Statistics

① Statistics is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data.

Types of Statistics
① Descriptive stats
② Inferential stats

Data = "facts or pieces of information"
eg = Age = 20, height = 5.6ft

the data can be in
population
sample

Descriptive stats

① Analyzing data, summarizing data. Organizing data in the form of Number & Graph.

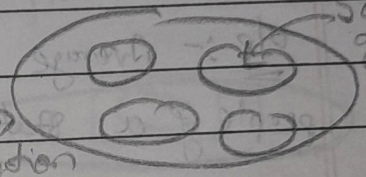
② BAR plot, Histogram, Pie-chart, PDF, CDF, Normal Distribution.

③ Measure of Central Tendency. Mean, median, Mode.

④ Measure of variance
SD, Variance

eg = There are 20 classes at a university and you have collected the ages of students in one class
Ages = {20, 21, 18, 35, 26, 24, 22, 21}
Q what is the most common age of student in your class?
Ans ⇒ {21} ← mode

Inferential stats



From the entire population we are just taking a sample and we are inferencing some information.

eg
① • Z-Test ⇒ Hypothesis testing
• T-Test
• Chi Square Test

Definition:
Consists of using data you have measured to form conclusions / inferences

* Population Vs Sample

* Statics

(1) Population → Population Mean] measures of central tendency
(2) Sample → Sample Mean] other are (Mean, Median, Mode)
(3) Random variable

Discrete Random variable

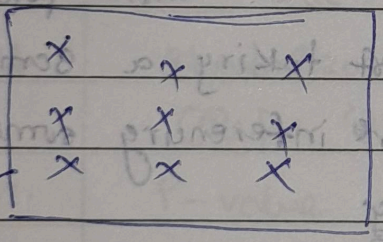
Continuous Random variable

* Population

O/P :- Average Height of all the people of a State → Karnataka

⇓ considering
≈ 1 million

μ = mean



Population

⇒ μ = Average height of all the people

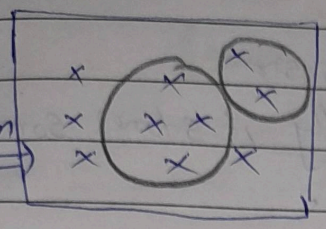
population $\mu = \frac{1}{1M} \sum_{i=1}^{1M} x_i$

Note :- When we have large population, it is very difficult to find Each and Every people height. So, we are using Sample

Population count => N | population mean = μ
Sample count = n | sample mean = $\bar{x}$

Date :
Page :

Sample



=> Here we will take sample of state population

taking 10,000 population

Sample mean $\bar{x} = \frac{1}{10,000} \sum_{i=1}^{10,000} h_i$ → height of all people

Note:- The population are v.v. Large values.
so → we try to take sample from population.

Eg:- EXIT Polls (Election)

Population (N)	&	Sample (n)
population mean $\mu = \frac{1}{N} \sum_{i=1}^{N} h_i$		sample mean $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} h_i$

population mean > Sample mean

Note:- when the sample size increases Sample mean ≈ population mean (equal)

* Gaussian Distribution or Normal Distribution

1) Random Variable
   - → Discrete Random Variable
   - → Continuous Random Variable

2) Gaussian / Normal Distribution

Random Variable

Variable => int, float, str

Variable: a = 10 (Basically it stores some value in it.)
x = 10.2
x = "10.2"

→ Discrete => It will be whole number
It will not be floating number
eg:- No. of Bank Account a person has
→ 2, 3, 4, 5,
eg:- population of a state 1 million

→ Continous => within a range of values we can have any value
eg:- 10 ---to--- 15
it can be 10, 10.2, 10.3, 14.9, 14.8, 15
It can be whole number, Decimal.
eg:- height = 5.11 inch, 5.6 inch
eg:- salary, loan amt

Smoothening the histogram

(2) Gaussion Distribution / Normal Distribution

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

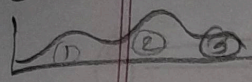
(3 sigma limit)

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

∴

SD → $\sigma = \sqrt{\text{Var}}$

• Gaussian Distribution Model MU = μ sigma = σ



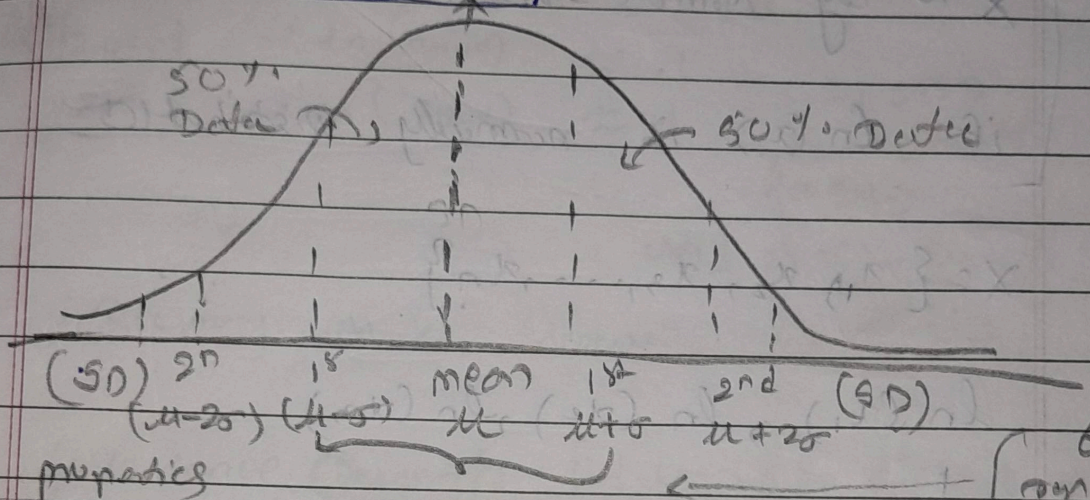
Date :
Page :

Random variable mean SD

$x \sim G.D\ (\mu, \sigma)$

specifically continuous variable → belongs → with same value SD or Variance

Bell curve



Three properties

Empirical formula in Gaussian Dist

probability

① $Pr(\mu - \sigma \leq x \leq \mu + \sigma) \approx 68\%$

68% of random variable will be falling in this first standard Deviation

x which is a part of → X (random variable)

The total of all elements from the X present between 1st SD to the left and 1st SD to the right

② $Pr(\mu - 2\sigma \leq x \leq \mu + 2\sigma) \approx 95\%$

③ $Pr(\mu - 3\sigma \leq x \leq \mu + 3\sigma) \approx 99.7\%$

(3) * Log Normal Distribution (Refer Live Class)

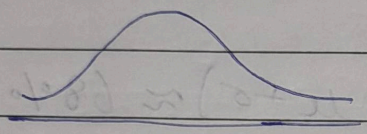
$X \sim$ Log Normal Distribution

if $\ln(x)$ is normally Distributed

$X = \{x_1, x_2, x_3, \ldots, x_n\}$

$\ln(x_1), \ln(x_2), \ln(x_3)$

↳ normally Distributed



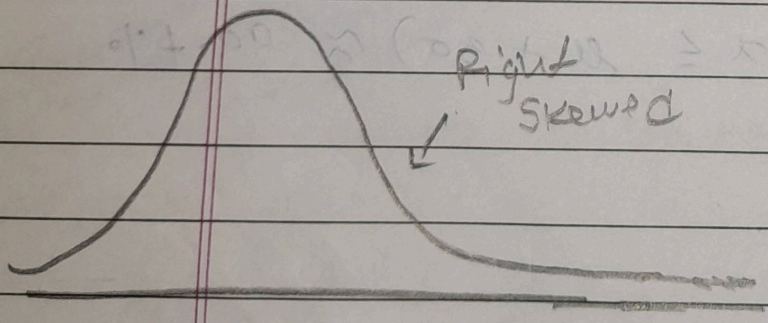
* Log normal

$X \sim$ Log Normal

if $\ln(x) \sim N(\mu, \sigma)$

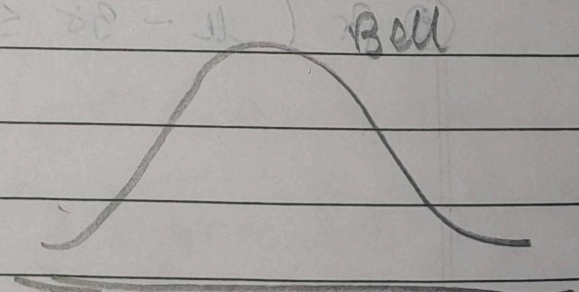
↑ G/N

Log Normal



eg:- 1) Income of the people
2) Amazon product reviews
- Sentiment Analysis usually follows log normal Distribution
- Hospitalization Days

Gaussian Dist



eg:- Distribution of height
- IQ score of 100

Monthly

1500$ → ln(1500) ≈ N(μ, σ)

2000$ → ln(2000) ≈ N

3000$ → ln(3000) ≈ N

log of this value will follow GD

⇒ SND

then convert to SND

why :) $\frac{x_i - \mu}{\sigma}$

∴ Accuracy will be increase.

(4)* Covariance Quantify Relation → for find out some exact relationship through some numerical number we use concept like cov, pearson, ~~t-test~~ spearman

Covariance :- It is one of the very imp topic when we consider Data preprocessing, Data Analysis.

eg: Consider two Random variable

Size	Price
1200 sqr	100K$
1500 sqr	200K$
1800 sqr	300K$

Quantify relation

Size ⟷ Price

Quantify relation is basically means finding relationship between size and Price

∴ If, S↑ P↑ (increase)

S↓ P↓ (decrease)

Covariance

$$\left[cov(Size, Price) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_x) * (y_i - \mu_y) \right]$$

x, y → random variable

n → sample

μ_x → mean of random variable x

μ_y → random variable y

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

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x) * (x_i - \mu_x)$$

$\therefore Cov(x, x) = Var(x)$

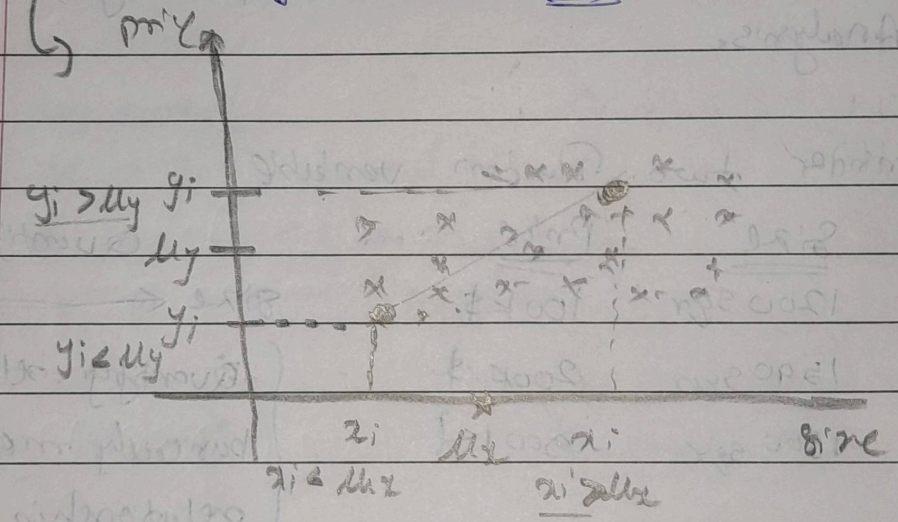
If we are trying to find out cov relation between x & x it is nothing but variance of x

(5)

$$Cov(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x) * (y_i - \mu_y)$$

* $x \uparrow \quad y \uparrow = \square$ +ve

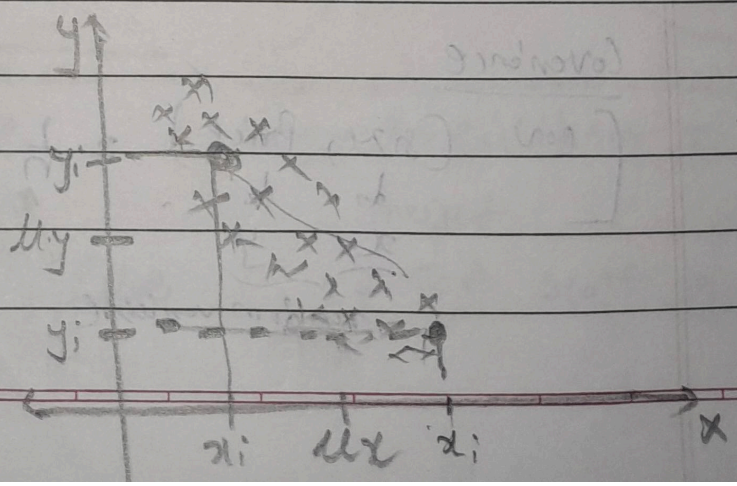
~~$x \uparrow \quad y \downarrow = \square$ -ve~~



$Cov(+) = (+ve) * (+ve) = +ve$

$Cov(-) = (-ve) * (-ve) = +ve$

* $x \uparrow \quad y \downarrow$



$\Rightarrow (+ve) * (-ve) = -ve$

$(-ve) * (+ve) = -ve$

In covariance

$x \uparrow \; y \uparrow \Rightarrow +ve \Rightarrow$ How much +ve (It doesn't tell)

$x \uparrow \; y \downarrow \Rightarrow -ve \Rightarrow$ How much -ve (It doesn't tell)

to overcom this Disadvantage we use

- Because of thise we use Pearson correlation Coefficient

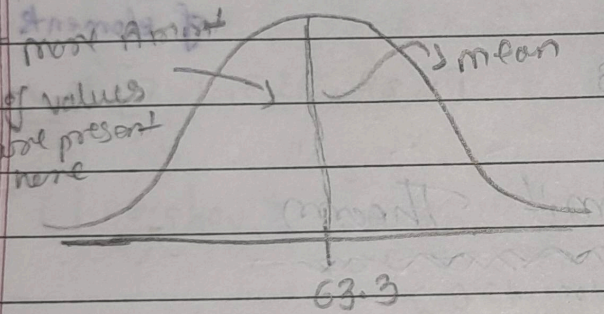
(5) * Mean, Median and Mode

⇒ Sample of height {168, 170, 150, 160, 182, 140, 175} cm

$x_1 \; x_2 \; x_3 \; x_4 \; x_5 \; x_6 \; x_7$

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{[\;]}{7} = 63.5$$

n = no. of sample (7)



⇒ This mean specifies measures of central Tendency

- Median

⇒ [1, 2, 3, 4, 5] $\quad \therefore \frac{15}{5} = 3$

$\mu = 3$

(If i add outlier)

⇒ [1, 2, 3, 4, 5, 50]

$\mu = 13$

After i added outlier

$\therefore \mu = 3 \qquad \mu = 13$

This very hagful for doing any statistical analysis

Note:- for Median the Number should be in shorted order

[1, 2, 3, 4, 5, 50]

Even: $\Rightarrow \frac{3+4}{2} = 3.5$

$\mu = 3.5$

If i have some outlier i can basically use median and mode

Mode:-

Mode Basically find out the maximum number of numbers in the this particular Distribution

repeated many times

[1, 2, 3, 3, 3, 4, 5, 50] Max no. of occurrence of elements

$\therefore \mu = 3$ [1, 2, 2, 3, 3, 3, 4, 5] $\Rightarrow \mu = 2 \& 3$

(6) * Central Limit Theorm

may or may not belong to G.D

* $X \not\sim G.D(\mu, \sigma^2)$

Suppose i know my mean and variance value

μ σ^2

Central Limit theorem specifies, two different point, suppose i take sample from this (X) random variable X.

Then,

Taking Sample as S_1

from this random variable (x). and considering n = 30.

$\therefore \boxed{n \geq 30}$ (n is greater than or Equal to 30 data points)

- Every time selecting 30 random sample value.

S_1 x_1 x_2 ..... $x_{30} = \bar{x}_1$

S_2 x_1 x_2 ..... $x_{30} = \bar{x}_2$

S_3 x_1 x_2 ..... $x_{30} = \bar{x}_3$

S_4 x_1 x_2 .....

S_5 :

S_6 :

:

S_{100} x_1 x_2 ..... $x_{30} = \bar{x}_{100}$

$$\boxed{\bar{x} \approx GD\left(\mu, \frac{\sigma^2}{n}\right)}$$

(variance, condition)

(f) Chebyshev's Inequality

$$\boxed{x \approx GD(\mu, \sigma)}$$

x, random variable belongs to GD with some value of mean (μ) and standard Deviation or variance (σ). If i see consider variance it will be sigma square (σ^2) then we know that based on the Empirical formula.

$Pr(\mu - \sigma \leq x \leq \mu + \sigma) \approx 68\%$

$Pr(\mu - 2\sigma \leq x \leq \mu + 2\sigma) \approx 95\%$

$Pr(\mu - 3\sigma \leq x \leq \mu + 3\sigma) \approx 99.7\%$

* If my ~~g~~ ≠ GD , in this we use Chebyshev's Inequality

$$\boxed{y \neq GD}$$

⇒ Chebyshev's Inequality, ~~SD~~ says that if you want to find out the probability of a random variable falling with the range of 1st SD will ~~always~~ always be greater than or Equal to

$$\boxed{1 - \frac{1}{K^2}}$$

K = Specifies for which range of SD, i have to find out this particular value, ~~i.e~~

i.e, is my percentage of Data points

Explanation → If my random variable does not belong to GD, i am basically able to find out ~~the~~ data ~~to~~ % of the data point will be falling within the range of variance

$$Pr(\mu - K\sigma < x < \mu + K\sigma) > 1 - \frac{1}{K^2}$$

$\therefore K = 2$

$$Pr(\mu - 2\sigma < x < \mu + 2\sigma) > 1 - \frac{1}{K^2} = \frac{1-1}{4} = \underline{\underline{75\%}}$$

*

(8) Pearson Correlation Coefficient

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

eg:- x (Height) y (weight)

cov(x,y) : :

$x \uparrow$ $y \uparrow$ = +ve

$x \uparrow$ $y \downarrow$ = -ve

But, in covariance it doesn't tell how much +ve and how much -ve may be.

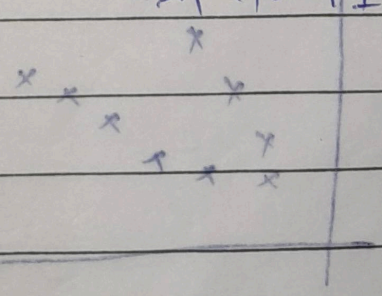
② Pearson correlation coefficient.

→ It says that based on your variance of x and y it will be able to tell you this parameter i.e ① strength (How strong this is correlated) ② Direction y Relationship

- Here, it is combining both the things ① strength and ② Direction of Relationship

$$\text{Pearson CL} = \rho(x,y) = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

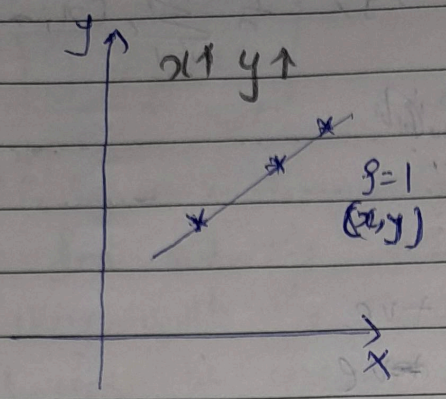
$\rho(x,y) = 0$ between $-1 \le \rho \le 1$ (the value ranges between -1 to +1) in Pearson

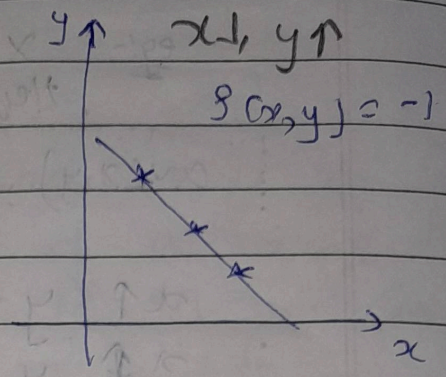


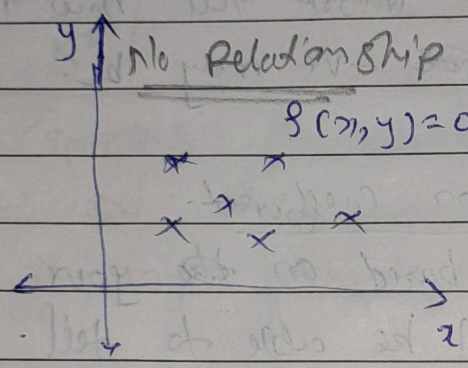
$$\rho(x,y) = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

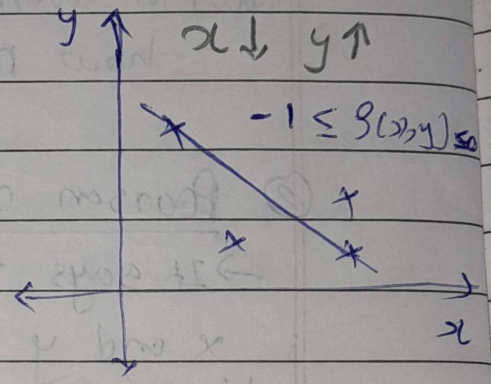
$$-1 \le \rho \le 1$$

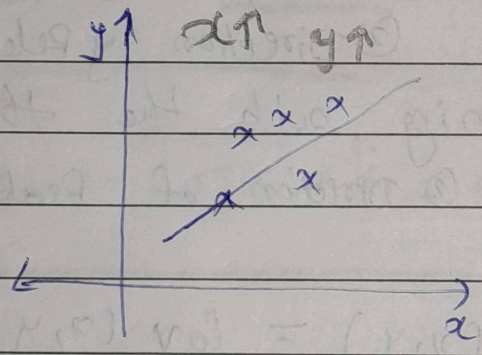
Correlation ⇒







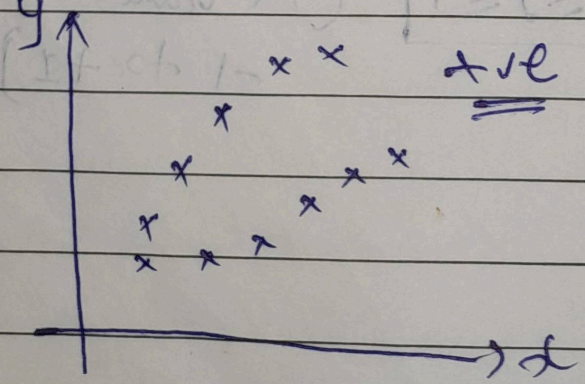




$$\rho(x,y) = 0 \le \rho(x,y) \le 1$$

Note:- This technique is basically used in feature selection.

0 to 1



(a) Spearman's Rank correlation coefficient

The Spearman correlation coefficient is defined as the Pearson correlation coefficient between the rank variable

Here we are trying to find out person of correlation of rank of x and rank of y

$$r_s = \rho_{rgx, rgy} = \frac{cov(r_{gx}, r_{gy})}{\sigma_{rgx}\ \sigma_{rgy}}$$

where,

ρ => denotes the usual Pearson correlation coefficient, but here applied to the rank variable

$cov(r_{gx}, r_{gy})$ => is the covariance deviations of the rank variable

$\sigma_{rgx}, \sigma_{rgy}$ => are the standard deviations to the rank variable

Only if all n ranks are distinct integers, it can be computed using the popular formula

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$$

$\therefore d_i = r_g(x_i) - r_g(y_i)$

n = no. of observation

eg Calculate the correlation between IQ of a person with the number of hours spent in front of TV per week.

IQ, x_i	Hours of TV per week, y_i
106	7
86	2
86	2
101	50
99	28
103	29
97	20
113	12
112	6
110	17

sort in ascending order x_i

rank y_i

$x_i - y_i$

⇒

IQ x_i	Hours of TV per week, y_i	rank x_i	rank y_i	d_i	d_i^2
86	2	1	1	0	0
97	20 (6th highest element)	2	6	-4	16
99	28 (8th highest element)	3	8	-5	25
100	27 (7th highest element)	4	7	-3	9
101	50 (10th highest element)	5	10	-5	25
103	29	6	9	-3	9
106	7	7	3	4	16
110	17	8	5	3	9
112	6	9	2	7	49
113	12	10	4	6	36

$\sum d_i^2 = 194$

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 194}{10(10^2-1)}$$

$$\rho = \frac{-29}{165} \approx -0.175757\ldots$$

Conclusion ⇒ The value is close to 0, the correlation between IQ and TV hour is very low.

(10) ✱ finding Outlier in Dataset using Z-score and IQR

Outlier: An outlier is a data point in a dataset that is distant from all other observation. A data point that lies outside the overall distribution of the dataset.

(1) Z-score

Consider Data points,

$x = [1, 2, 3, 4, 5]$ → if i want to convert this distribution in SND

$\mu = 3 \quad \sigma = 1$

for SND property we get,

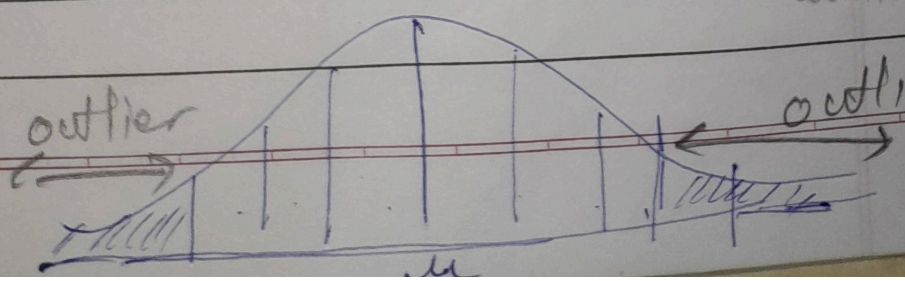
SND ⇒ $\mu = 0, \sigma = 1$

∴ Here, in order to convert this, we usually apply Z-score

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

observation, mean, SD

• Z-score tell that in SND to outlier the far Dot full after the 3rd SD are considered as outlier.



③ IQR

75%, 25% values in a dataset

Steps:-
1) Arrange the data point in increasing order
2) Calculate first (q1) and third quartile (q3)
3) find Interquartile range (q3 - q1)
4) find lower bound q1 * 1.5
5) find upper bound q3 * 1.5

- Anything that lies outside of lower and upper bound is an outlier.

- Various ways of finding outliers
1) Scatter plot
2) Box plot
3) Z-score
4) Using IQR