Statistics Basics

Mathematical science including methods of collecting, organising & analysing data in such a way that meaningfull conclusions can be drawn from it.

Data ⇒ facts or pieces of info. that can be stored, measured & re-accessed.

Ultimate goal (in case of business) To make revenue to grow, from data insights.

Collecting data ⇒ manual, survey, internet clicks, ratings.

Organising data ⇒ SQL & NoSQL DB.

Analysing data ⇒ Python libraries (Pandas, numpy, matplotlib, pyplot, bokeh).

Ex., for data

1) Heights of students in class 12 'A' section.
2) No. of units a product has sold ₹ in each month over the year.
3) No. of products units produced by the company over several years.

Types of statistics :-

1) Descriptive statistics
   ↳ It consists of organising & summarising the complete data (Population).
   Ex; Avg. weight of students in classroom, strike rate of M.S. Dhoni.

2) Inferential statistics

You can't count all the trees present in all of Indian forests or calc. avg. weight of people, by taking avg. of all 150cr+ weights.

It uses sample data measured to form conclusion of population.

Sample is a subset of population.

Ex., calculating avg. weight of entire people of a country.

Techniques of Descriptive statistics

1) Central Tendency (Mean, Median, Mode)
2) symmetricity (skewness, Kurtosis)
3) Dispersion (std. Deviation, Variance)

Types of Sampling (Inferential statistics)

1) Simple Random Sampling
2) Stratified sampling
3) cluster sampling
4) systematic sample

} Most commonly used

→ used mostly in research studies

Sampling Types

INFERENTIAL STATISTICS

1) Simple Random Sampling :-

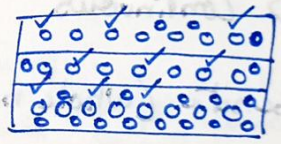
→ Every member of Population (N) has equal chances of getting selected.

Ex:- If N = 1000, The prob. of each & every member getting selected is 1/1000.

Disadvantage : If a sample is collected from all the states of India, the prob. of people getting selected from less populated states is less to none.

So, Not Equal Representation.

2) Stratified Sampling :-



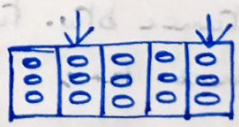
Strata → layers / groups

Diff: distinct categories are seperated.

A Simple Random sample would de chose from each category (strata / layer / group).

Adv. : Equal Representation.

3) Cluster Sampling :-



clustering the population or grouping the population into clusters & then the sample is obtained / formed from elements of each cluster randomly.

Ex:- States with similar population are made into one cluster & sample is formed by random selection of states from the cluster.

4) Systematic Sampling

→ Every n^{th} element will be selected.

Ex:, selection based on
- odd roll no. (3, 5, 7, ...)
- odd year of birth (2001, 2003, 2005, ...)

Types of Data

Quantitative (Numerical) (Measurable)
- Discrete
- Continuous

Discrete (whole no.)
No. of children playing in park. Ex.,

Continuous (range) (Real no.)
distance btw. two places. Ex.,

Qualitative (categorical)
- Nominal
- Ordinal

Nominal (No order is there, naturally)
Ex., gender, blood group.

Ordinal (Order is there)
Ex., grades, ranks.

Sometimes, conversion from categorical data to Numerical data.

Scales of Measurement

i) Nominal scale
ii) ordinal scale
} Qualitative

iii) Interval scale
iv) Ratio scale
} Quantitative

i) Nominal scale data

→ Basic counting using numbering or percentage.
→ Ex:, No. of females in Room A - 350. Pie, Bar Plot
No. of Males in Room A - 209.

ii) Ordinal scale data

→ order & rank has a meaning.
→ Difference can't be measured. (1st rank - 2nd rank ⇒ not meaningful)
→ Pie chart, Bar plot.

iii) Interval scale data

→ order & rank has meaning
→ Diff. can be measured (Excluding Ratios)
→ "0" need not be the starting value
Ex: Length, weight, height.
Histogram, Scatter plot, Line graph.

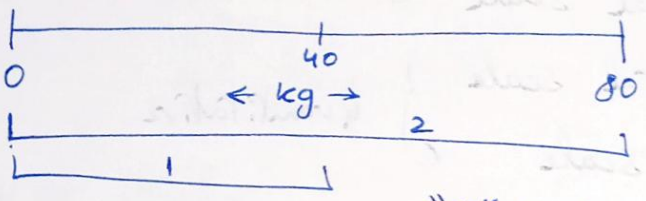
iv) Ratio scale data

→ order & rank has meaning.
→ Differences & ratio are measurable.
→ starting pt. is zero.

Ex., ~~Temp~~. weight

80 kg B

40 kg A



$\frac{B}{A} = \frac{80}{40} = \frac{2}{1}$

B:A = 2:1

"0" is present

B is twice has heavy as A.

DATA	Nominal	ordinal	Interval	Ratio
Labelled	✓	✓	✓	✓
Meaningful order	X	✓	✓	✓
Measurable difference	X	X	✓	✓
Zero as starting point	X	X	X	✓
Example	Gender, Religion, Location	Rating, Grade, rank	IQ, Temp, Height, Score, weight	height, weight, length.

I) Measures of Central Tendency

DESCRIPTIVE STATISTICS

"Central" ⇒ what is the one value around which all data is revolving?

Mean, Median, mode → EDA, Data preparation, feature engineering.

① Mean (Avg. or Arithmatic Mid value)

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

$\sum$ ⇒ Summation

Summing up all observations.

No. of observations.

② Median (Physical Mid. point of data)

→ sort the data points
→ Divide the No. of data points by 2
→ The output no. is the position of mid value.

The above is ok for odd no. of values.

Even no. of observations : Case

P = { 1 2 3 4 }

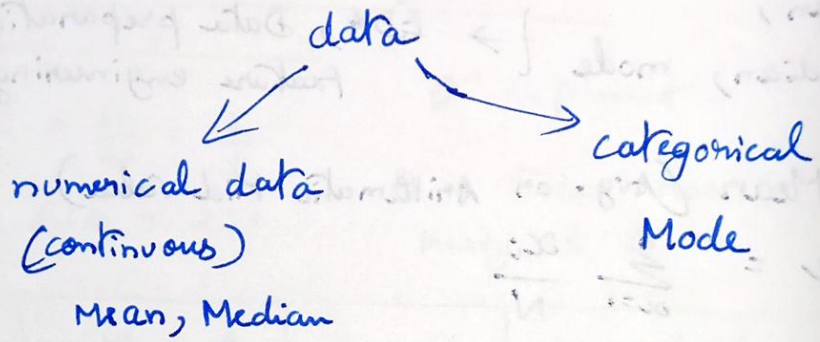
$\frac{2+3}{2} = 2.5$ (median).

Median is not affected by outliers whereas Mean is affected by outliers.

(3) Mode (Highest occuring data point)

1, 2, 9, 10, 10, 8, 7, 7, 8, 8

Here ⇒ 8 is the mode.



USE CASE OF CENTRAL TENDENCY
(Missing Value Imputation)

AGE	GENDER	WEIGHT(kg)
25	M	80
26	M	70
24.75	M	30
23	M	75
25	F	1000

AGE is continuous variable

↳ Impute / Replace the missing / null value with mean if there is no outlier

$$\frac{25 + 26 + 23 + 25}{4} = 24.75$$

GENDER is categorical variable

↳ Impute with Mode

M, M, M, F

Mode = M

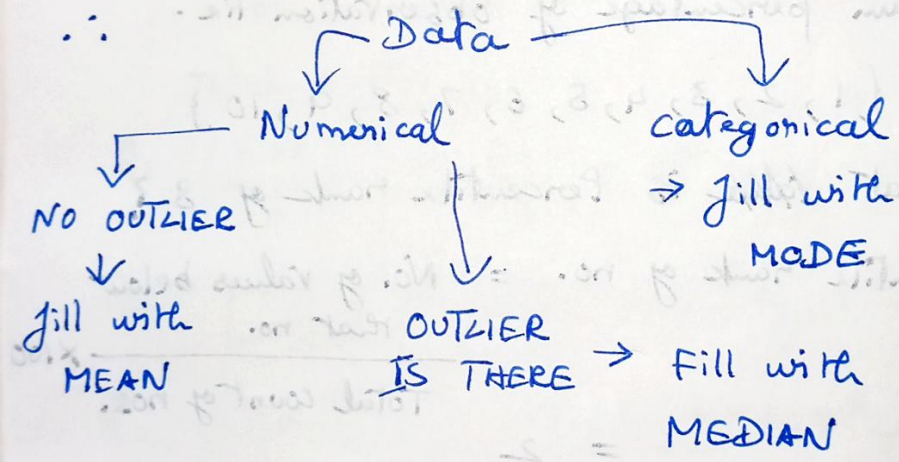
WEIGHT is continuous variable

↳ Impute with Median as outlier (1000) is present

30, 70, 80, 1000

$\frac{70+80}{2} = 75$

∴



II) Measures of Dispersion

How the data is spread out?

1) **Range** → Diff. between max. & min. value.

Ex: {1, 2, 3, 4}

Range = 4 − 1 = 3

Ex: {1, 2, 3, 1000}

Range = 1000 − 1 = 999

★ OUTLIER AFFECTS RANGE ★

2) Percentage

Ex:, 1, 2, 3, 4, 5

what is % of no. of odd numbers?

$3/5 \times 100 = 60\%$

60% of the numbers in the given data is odd.

3) Percentile

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

Ex:, {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}

i) what ~~value~~ is Percentile rank of 3?

$$\text{Percentile rank of no.} = \frac{\text{No. of values below that no.}}{\text{Total count of nos.}} \times 100$$

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

$$= 20^{th} \text{ Percentile.}$$

ii) what value exists at 75th Percentile?

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

$$= \frac{75}{100} \times (10+1)$$

$$= \frac{3}{4} \times 11 = 8.25$$

8

4) Quantiles

Quantiles are values that divides a list of nos. into quarters (4 divisions).

Ex.,1. {5, 9, 13, 14, 21, 22, 39}

$Q_1 = 9$ $\quad Q_2 = 14$ $\quad Q_3 = 22$

Ex.,2. {1, 1, 1, 1, 2, 2, 2, 3, 3, 3, 4}

Case 1 : odd count of nos.

Total = 11

$$Q_1 = \left(\frac{n+1}{4}\right)^{th} = \frac{11+1}{4} = 3^{rd} \text{ no.} = \underline{1}$$

$$Q_2 = \left(\frac{n+1}{2}\right)^{th} = \frac{11+1}{2} = 6^{th} \text{ no.} = 2$$

$$Q_3 = \left(\frac{3(n+1)}{4}\right)^{th} = \frac{3 \times (11+1)}{4} = 9^{th} \text{ no.} = 3$$

Ex.,3. {25, 103, 921, 2517, 3, 19, 31}

case 2 : even count of nos.

Total = 6

$$Q_1 = (n/4)^{th} = 6/4 = 1.5^{th} = \frac{25+103}{2} = 64 \quad (Q_1)$$

$$Q_2 = \frac{(\frac{n}{2})^{th} + (\frac{n}{2}+1)^{th}}{2} = \frac{\frac{6}{2} + (\frac{6}{2}+1)}{2} = \frac{3^{rd} + 4^{th}}{2}$$

$$= \frac{921 + 2517}{2} = 1719 : Q_2$$

$Q_3 = \left(\frac{3n}{4}\right)^{th} = \frac{3\times 6}{4} = \frac{18}{4} = 4.5^{th}$

$\Downarrow$

avg. of 4^{th} & 5^{th} number

$\neq$ ~~3/4~~ $\neq$ ~~3/8~~

$= \frac{2517 + 319}{2}$

$= 1418$ //.

Five pt. Summary

$Q_0 \Rightarrow$ min. no. $\Rightarrow$ (1) $\Rightarrow 0^{th}$ percentile

$Q_1 \Rightarrow 25^{th}$ %ile $\Rightarrow$ (2.5)

$Q_2 \Rightarrow$ median $\Rightarrow 50^{th}$ %ile $\Rightarrow$ (5.5)

$Q_3 \Rightarrow 75^{th}$ %ile $\Rightarrow$ (7.5)

$Q_4 \Rightarrow$ max. no. (10) $\Rightarrow 100^{th}$ %ile

Ex., data, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

Q_0 Q_1 Q_2 Q_3

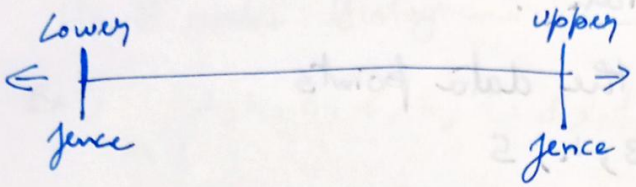
5) Inter Quartile Range

$IQR = Q_3 - Q_1$

Not affected by range & outliers

Main usage of IQR is to find out outliers.

Outlier is extreme values



$$\text{Lowerfence} = Q_1 - 1.5 \times IQR$$

$$\text{Upperfence} = Q_3 + 1.5 \times IQR$$

If any value is present & is greater than the upper fence or lesser than the Lower fence, it is an OUTLIER.

For the previous example,

$$\text{Lowerfence} = 2.5 - 1.5 \times IQR$$

$$(IQR = Q_3 - Q_1 = 7.5 - 2.5 = 5)$$

$$= 2.5 - 1.5 \times 5$$

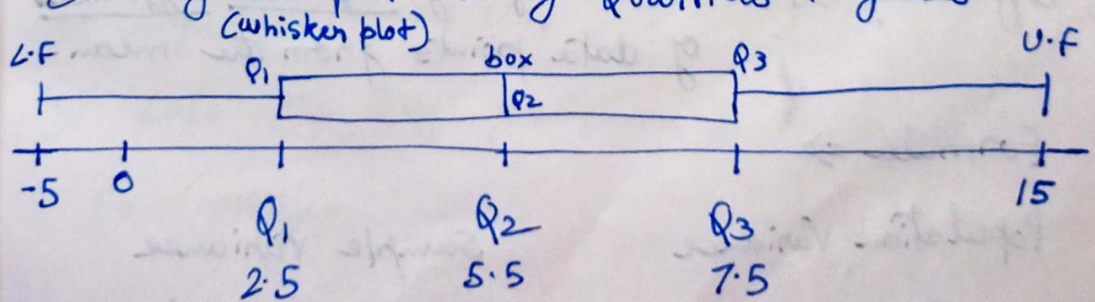
$$\text{Lowerfence} = -5.$$

$$\text{Upperfence} = 7.5 + 1.5 \times 5$$

$$= 7.5 + 7.5$$

$$\text{Upperfence} = 15.$$

Drawing box plot using Quartiles & fences....



* NOT TO SCALE

L.F => Lower fence/whisker

UF => Upper Fence/whisker

Main Concepts of Measures of Dispersion

6) Mean Deviation :-

Ex., Let's take the data points

1, 2, 3, 4, 5

Here, the mean is 3.

1 is 2 points away from 3
2 is 1 point away from 3
4 is 1 point away from 3
5 is 2 point away from 3

$$\therefore \quad \frac{2+1+0+1+2}{5}$$

$$= 6/5 = 1.2 \Rightarrow \text{Mean Deviation}$$

Formula $\Rightarrow$ $\dfrac{\sum_{i=1}^{n} |x_i - \mu|}{N}$

Definition $\Rightarrow$ The average of difference of each data point from the mean.

7) Variance :-

Definition $\Rightarrow$ The average of squared differences of data points from the mean.

Formula $\Rightarrow$

Population Variance

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

Sample variance

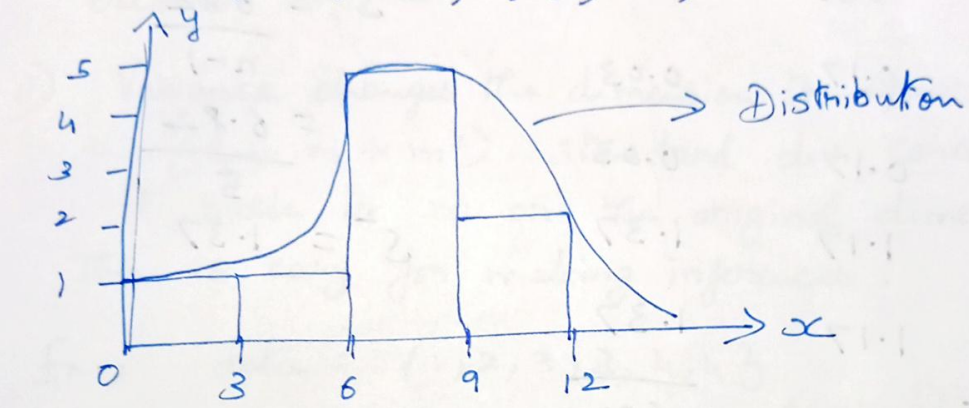
$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{N-1}$$

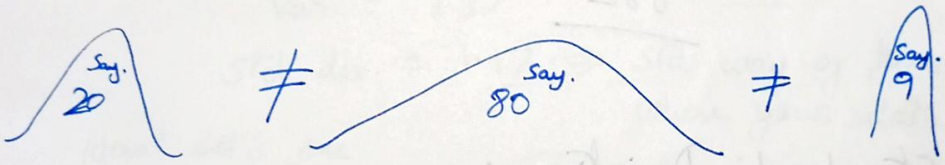
where, μ => Population mean

$\bar{x}$ => Sample mean

How to make Histogram ?

Ex:, 2, 4, 6, 6, 6, 8, 8, 10, 11





Variance represents the spread.

Spread ↑, Variance ↑

How To Calc. Variance ?

1) Calculate Mean
2) For each no. in data, Subtract the mean & the no.
3) Square the (difference)
4) Calc. the avg. of Square of difference.

Ex:, problem.

data = {1, 2, 3, 3, 4, 4} $\bar{x}$ = 17/6 = 2.83

x	$x - \bar{x}$	$(x - \bar{x})^2$
1	-1.83	3.34
2	-0.83	0.68
3	0.17	0.03
3	0.17	0.03
4	1.17	1.37
4	1.17	1.37
17		6.82

Sample Formula

$$S^2 = \frac{6.82}{n-1}$$

$$= \frac{6.82}{5}$$

$$S^2 = 1.37$$

8) Standard Deviation :-

S.D is a measure of how spread out numbers are.

↳ square root of Variance.

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

from previous example,

$$S = \sqrt{1.37} = 1.17.$$

formula :

Population

$$\sigma = \sqrt{\text{var}_p}$$

sample

$$S = \sqrt{\text{var}_s}$$

Why Std. Dev.?

i) Variance can be a huge no. because it talks about spread at an overall level. Comparison of each data point w.r.T to variance becomes difficult.

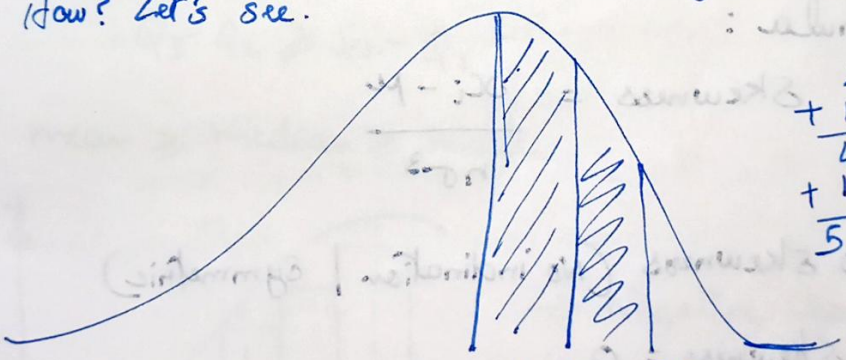
ii) Variance changes the dimension to square (Ex., $m \rightarrow m^2$). Standard dev. converts it back to m on the original dimension. This is easy for making inferences.

Ex., data = {1, 2, 3, 3, 4, 4}

Var. = 1.37

Std. dev. = 1.17 ⇒ Std. way of knowing where your data lies.

How? Let's see.



$\bar{x}$

2.83 + 1.17 = 4.00

4.00 + 1.17 = 5.17

Impirical formula :-

1 std. dev. away from mean

2 std. dev. away from mean

68.1. of data - within 1 std. dev.

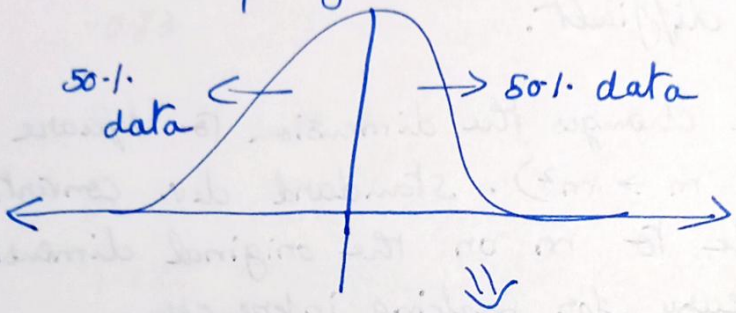
95.1. of data - within 2 std. dev.

99.7.1 of data ⇒ within 3 std. dev.

III) Measures of Symmetricity

↓

Shape of data



⇓

Symmetric

1) Skewness :-

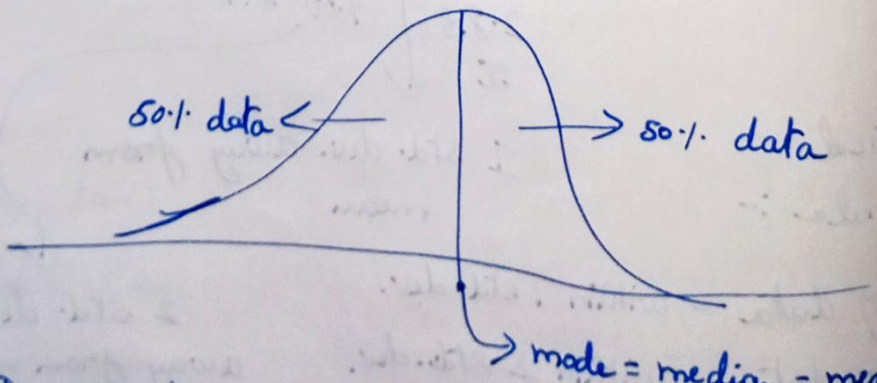
Inclined to some side

formula :

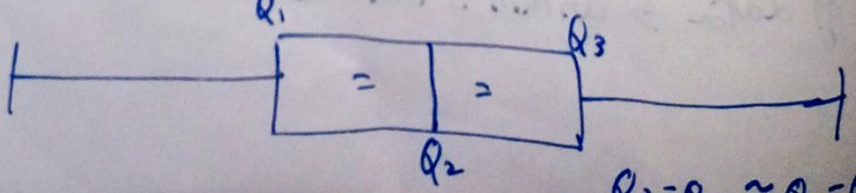
$$\text{Skewness} = \frac{x_i - \mu}{n\sigma^3}$$

i) No skewness (No inclination / symmetric)

Skewness = 0

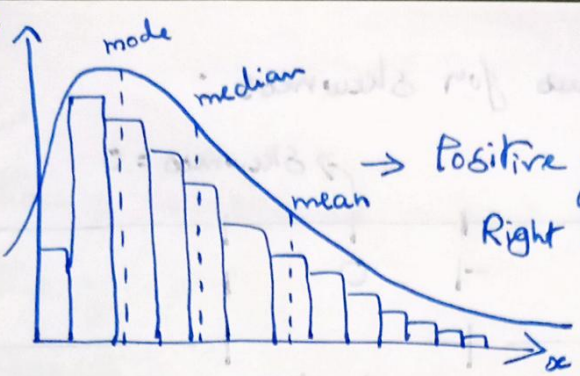


(Box plot)



$Q_3 - Q_2 \approx Q_2 - Q_1$

ii)

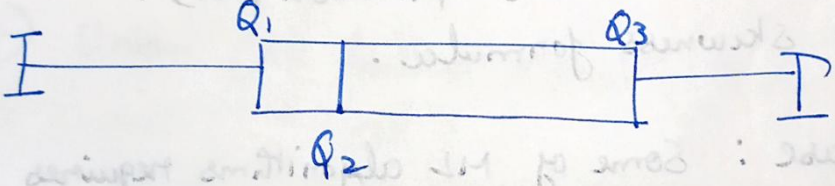


→ Positive skewed data
(or)
Right skewed data

→ Tail is on right side of distribution

→ Most of data is in the right side of the distribution.

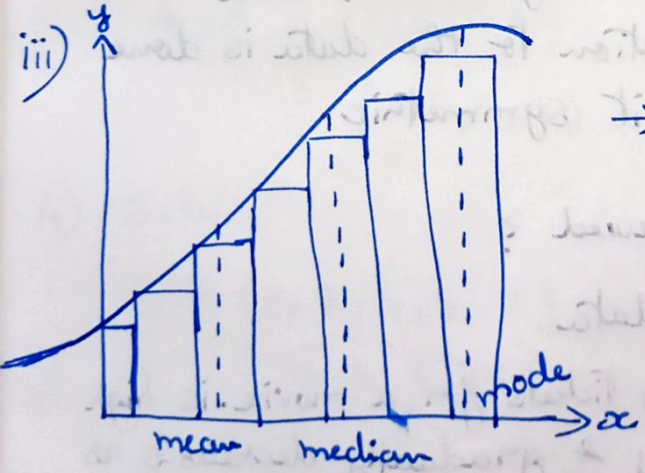
(Box plot)



$Q_3 - Q_2 \geq Q_2 - Q_1$

mean ≥ median ≥ mode

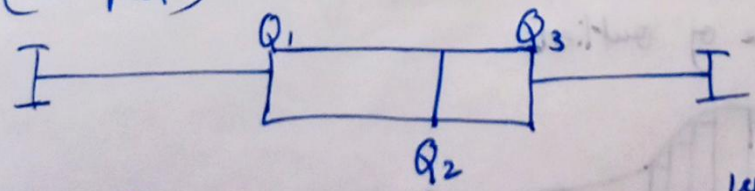
iii)



→ Negative skewed data
(or)
Left skewed data

mode ≥ median ≥ mean

(box plot)

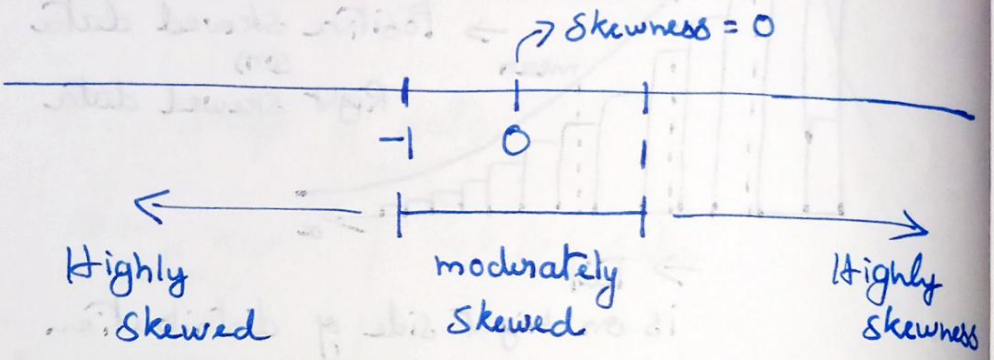


$Q_2 - Q_1 \geq Q_3 - Q_2$

Tail on the left side of distrib.

Most of data on left side of distribution.

Possible Values for skewness :



To know if data is skewed or not ?

→ Visualisation (distribution plot).

→ skewness formula.

Use Case : Some of ML algorithms requires Symmetric data to build the ML model.

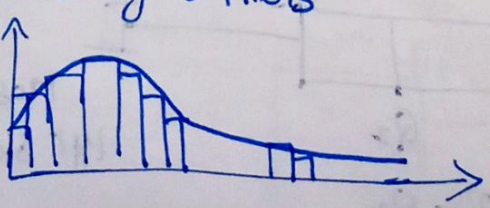
If the data is asymmetric, some transformation to the data is done to make it symmetric.

Why data is skewed ?

i) Nature of data

Ex., Ex., sale of tickets for a movie. is high initially & gradually decreases as days pass on.

ii) Presence of outliers



SET

Collection of un-ordered unique elements.

Properties of Set :

1) Intersection (common elements)

$A = \{1, 5, 7, 9, 13\}$

$B = \{13, 15, 17\}$.

$A \cap B = \{13\}$

2) Union (All distinct elements from both sets)

$A \cup B = \{1, 5, 7, 9, 13, 15, 17\}$

3) Difference

$A - B = \{1, 5, 7, 9, 13\} - \{13, 15, 17\}$

$= \{1, 5, 7, 9\}$.

4) Subset

$C = \{1, 2, 4, 5, 7\}$

$D = \{2, 4, 5\}$

D is subset of C as all the elements of D is present in C.

5) Superset

From the C & D example, C is the superset of D as C contains all the elements of D.

6) Symmetric Difference (opposite of Intersection)

$E = \{1, 20, 200, 4000\}$.

$F = \{20, 40, 80, 100, 200\}$

$E \Delta F = \{1, 200, 4000, 40, 80, 100\}$.

All the elements of both sets except the common ones.

(Covariance & Correlation)

Understanding the relationships

i) Direct Relationship:

If one variable ↑, another variable ↑, then they are in a direct relationship.

Ex.; Predict price of house based on area of house.

If the area of house increases, price of house also increases.

ii) Indirect relationship:

If one variable ↑, another variable ↓, or vice versa, then they are in indirect relationship.

Ex.; As you drive longer, the distance covered will be increasing but the fuel level in the vehicle will be decreasing.

To Quantify / Measure The Relationships :-

① Covariance

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

where, x & y are two features.

Covariance ⇒ relationship b/w two variables

$x \uparrow y \downarrow$ (or) $x \uparrow y \uparrow$

$x \downarrow y \uparrow$ $x \downarrow y \downarrow$

-ve covariance +ve covariance

Ex:-

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$
2	3	-1	-2.5	2.5
3	5	0	-0.5	0
6	6	3	0.5	1.5
1	8	-2	2.5	-5
				+
				-1

$\bar{x} = 3$

$\bar{y} = 5.5$

$$cov(x,y) = \frac{-1}{4-1}$$

$$= -1/3 = -0.33$$

Interpretation ⇒ x & y are negatively related.

Disadvantage of Covariance

→ The range of Covariance is $-\infty$ to ∞.

→ This means the comparison of strength of relationship is not possible.

→ There is no standardised scale to interpret the strength.

→ Covariance has dimension.

Ex., Cov(Price, height) = Rs.ft = 450 Rs.ft.

Cov(height, ~~Height~~ weight) = ft.kg. = 600 ft.kg.

We cannot say that Cov(ht., wt.) is larger than Cov(Price, ht.) as both are diff. units.

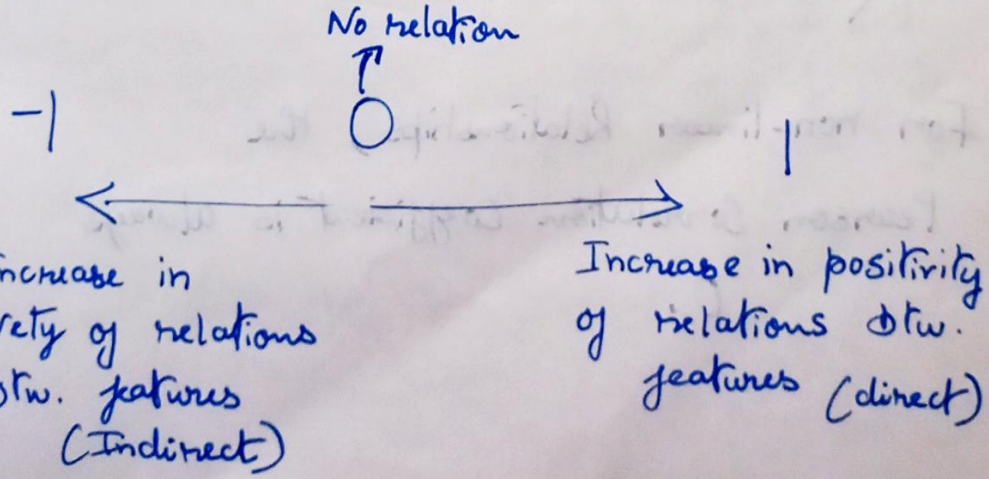
(The above example is for height, weight & Price of Tiles (floor Tiles))

To overcome these disadvantages, we use, ∴

(2) Pearson Correlation Coefficient (-1 to 1)

$$\rho_{(x,y)} = \frac{Cov(x,y)}{\sigma_x \cdot \sigma_y}$$

Diamensionless quantity as the diamensions are cancelled out by the standard deviations of the factors / features.



$\int_{(x,y)} = 0.4$ $\int_{(a,b)} = 0.8$

Features a,b is highly +vely correlated as compared to Features x,y.

$\int_{(c,d)} = -0.2$ $\int_{(o,p)} = -0.5$

Features o,p are more -vely correlated as compared to Features c,d.

IMP :: Pearson correlation coefficient always measures the linear relationships.

Correlation is not dependent on the angle of the slope.

/ $\Rightarrow 1$ / $\Rightarrow 1$ / $= 1$

— $\Rightarrow 0$

\ $\Rightarrow -1$ \ $\Rightarrow -1$ \ $\Rightarrow -1$

For non-linear Relationships, the Pearson Correlation Coefficient is always 0.

To understand non-linear relationships, we use,

(3) Spearman Rank Correlation

$$\gamma_s = \frac{Cov(R(x), R(y))}{\sigma(R(x)) \cdot \sigma(R(y))}$$

where, $R(x)$ ⇒ Rank of x.

$R(y)$ ⇒ Rank of y.

Ex:,

x →	5	7	8	1	2
y →	6	4	3	1	2
$R(x)$ →	3	2	1	5	4
$R(y)$ →	1	2	3	5	4

$x \rightarrow 5, 7, 8, 1, 2$ ⇒ Sort the values in descending order

8, 7, 5, 2, 1 → 5th

1st ↙ 2nd ↙ 3rd ↓ 4th ↓

$y \rightarrow 6, 4, 3, 1, 2$

6, 4, 3, 2, 1

1st ↙ 2nd ↙ 3rd ↓ 4th ↓ 5th ↓

Then, we use the formula to find out the correlation.

Use case for correlation => used for feature selection in ML modelling.

Ex.; Let's take factors

$x_1, x_2, x_3, x_4, x_5, \ldots\ldots, x_{1000}$

&

y (price)

Correlation is used to find which factors/features from x_1 to x_{1000} correlate well with feature y (price) in-order to use it for Machine learning modelling & predictive modelling.