PAGE No.
DATE 20 8 2015

Statistics

- intro to statistics

Descriptive
- measure of central tendency
- measure of dispersion
- measure of symmetricity

Inferential
- Probability Distribution
- PMF and PDF & CDF
- Central limit theorem
- Statistical test

Definition → Statistics is a mathematical science of including methods of collecting, organizing and analysing data in such a way that meaningful conclusion can be drawn from them.

Why stat in data science →
In DS/ml → We try to learn pattern in data

→ it is raw fact and pieces of information that can be stored, measured and re-accessed.

*** data is used to bring insights to increase the company's revenue ***

collecting → source of data
organizing → to reaccess it effeciently
& analysing → finding insights from data

PAGE No.
DATE

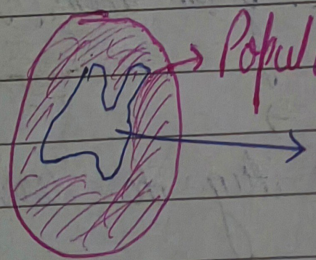
Types of statistics

1> Descriptive Statistics → it consist of organizing and summarizing the complete data / population. → {population summarization}
(describe)

2> Inferential statistics → you can not count no of trees in a jungle if someone ask you this. you will use your inference.
(Inference)

→ It consist of using ~~data~~ sample data that has been measured to form conclusion about a population.

With given sample we conclude something about a population



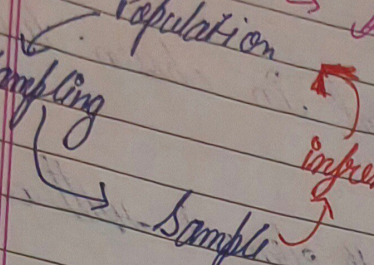
NOTE → When we want exact statistics to take any decison we use Descriptive Statistics

Techniques of descriptive stat

1> Measure of central tendency (mean, median, mode)
2> Measure of symmetricity (skewness, kurtosis)
3> Measure of dispersion (std deviation, Variance)

Techniques of inferential stat

↳ when population is very large
↳ time and resource constraints



* in inferential stat taking sample is very crucial part.

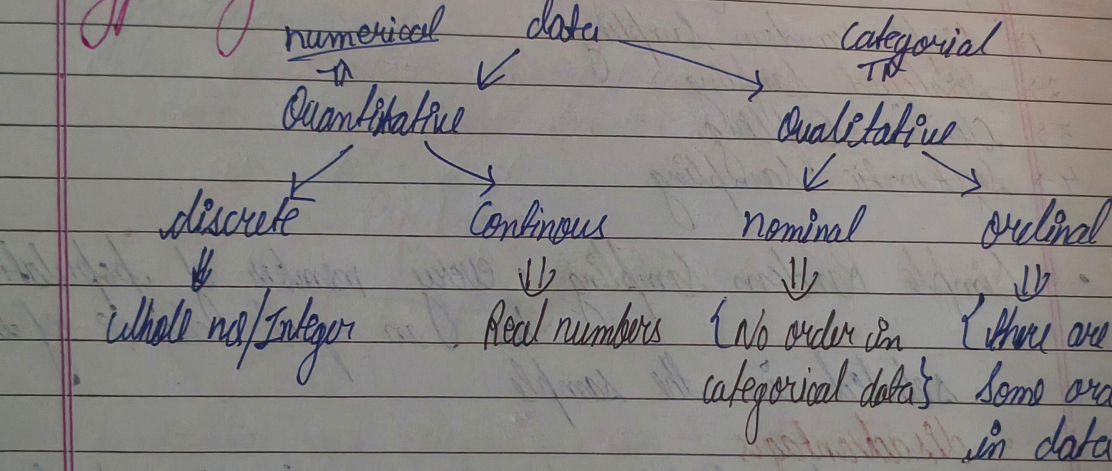
* Types of sampling →

1→ Simple random sampling
2→ Stratified sampling
3→ Cluster sampling
4→ Systematic sampling

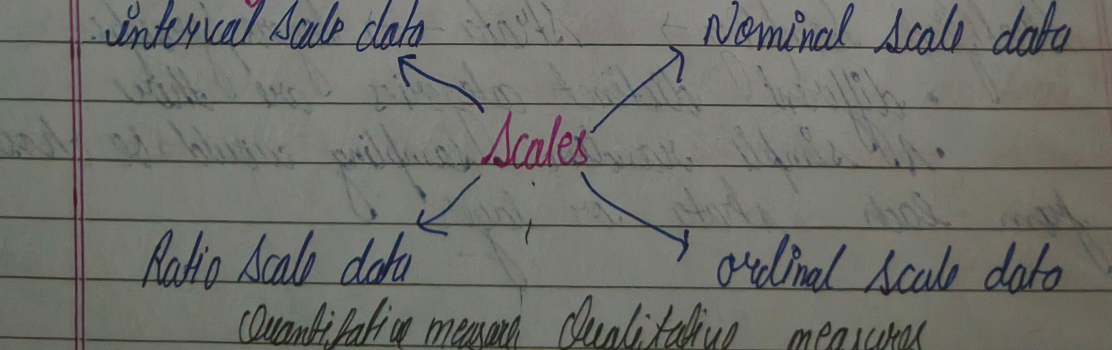
- Simple Random sampling • every member of population(os) has an equal chance of being selected in the sample
  - disadvantages
    - A possibility of members not being part of sample from a certain group.
- Stratified sampling → Strata → layers/groups.
  - different distinct categories are there
  - A simple random sampling would be chosen from each strata or layer.

- Cluster sampling → divides population into group or cluster, then some of these cluster are randomly selected
  → then all individual in chosen cluster are selected in the sample
- Systematic sampling → every n^{th} element will be selected

* Types of data



* Scales of measurement →



- Nominal Scale data
  - → Qualitative or categorical data
  - → No order in data.
  - → Gender, color, loc etc.
  - → Charts → pie, bar, ~~orch~~
  - May → Count → show percentage → use pie chart to represent.
- Ordinal Scaled data
  - → Here order or rank has a meaning
  - → diff can not be measured.
  - plots → pie, bar, Jug or count
- Interval Scaled data
  - → The rank and order has a meaning
  - → diff. can be measured (excluding ratio)
  - → it doesnot have zero starting value.
  - plots → Histogram, Scatter plot, ~~Etc~~ line chart.
- Ratio Scaled data
  - → order and rank has a meaning
  - → Difference and ratio are measured
  - → It does have a 0 starting point compulsorily.

PAGE NO.
DATE 25/08/2023

Measure of Central tendency

↳ Descriptive stat
↳ Summarization of data without removing or subtracting any value or instance

** Central tendency tries to represent the center point of a data set.
- → mean
- → median
- → mode

mean → symbolic representation is μ for mean.

$x = \{1, 2, 3, 4, 5\}$

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

↳ Summing up all the values of observation and dividing them with the no. of observation.

median → physical mid point of data

1. Sort the data
2. if count is even → median = avg of two middle value
   if count is odd → median = $\left(\frac{n}{2}\right)^{th}$ value

Note → mean is affected by outliers whereas median is not affected at all.

mode → maximum frequency (the element that repeated highest no. of time)

** with mean, median, mode we are trying to represent the centeral most value in our data.

Note if data is numerical data (continous) → mean, median
if data is categorical data → mode.

Measure of dispersion

before understanding the range we will understand about some topic.

1→ Range → difference between the minimum and the maximum value.

Range ⇒ Max − Min

Note → outlier affect Range that's why we come upto another topic which is Quatrile.

• Percentage → {1, 2, 3, 4, 5}

% of odd No = $\frac{3}{5} \times \frac{20}{100} = 60\%$

60% of the ~~no~~ numbers are odd

* percentage = $\frac{\text{favoured}}{\text{Total}} \times 100$

Percentile → A percentile is a value below which a certain percentage of observation lie.

1, 2, 3, 4, 4, 6, 7, 8, 9, 10

* What is percentile rank of 3.

no below 3 → 2 , Total no = 10

$\frac{2}{10} \times 100 = 20\%$

Here 20 percentile of data lie below 20.

* What value exist at 75th percentile

Value = $\frac{\text{percentile} \times (\text{total value} + 1)}{100}$

Value = $\frac{75}{100} \times 11 = \frac{33}{4}$

Value = 8.25

Here 8th value in data set is value below which 75% of data exist

Quartile → Quartiles are the values that divide a list of numbers into quatres.

- first the no. in order
- Then cut the no. into 4 equal parts
- The quartiles are at the cut, i.e 1st, 2nd, 3rd, 4th

1 | 2 | 3 3 | 4 | 5 | 6 7

1st 2nd 3rd 4th

Q1 → 2
Q2 → 4
Q3 → 6

quartile 1st
quartile 2nd
quartile 3rd

if total no is odd

$Q_1 = \left(\frac{n+1}{4}\right)^{th}$ no.

$Q_3 = 3\left(\frac{n+1}{4}\right)^{th}$ no.

$Q_2 = \left(\frac{n+1}{2}\right)^{th}$ no.

Q_2 ↓ median

if total no. is even

$Q_1 = \frac{n^{th}}{4}$ no

$Q_3 = \frac{3n^{th}}{4}$ no

$Q_2 = \frac{\left(\frac{n^{th}}{2} \text{ no}\right) + \left(\frac{n+1}{2}\right)^{th}}{2}$

Quartile 1	Quartile 2	Quartile 3	Quartile 4

Q_0 Q_1 Q_2 Q_3 Q_4

↳ divide into 2 equal po

- Five point summary

Q_0 (min)	0	→ 0%	→ 0^{th}	percentile
Q_1	1	→ 25%	→ 25^{th}	percentile
Q_2	2	→ 50%	→ 50^{th}	percentile
Q_3	3	→ 75%	→ 75^{th}	percentile
Q_4 (max)	4	→ 100%	→ 100	percentile

Transaction amount

1000 → min = 1000
2000
⋮
$Q_1 = 5000$
↳ this means 25% of transactions amount is equals to or below 5000 in data

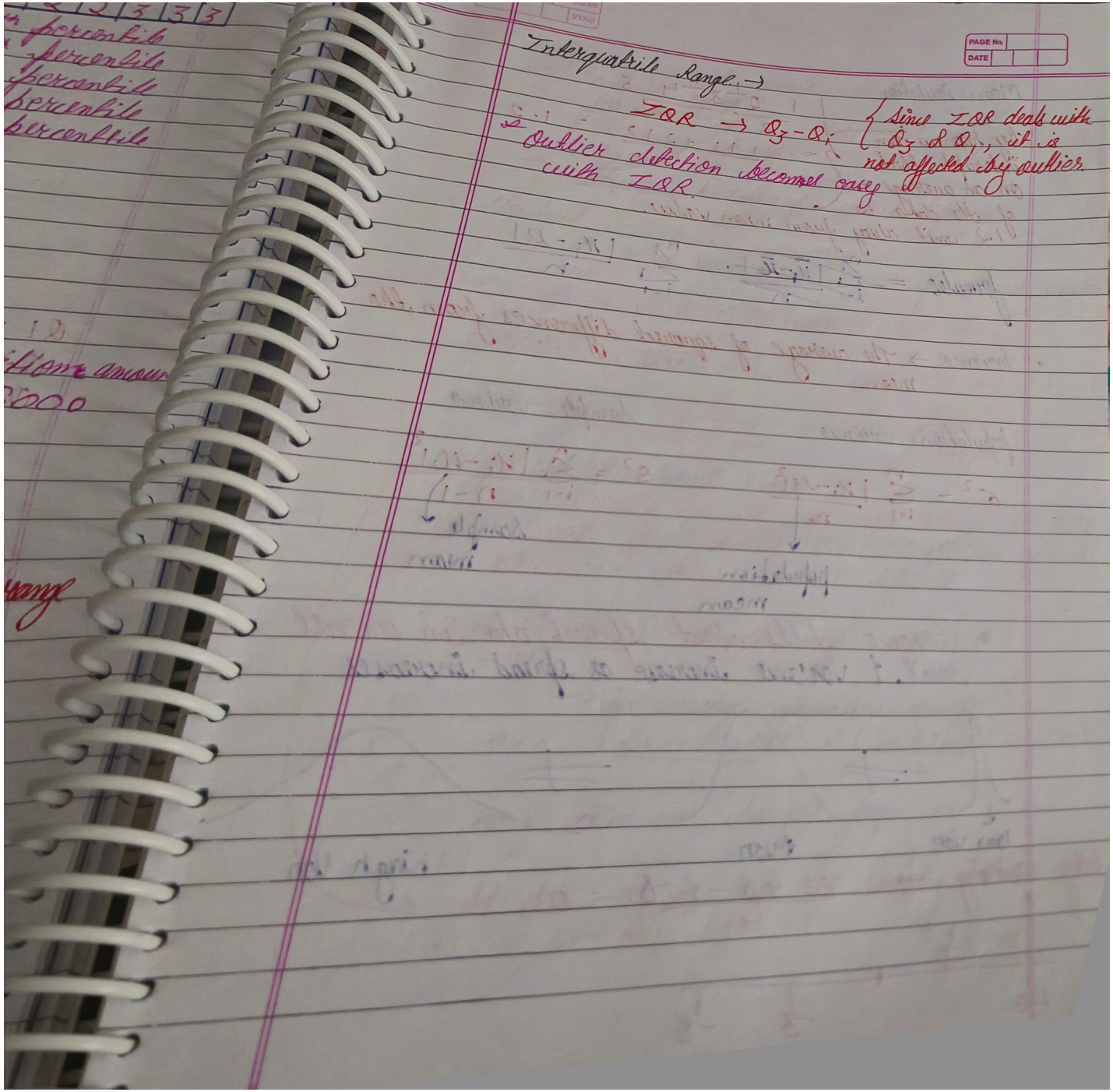
Similarly for $Q_2 = 10000$
$Q_3 = 18000$

- Disadvantages of range → outlier affects the range

Quartile 1	Quartile 2	Quartile 3	Quartile 4

Q_0 Q_1 Q_2 Q_3 Q_4

These are the outlier ranges
min ∈ Q_1 max ∈ Q_4


Interquartile Range →
IQR → Q3 - Q1
{ Since IQR deals with Q3 & Q1, it is not affected by outlier.
∴ Outlier detection becomes easy with IQR.

Mean Deviation

1 2 1/2 3 1/2 4 5

$= \frac{1+1+2+2}{5} = 1.2$

by here we can interept that on an average each of the data is 1.2 unit away from mean value.

formulae $= \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$ $\frac{\sum_{i=1}^{n} |x_i - \mu|}{n}$

- Variance → The average of squared differences from the mean.

population variance

$$\sigma^2 = \sum_{i=1}^{n} \frac{|x_i - \mu|^2}{n}$$

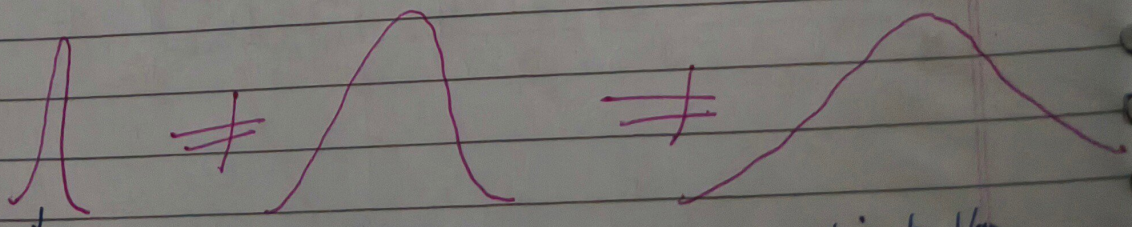
↓ population mean

Sample variance

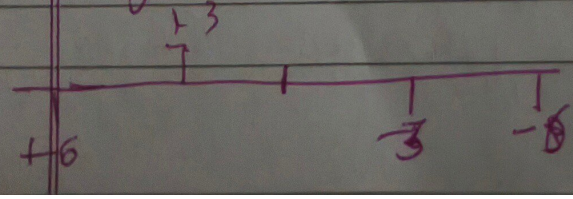
$$s^2 = \sum_{i=1}^{n} \frac{|x_i - x_2|^2}{n-1}$$

↓ Sample mean

* Variance talks about spread at an overall level. ↑ Variance increase as spread increases



Why squaring → $\frac{+3+6-3-6}{4} = 0$



• steps to calculate variance
→ calculate mean
→ for each no. in data, calculate diff b/w the mean and no.
→ squared the difference
→ calculate average of squared difference.

• Standard Deviation
it is a measure of how spread out number are.
↳ square root of variance

$\sigma / S = \sqrt{var}$

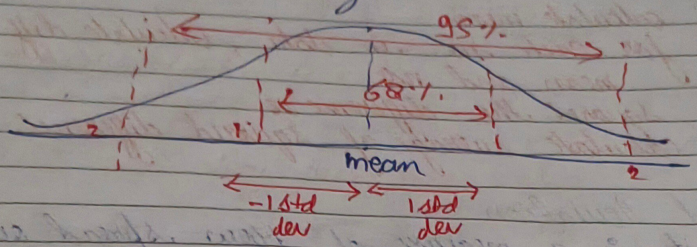
$\sigma = \sqrt{\sigma^2}$ population. $S = \sqrt{S^2}$ sample.

why std. deviation?
1. Variance can be huge no. because it talks about spread of an overall level. Comparision of each no. w.r.t variance becomes difficult

2. $\frac{\sum_{i=1}^{n} |x_i - \mu|^2}{n}$ → dimensions are squared
↳ in variance dimension changes
eg → $\frac{|dis - Avg\ dis|^2}{2} = \frac{|3m - 2m|^2}{2} = m^2$
Here dimension changed from m to m^2

→ std dev. → std way of knowing where your data lies.

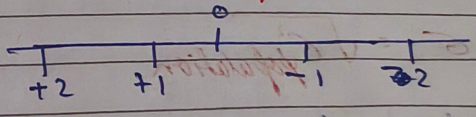
if data is normalize



Why we using squared diff in variance

i→ it negtate the values if not sq.

eg → $\mu = 0$

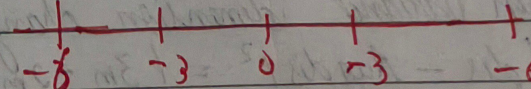


$= \frac{+2+1-1-2}{4} = \frac{0}{4} = 0$

→ Here is the spread = 0, nd.
if it is zero then all values = mean

ii→ if we make absolute instead of sq.

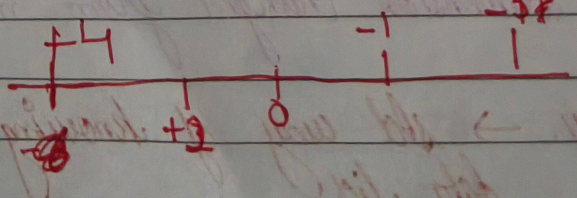
SC-1



$= \frac{3+3+6+6}{4}$

$= 4.5$

SC-2



$= \frac{1+2+4+5}{4}$

$= 4.5$

* Here if we see we get same value 4.5

But is both dataset have same variance?
→ NO

→ That's why we use squared difference

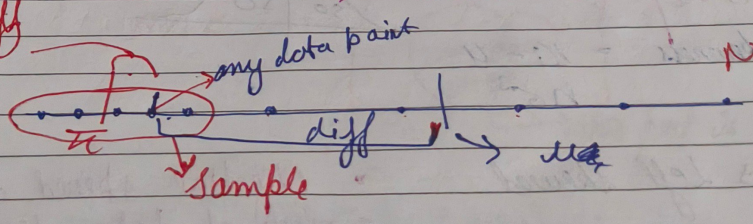
Why in population variance it divides by N but in sample variance it ~~is~~ is n-1

When we calculate any sample stat.
↳ then we do not have access to complete population

we use this because sample variance will be unbiased estimator

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$ → Bessel Correction

near diff

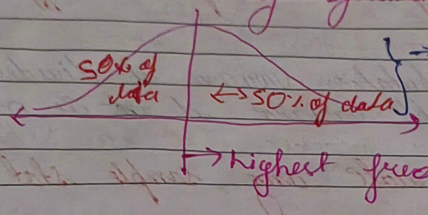


$$(x_i - \bar{x})^2 < \frac{\sum |x_i - \mu|^2}{n}$$

So to get ans near to μ we reduce n = n+1 as it is denominator
(denominator ↓, ans ↑)

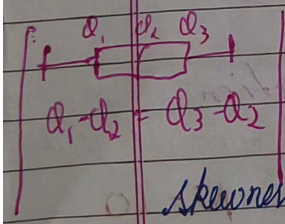
* Measure of symmetricity → talks about shape of data

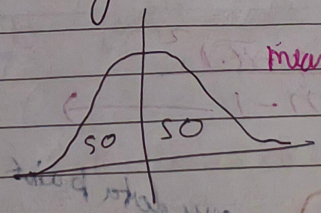
skewness → measure of symmetricity



Skewness → inclined to some side

1 → No skewness ⇒ { symmetric }
{ skewness = 0 }

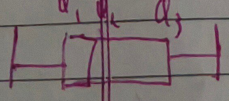




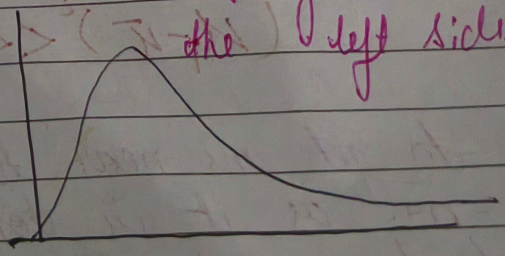
mean = median = mode

Skewness = $\frac{\bar{x}_i - \mu}{n\sigma^3}$

2 → Left skewed



$Q_3 - Q_2 > Q_2 - Q_1$

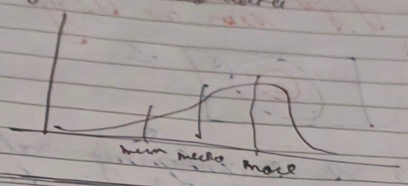


• Negative skewed data
• most of data lies at the left side.

3 → Right skewed

{ mean > median > mode

3 → Right skewed data



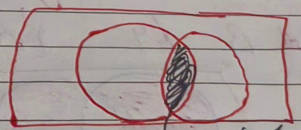
mean < median < mode

Set → collection of unordered unique elements)

Properties of set

1 → intersection → common element

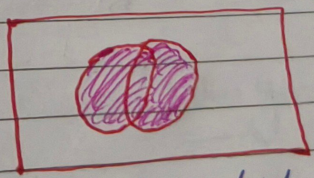
$A \cap B$ =



→ This part is intersection

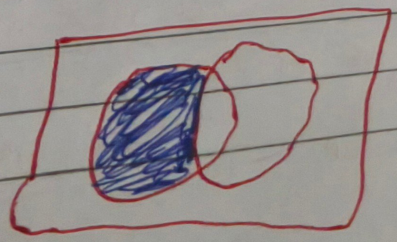
2 → Union → all distinct elements from both sample

$A \cup B$ =

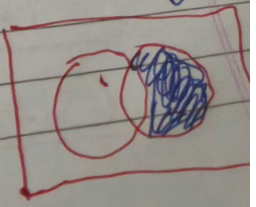


3 → difference → elements which are present only in

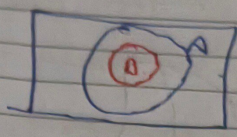
$A - B$ =



$B - A$ =



4→ Subset → if all elements of B present in A then we say B is subset of A or B⊂A

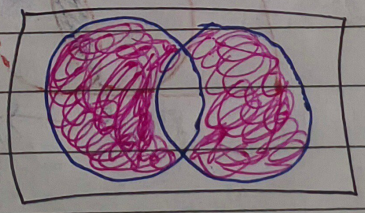


5→ Superset → A is containing all element of B, then A is superset of B

6→ Symmetric Diff → opp of intersection

* the element of that are distinct in both excluding of intersection

A Δ B or {A∪B − A∩B}

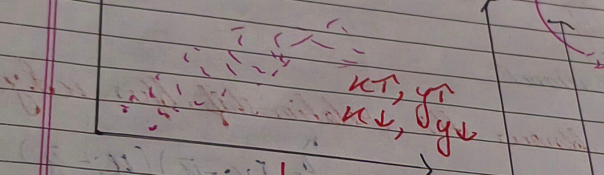


Covariance And Corelation

Here we say as x increase y also increases

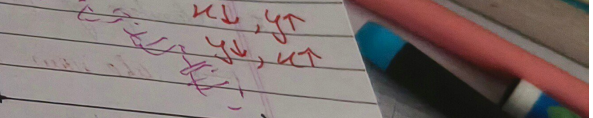
x Total transaction count	y Total of transaction amount
5	10K
4	8K
3	3K

* understanding the relationship



↓ direct relationship

ex → price of house based on area / locality



↓ indirect relationship

ex → No. of years and alcohol consumption

* Quantifying / measuring relationship

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

x, y are features

Co-variance → spread

exist regative

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

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

So i can say that variance is the relationship of feature with itself.

So we can say it is telling relation ship with itself.

Similarly Covariance is nothing but we are trying to calculate relationship of any feature with ~~at~~ respect of other.

Advantage of Covariance

1 → We now know the relationship b/w x & y.

disadvantage $\left[Cov(x,y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1} \right.$

x	y
Cov(x,y) = 50	

A	B
Cov(A,B) = 100.	

Can we say $Cov(x,y) = \frac{1}{2} Cov(A,B)$?

→ No.

Why ⇒ Cov(x,y) → range → $(-\infty, +\infty)$

The we can not compare strength as both may have different starting point. or '0'.

2 → No comparision of strength of relation, No any standerlised scale to interept the strength.

2-> Covariance has dimension

$$Cov(x,y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y - \bar{y})}{n-1} \quad \text{Rs} \cdot \text{ft}$$

Cov(transaction, height) = Rs.ft = 450 Rs.ft

↳ we can not compare two different dimensions ⇒

Correlation

Pearson ↑ Coefficient [-1 to 1]

$$f(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y} = [-1 \text{ to } 1]$$

↳ it solve dimension problem as it cancelled and become dimensionless

-1 ←——— 0 ———→ 1

less Correlated. The more positively correlated features are

Note -> • Pearson Correlation Coefficient always measures the linear relationship.
• Correlation is never a slope
• Non linear data have pearson Correlation = 0

for non linear data relation ship we have

⇒ Spearman Rank correlation

$$\rho_s = \frac{cov(R(x), R(y))}{\sigma(R(x)) \cdot \sigma(R(y))}$$

R(x) → Rank of x
R(y) → " " y

Rank →
1 → Sort value
2 → highest will 1st then it will increase
3 → use the Rank in position of data points.