = \frac{\text{Percentile}}{100} \times (n+1)$$

$$= \frac{25}{100} \times 21 = 5.25$$ (This the Index value)

now at 5 position there is number 5 and at here its 5.25 so, will consider next index as well so, it will be 5 again will do the average of both as $\frac{5+5}{2} = 5$

so, the value will be 5

* Quartiles :- That means 1/4th

Also known as five number summary

minimum
25 % → 1st quartile (Q_1)
median
75 % → 3rd quartile (Q_3)
maximum

Inter quartile range (IQR) = $Q_3 - Q_1$
This help to find outliers.

* Scales of Measurement :-

It defines how the data is classified, compared & used mathematically.

* Properties of measurement levels

- Qualitative
  - (a) Nominal
  - (b) Ordinal
- Quantitative
  - (a) Discreate
  - (b) Continueous.

* Defⁿ of Measurement :- Assignment of numbers to characteristics of objects or events according to rules.

* Properties of scales of measurement :-

(1) Identity :- Each value has a unique meaning or category.
(2) Order :- Values can be ranked or ordered logically.
(3) Equal Intervals :- The difference between values is meaningful and equal.
(4) True zero :- A value of zero means the absence of the quality. (eg. zero weight = no weight)

Types of data from pdf

Nominal :- categorizes data without a specific order. eg. Gender, blood type.

ordinal :- categorizes data with a meaningful order. eg. education level, ranking.

Interval :- Numeric data with ordered categories and equal intervals but no true zero. eg. Temp., IQ score

Ratio :- like interval but with a true zero point: eg. Height, weight, age, income.

(*) Interquartile Range. (IQR)

$IQR = Q_3 - Q_1$

$Q_3 = 9 \quad Q_1 = 5$

$IQR = 4$

rough example

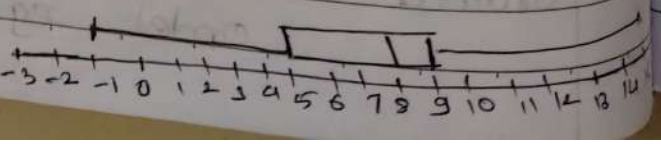
min → lower fence $= Q_1 - 1.5(IQR)$
$= 5 - 1.5(4)$
$= 5 - 6$
$= -1$

max → higher fence $= Q_3 + 1.5(IQR)$
$= 9 + 1.5(4)$
$= 9 + 6$
$= 15$

$x < -1$ = Outlier
$x > 15$ = Outlier.

Box plot
min = -1
25th = 5
median = 8
75th = 9
max = 15

$Q_3 - Q_1$



2nd example :-

$x = \{1, 2, 2, 2, 3, \underline{3}, 4, 5, 5, 5, 6, 6, 6, 7, 8, \underline{8}, 9, 15, 12, 7\}$

$$Q_1 = \left(\frac{25}{100} \times 20\right) + 1$$

$= 6^{th}$ Index

$=$ On 6^{th} Index $= \boxed{3}$

$$Q_3 = \left(\frac{75}{100} \times 20\right) + 1 = 16$$

$= 16^{th}$ Index

$= \boxed{8}$

$IQR = Q_3 - Q_1$

$= 8 - 3$

$= 5$

Min - Lower fence $= Q_1 - 1.5(IQR)$

$= 3 - 1.5(5)$

$= -4.5$

Max - higher fence $= Q_3 + 1.5(IQR)$

$= 8 + 7.5$

$= 15.5$

min $= -4.5$

$Q_1 = 3$

Median $= 5.5$

$Q_3 = 8$

Max $= 15.5$

Date:

* Descriptive statistics. *

Defⁿ Descriptive statistics is a branch of statistics that summarizes and describes the main feature of a dataset. It helps us understand the basic characteristics of data through numerical measures and graphical representations.

(A) Types of D.S. :-

1) Measure of central Tendency :-
This refers to the measure used to determine the "center" of distribution of data.

(a) Mean :- The average of the data.
(b) Median :- The middle value when data is ordered.
(c) Mode :- The most frequent value.

2) Measures of Dispersion (spread) :-
These show how spread out the data.

(a) Range :- Difference between the maximum & minimum values.

(b) Variance :- Average of the squared differences from the mean.

(c) standard deviation :- Square root of the variance shows average distance from the mean.

3)

(B)

Mea
(e

(d) Interquartile Range (IQR) :-
Difference between the 75th percentile (Q_3) and 25th percentile (Q_1)

3) Shape of the Distribution

(a) Skewness :- Measures the asymmetry of the distribution.

(b) Kurtosis :- Measures the "tailedness" of the distribution.

(B) Data Visualization Tools :-

1) Histogram :- shows the distribution of data.
2) Bar chart :- Useful for categorical data.
3) Pie chart :- shows proportions of categories

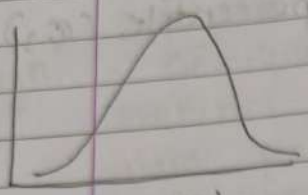
Descriptive Statistics

Measure of Central Tendency
1) Mean
2) Median
3) Mode

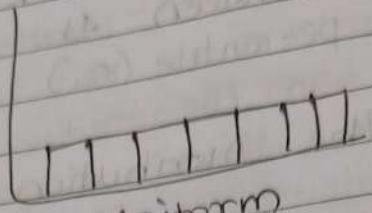
Measure of Variability / Dispersion
1) Range
2) IQR
3) variance
4) std. deviation

Shape / Freq. Distribution
1) Skewness
2) Kurtosis.

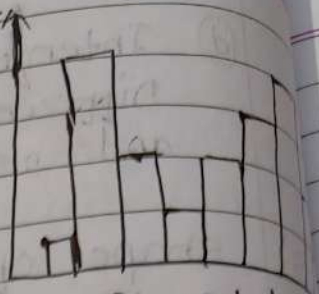
* Different shapes of Distribution :-



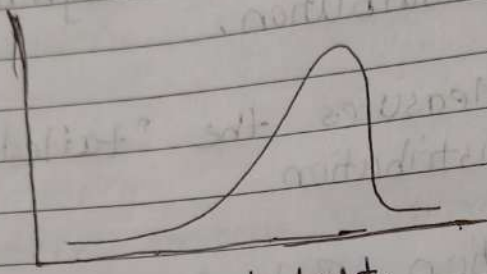
symmetric



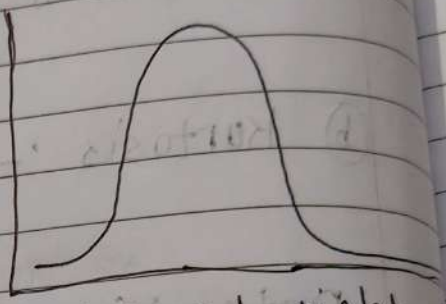
Uniform



Bimodal



skewed left.
Mean > Median
(+vely skewed)

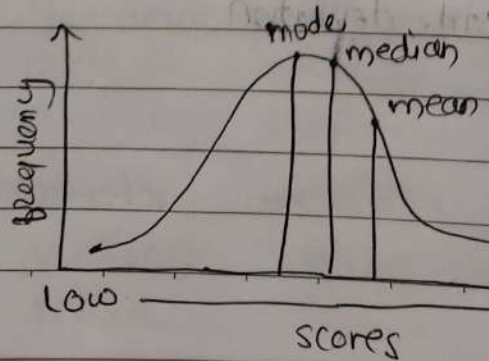


skewed right
Mean < Median
(-vely skewed)

(A) Skewness :- measures lack of symmetry in data distribution.
(less symmetric data, more the skewness)

* for normal distribution skewness is zero.
* Any kind of symmetrical data shows skewness close to zero.

(a) Right - skewed distribution (+ve skewness)



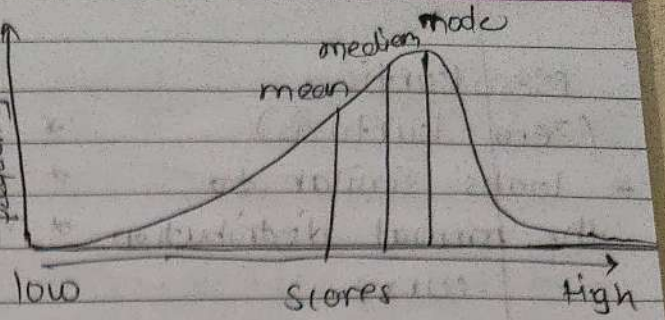
(a) most of the value are around left of the distribution.

(b) Mean, median greater than the mode.

(b) Left - Skewed distribution (Negative skewness).

→ @ most of the values are clustered around right.

ⓑ Mean, median less than the mode.



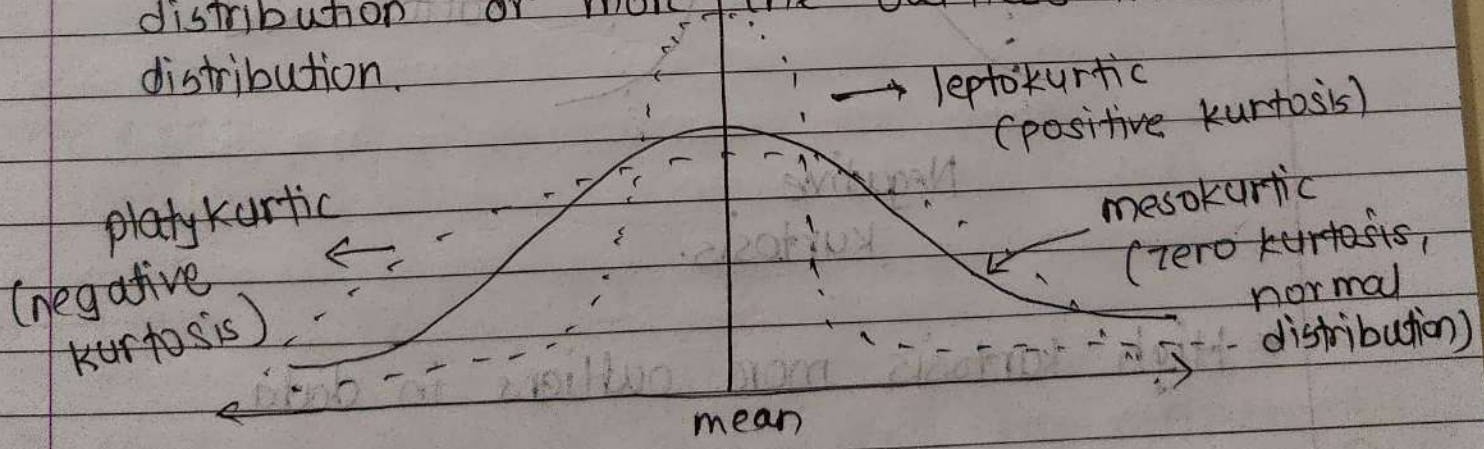
(*) If skewness is less than -1 or greater than 1 :- Highly skewed.

(*) If skewness is between -1 and -0.5 or between 0.5 and 1 :- Moderately skewed

(*) If skewness is in between -0.5 and 0.5 :- Approximately symmetric.

(B) Kurtosis :- measures ~~we~~ wheater data is heavy tailed or light-tailed.

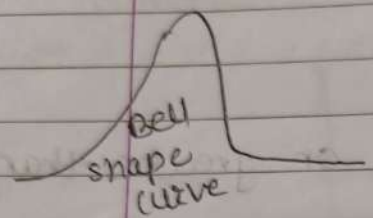
Higher the kurtosis, heavier the tail of data distribution or more the outliers in data distribution.



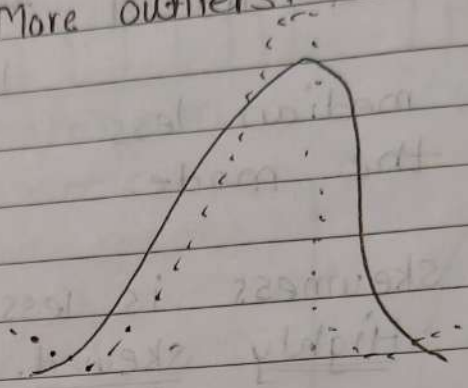
kurtosis distribution

* negative (-ve)
* Zero
* Positive (+ve)

Mesokurtic
(Zero kurtosis)
* looks similar to the normal distribution curve.



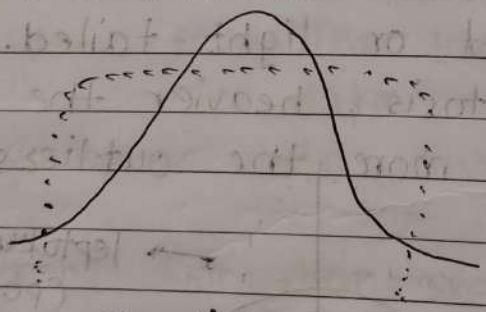
Leptokurtic
(+ve kurtosis)
* longer distribution
* Sharper peak & heavier tails
* More outliers.



Probability of having outliers or extreme values is zero or close to zero.
kurtosis = 0

Positive kurtosis
kurtosis > 3

Platykurtic
* shorter distribution
* flatter peak & Thinner tails
* Less outliers.



Kurtosis < 3

Negative kurtosis.

* High kurtosis more outliers in data.

(*) When is kurtosis useful :-
- Only when its used in conjunction with standard deviation.

→ It is possible that a random variable might have a high kurtosis (bad) but the overall standard deviation is low (good).

→ Conversely, one might see another random variable with a low kurtosis (good), but the overall standard deviation is high (bad).

* Does sample size have an impact on skewness & kurtosis?
They are always increased with increasing sample size.

(A) Skewness :-

$$\frac{n}{(n-1)(n-2)} \sum \left[\frac{(x_i - \bar{x})^3}{S.D.} \right]$$

x = data point
$\bar{x}$ = mean
S.D. = standard deviation
n = no. of observation

(B) Kurtosis :-

$$\frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \frac{(x_i - \bar{x})^4}{S.D.} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

* Data Visualization Tools :-

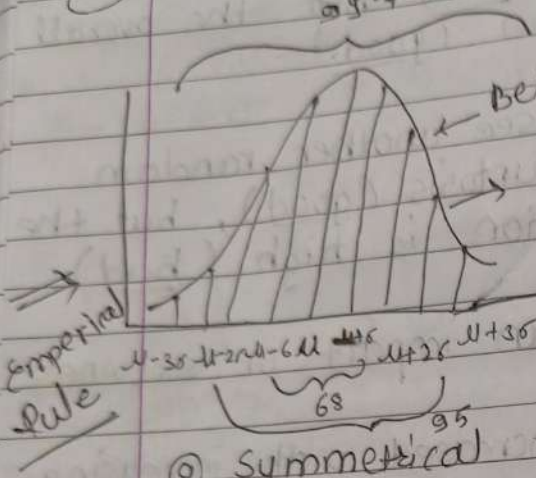
(1) ~~How to construct a box plot for outliers~~ :-

(2)

Date: YOUVA

⊛ Normal Distribution / Gaussian Distribution

68 - 95 - 99.7
(empirical Rule)
(3- sigma rule)



After smoothing it known as pdf (probability density function)

② Symmetrical (equal distribution)

In normal distribution follows 3 sigma rule.

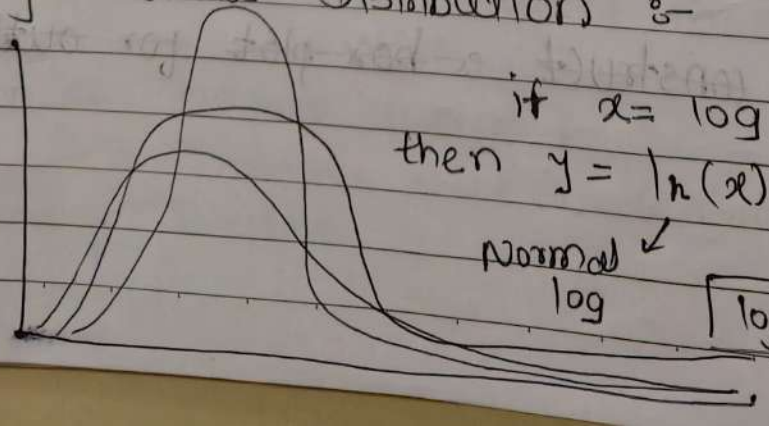
eg. Hight, weight

Iris Dataset ⇒ {sepal legth, petal length}

⊛ Central Limit Theorem :-

whether your distribution is normal distribution or not normal distribution if we take the sample where $n \geq 30$ and if we take all the sample and when we popular or when we plot it we get a normal distribu

⊛ Log Normal Distribution :-



if x = log normal distribution
then $y = \ln(x)$ ← whatever val
Normal log
loge
will get n y
all it's coordina
with x it will be

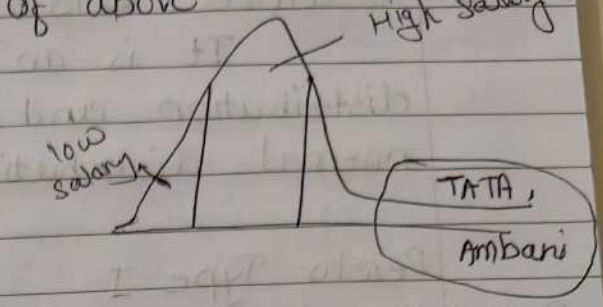
And the transfor known as box co Transfo

Also, we can write instead of above

$x = \exp(y)$

↓

exponencial.

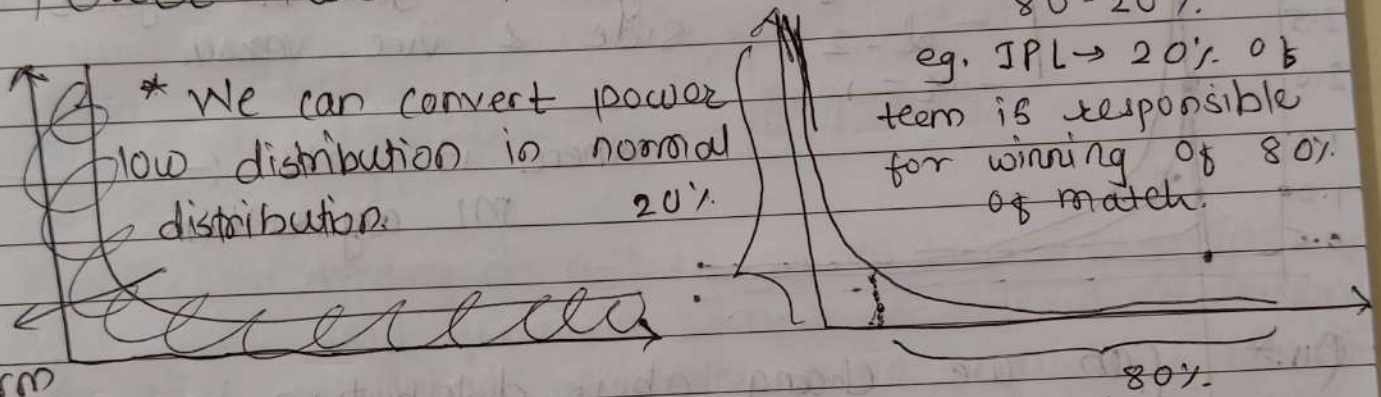


eg. (a) wealth distribution

(b) comments in

where we use :- (1) Machine learning

(*) Power law Distribution :-



* We can convert power law distribution in normal distribution.

And the transform known as box cox Transform.

eg. 80 % of wealth is distributed with 20% of the total population.

80% of the total oil is distributed with 20% of nations.

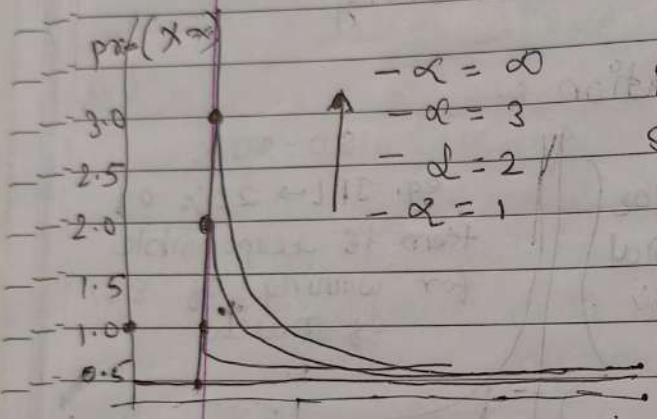
When we convert power law distribution & pareto distribution in normal distribution it's known as box cox transform.

⊛ Pareto Distribution :-

It is an example of power law distribution and this distribution are non normal distribution.

Pereto Type I

Probability density function.



As the element number goes higher it goes on higher side & vice versa.

where it works :-
ML algorithm

Qu.* Can we chang above distribution in normal distribution.

ans. Yes, we can transform in normal distribution The process of transformation known as Box Cox transformation.

eg. ① 20% of the product in amazon is responsible of 80% of sales.

② 20% of defects solves the 80% of upcoming defects.

③ 20% of team is responsible in deliverable 80% of projects.

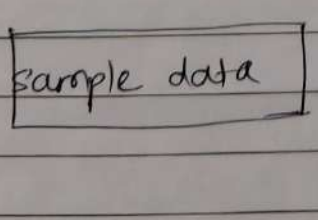
Interview Question :-
Relation betn log normal distribution & pareto distribution?

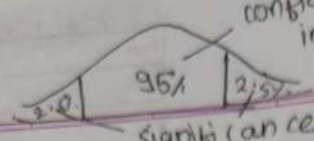
"Hypothesis TESTING"

Defn Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter or a population probability distribution.

* What is hypothesis testing :-
Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter. The methodology employed by the analyst depends on the nature of the data used and the reason for the analysis.

* Inferential statistics :-

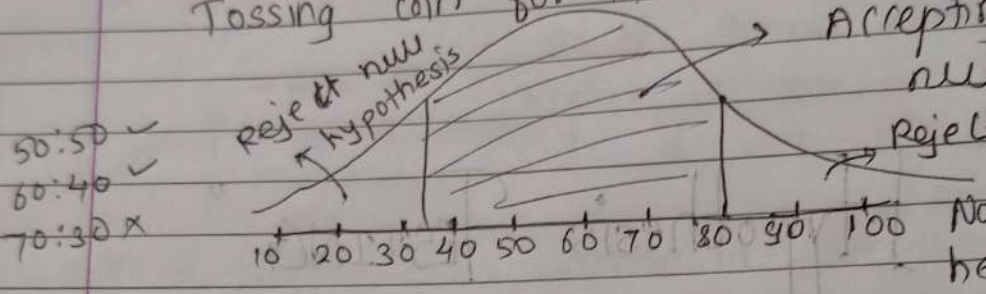




(I = 95%
S.V. = 1 - 0.95 = 0.05 = 5%

Hypothesis testing :- coin is fair or not.

① Null hypothesis (H_0) = Coin is fair.
② Alternate hypothesis (H_1) = Coin is not fair.
③ Experient = where null hypothesis accepted or not accepted.



Domain expert decide the

Confidence interval :- (percentage)
A confidence interval is a range of values used to estimate an unknown statistical parameter, such as a population mean. (Decision boundry)

* 2 Tail Test :-
A two-tailed test is a hypothesis test where the alternative hypothesis claims that the population parameter is different from the value specified in the null hypothesis, meaning it could be either greater than or less than.

* Distribution function :-

① Probability distribution function
② Probability Density function
③ Probability mass function
④ Cumulative distribution function

we understand distribution of data. ← Probability distribution function

Probability density function ↑ distribution
eg continuous Random variable

Probability mass function ↑ Distribution
eg Discreate random variable (whole number)

ⓐ Variables :- $x+5=7$ $3y+x=8$
$x=7-5=2$ $y=6/3=2$

In linear algebra we used to find x & y, etc. So, these are nothing but the variables.

ⓑ Random variable :-
Is a process of mapping the output of a random process or experiments to a number.
eg. ① Tossing a coin (experiment) ⇒ outcome of experiment.

The values in variable keep changing $X = \{0 \text{ if head}, 1 \text{ if Tail}\}$

② Rolling a die (experiment) $= \{0,1,2,3,4,5,6\}$

$y = \{\text{sum of the outcome of rolling a die 5 times}\}$

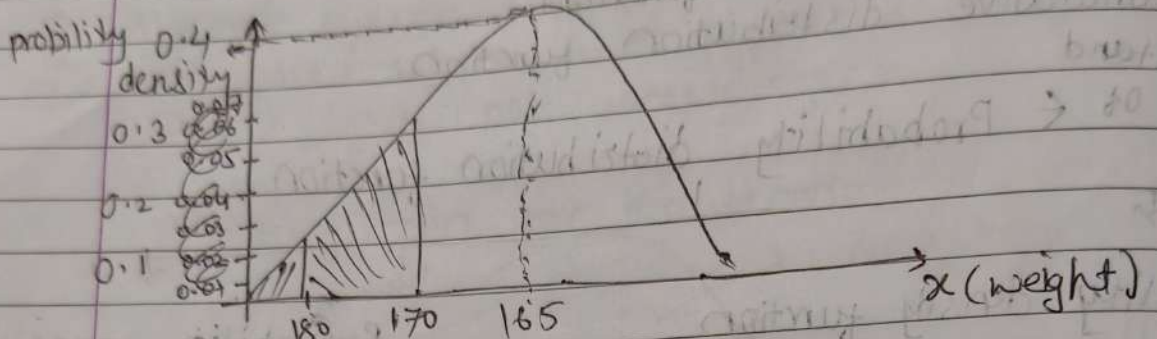
[probability $y \geq 15$] = to find

Interview question

Difference betn Pdf & Cdf

(*) Probability Density function (Pdf)

(@) Distribution of continuous Random variable

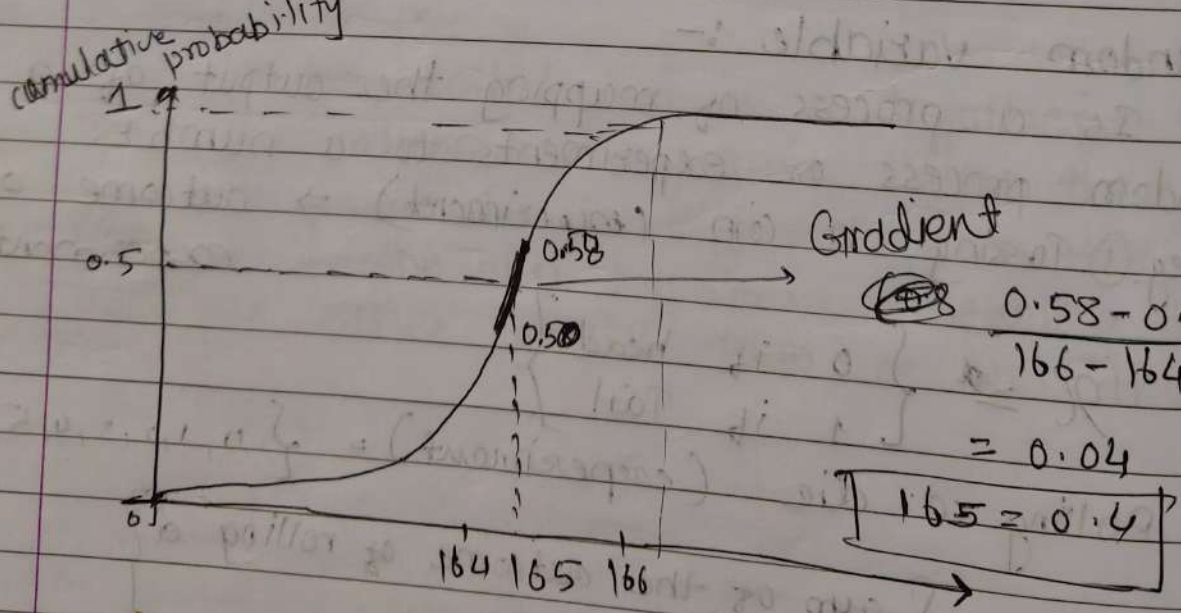


Area under the curve

probability Density = Gradient Decent of cumulative curve

Gradient = movent in the x axis / moment in the y axis

* Cumulative Density function :- (CDf)

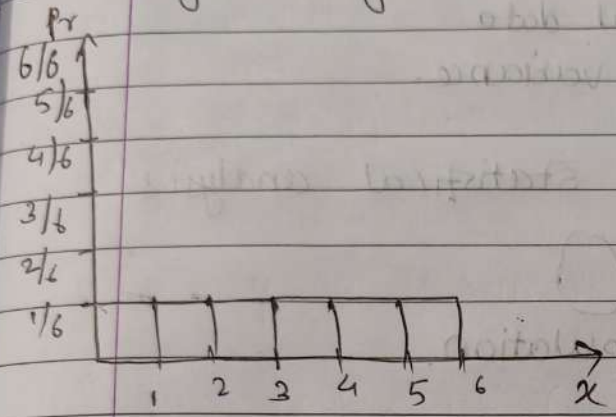


M T W T F S S
Page No.:
Date:
YOUVA

* Probability Mass function :-
Discreate Random variable.

eg. Rolling a dice {1,2,3,4,5,6}

Probability
$P_r(1) = 1/6$
$P_r(2) = 1/6$
$P_r(3) = 1/6$
$P_r(4) = 1/6$
$P_r(5) = 1/6$
$P_r(6) = 1/6$

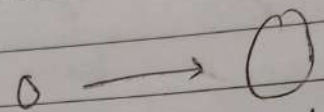


* Difference betn PMF & CDF :-

* Hypothesis Testing & Statistical Analysis:-

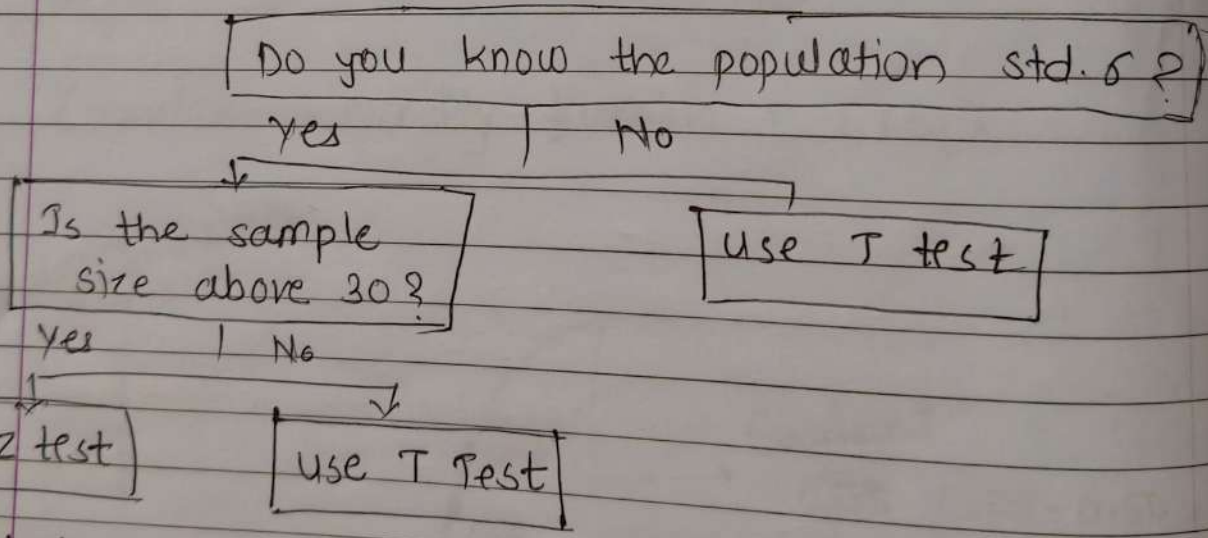
(a) Z test } Average / mean
(b) T Test
(c) CHI square ⇒ categorical data
(d) ANOVA ⇒ Analysis of variance.

This technique use for statistical analysis.



(a) Z Test :- Z Table

When to use T test Vs Z test



* Z-test

The average height of all residents in a city is 168 cm with a population std $\sigma = 3.9$.
A doctor believes the mean to be different.
He measured the hight of 36 individuals & found the average to be 169.5 cm