C&S electric

Basic Stats

Probability

1. Introduction to Basic Terms
2. Variables
3. Random Variables
4. Population, Sample, Population Mean, Sample Mean
5. Pop Distribution, Sample Distribution, Sampling Distribution
6. Mean, Median, Mode - Measures of Central Tendency
7. Range
8. Measure of Dispersion
9. Variance
10. Std. Deviation
11. Gaussian / Normal Distribution

Intermediate Stats

12. Std. Normal Distribution
13. Z score
14. Prob. Density Function
15. Cumulative Distribution function
16. Hypothesis Testing
17. Many different plotting graphs
18. Kernel Density Estimation
19. Central Limit Theorem
20. Skewness of Data

Adv. Stats

25. Q-Q Plot
26. Chebyshev's Inequality
27. Discrete & Continuous Distribution
28. Bernoulli & Binomial Distribution
29. Log Normal Distribution
30. Power Law Distribution
31. Box Cox Transform
32. Poisson Distribution
33. Application of Non Gaussian Distribution

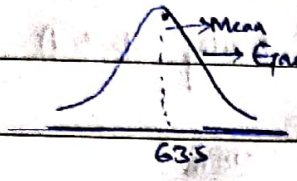
V. Mean, Median, Mode - Measure of

(i) Mean :- Measure of Central Tendency.

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

Sample of Height = {168, 170, 150, 160, 182, 140, 175}

$\mu = \frac{168 + \cdots + 175}{7} = 163.5$



1 2 3 4 5 50

$\sigma^2 = \frac{1}{6}[(1-10.83)^2 + (2-10.83)^2 + \cdots + (50-10.83)^2]$

308.472233 $\sigma = 17.56$

$\mu = 10.83$

(ii) Median :-

[1,2,3,4,5] → Mean = $\frac{1+2+3+4+5}{5} = 3$

Median = 1,2,3,4,5 = 3. (Middle)

-24.89 -6.83 28.39 45.95

10.83

83.3%

Outlier Added [1,2,3,4,5,50] → Mean = $\frac{1+2+3+4+5+50}{6} = 10.83$ (64%)

Median 1,2,3,4,5,50 = $\frac{3+4}{2} = 3.5$

Write in Ascending Order to find Median.

Size :- Odd No - Middle.

Even No - $\frac{m_1 + m_2}{2}$

Age
23
24
27
32
35
21
—
—

} Data Mining.

If I remove the missing Data. I may loose some information

In that case, I will fill Data using Mean, Median, Mode

Mean - If no outliers are present - Prefer Mean.

Median/Mode - More Outliers are present - Prefer Median/Mode.

(iii) Mode -

Statistics

I.

1. Population :-

Question : Average Height of all people in a state

$N \rightarrow$ Population.
$n \rightarrow$ Sample.
$\mu \rightarrow$ Population Mean
$\bar{x} \rightarrow$ Sample Mean

Solution :-

$$\mu = \frac{1}{10L} \sum_{i=1}^{10L} x_i \Rightarrow \boxed{\mu = \frac{1}{N} \sum_{i=1}^{N} x_i}$$

2. Sample :-

To get Data from all people is difficult (Population).

So, We consider Sample.

$$\bar{x} = \frac{1}{10000} \sum_{i=1}^{10000} x_i \Rightarrow \boxed{\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i}$$

$n << N$

When Sample Size increases $n \cong N$

II.

1. Random Variables :- A random Variable which stores Data. Types < Numerical R.V. / Categorical R.V.

Types of ~~R.V~~ Numerical R.V :-

(i) Discrete Variables - Whole Number.

eg:- No. of Bank Accounts a person has? ~~Ans~~

Ans:- 2, 5, 7... Can't be 2.5 (decimals)

(ii) Continuous Random Variable :- Within a range, any Value.

10.1, 10.2, 10.368....

Depends on Scaling eg:- Weight of a Person. 63 Kg.
63.2 Kg. (depends on scaling)
63.269 kg.
Interest Rate
Salary.

Categorical Variable :- for each record it is continuous repeated.

Gender is Categorical R.V.

eg:- Demographic Statistics

House Owner	Age	Gender M/F	No. of People in House	Height of House Owner
	Numerical C.R.V / D.R.V	Categorical	Numerical D.R.V.	Numerical C.R.V.

C&S electric

Distributions.

1. Gaussian / Normal Distribution :

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

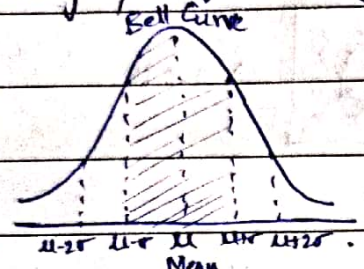
$X \sim GD(\mu, \sigma)$ (Mean, S.D.) or $N(\mu, \sigma)$

$GD(\mu, \sigma^2)$ or $N(\mu, \sigma^2)$ (Variance)

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

Std. deviation $\sigma = \sqrt{Var}$

Prob. Density Fun / Histogram



Std. deviation : From Mean how far the other element is distributed / deviated. (Right / left distance)

Emperical formula :

$Pr(\mu-\sigma \le X \le \mu+\sigma) \approx 68\%$

$Pr(\mu-2\sigma \le X \le \mu+2\sigma) \approx 95\%$

$Pr(\mu-3\sigma \le X \le \mu+3\sigma) \approx 99.7\%$

(Probability of having Data points in first σ deviation)

Bell Curve & Symmetrical.

Dataset eg: (i) Distribution of Height
(ii) IRIS Dataset – Petal length.

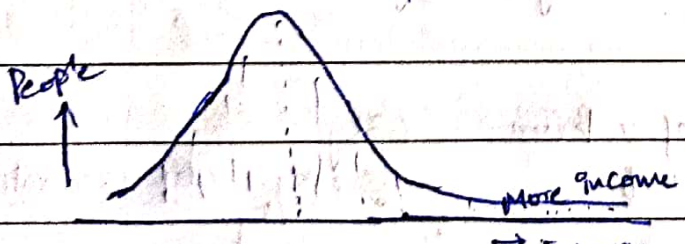
Std SND ⇒ $\mu = 0$, $\sigma = 1$

2. Log Normal Distribution

$X \sim$ Log Normal Distribution if $\ln(X)$ is Normally Distributed.

$X = \{x_1, x_2, x_3 \dots x_n\}$ $\ln(x_1), \ln(x_2) \dots \ln(x_n)$

$\ln(X) \sim N(\mu, \sigma)$



Similar to G.D, but Right Hand side goes more flatten.
⇒ Right Skewed.

Dataset eg: (i) Income of the people
(ii) Amazon Product Review Description Length

More Income people are very less.

Why ~~we~~ we use all types of this distributions?

Money Spent on

	R&D	Marketing	Profits states Company
1	10800/-	100000/-	
2	1/-	1/-	
3	1/-	1/-	
4			
5			
6			
	Gaussian	Log Normal	

As per Domain Knowledge, If we know R&D is Gaussian Distribution. & Marketing is Log Normal Distribution.

R&D → Scaling ~~do~~ to Std. Normal Distribution ($\mu=0, \sigma=1$)

$$\frac{\bar{x}_i - \mu}{\sigma}$$

Marketing →

Marketing	Log (marketing)
1800	ln(1500)
	ln( )
	ln( )

Log → will follow Normal Distribution.

ln ≈ Gaussian ($\mu=0, \sigma=1$)

$$S.N.D = \frac{x_i - \mu}{\sigma}$$ Scaling done.

It is called Log Normalization.

Now after Scaling both R&D, Marketing (comparing) → Analyzation with More Accuracy.

~~For~~

~~Covariance~~ :

C&S electric

Q.

Covariance :- Useful in Data Preprocessing.

Size	Price
1200 sqm	100/-
1500 sqm	200/-
1800 sqm	300/-

Relation b/w Size & Price

$S\uparrow \rightarrow P\uparrow$ or $S\downarrow$ or $P\downarrow$

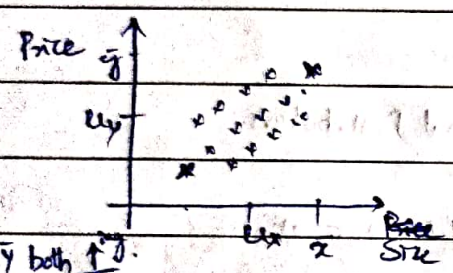
$$Cov(X,Y) \quad Cov(\text{size, Price}) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)$$

(x = size, y = price)

$Cov(X,X) = Var(X)$

$X\uparrow$ & $Y\uparrow \Rightarrow$ Cov is +ve

$X\uparrow$ & $Y\downarrow \Rightarrow$ Cov is −ve.



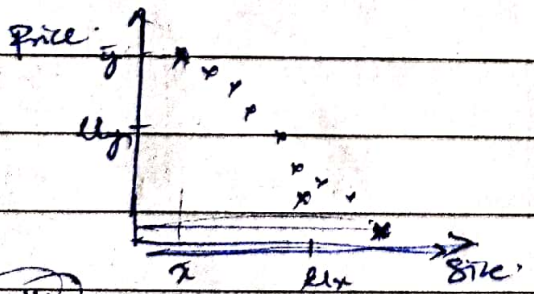
Here Covariance doesn't tell about how much Covariance b/w 2 Quantities.

That's why we use Pearson Correlation Coef.

$\bar{x}$ & $\bar{y}$ both $\uparrow$.

$\bar{x} > \mu_x$, $\bar{y} > \mu_y$ Cor. = +ve . +ve = +ve.

$\bar{x} < \mu_x$, $\bar{y} < \mu_y$ Cor. = −ve . −ve = +ve.



$\bar{x}\uparrow$ $\bar{y}\downarrow$ or vice versa.

$\bar{x} > \mu_x$, $\bar{y} < \mu_y$ Cor. = +ve . −ve = −ve.

$\bar{x} < \mu_x$, $\bar{y} > \mu_y$ Cor = −ve . +ve = −ve.

~~Population & Sample~~

~~(i) Population Mean~~

VI. Central Limit Theorem:-

$\underset{(R.V.)}{X} \not\sim G.D(\mu, \sigma^2) \quad n \geq 30$

This R.V. may/may not belong to a Gaussian Distribution

Taking sample from R.V. X (30 data points. Randomly selected)

		Mean
μ σ^2	S_1 $x_1\ x_2 \dots x_{30}$	$= \bar{x}_1$
	S_2 - - - - - -	$= \bar{x}_2$ (some other Random datapoints)
	S_3 $x_1 \dots$	$= \bar{x}_3$
	$\vdots$	
	$S_{100} \dots$	$= \bar{x}_{100}$

$\bar{x} \approx \mu$

According to Theorem ~~Mean~~ $\bar{x} \approx G.D\left(\mu, \frac{\sigma^2}{n}\right)$ $n \geq 30$.

Mean follows G.D.

VII. Chebyshev's Inequality

For suppose R.V. X following G.D

$X \approx G.D(\mu, \sigma)$

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

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

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

For suppose R.V. Y not following G.D.

$Y \not\sim G.D(\mu, \sigma)$

What % of datapoints in Y falling within range of first Std. deviation?

Chebyshev's Inequality.

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

$K=2$ 2nd Std. deviation. $Pr(\mu-2\sigma < x < \mu+2\sigma) > 1 - \frac{1}{2^2} \Rightarrow Pr(\mu-2\sigma < x < \mu+2\sigma) \geq \frac{3}{4}$

$\Rightarrow Pr(\mu-2\sigma < x < \mu+2\sigma) > 75\%$

$K=3$ 3rd Std. deviation $Pr(\mu-3\sigma < x < \mu+3\sigma) > 1 - \frac{1}{3^2} \Rightarrow \frac{8}{9} \times 100 > 88.9\%$

C&S electric

Pearson Correlation Coef.

Linear

Wikipedia

Useful for feature selection.

$$\text{Covariance} = \text{Cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_x)*(y_i-\mu_y)$$

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

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

Here Covariance gives only Direction of Relationship.

but doesn't Knows about Strength.

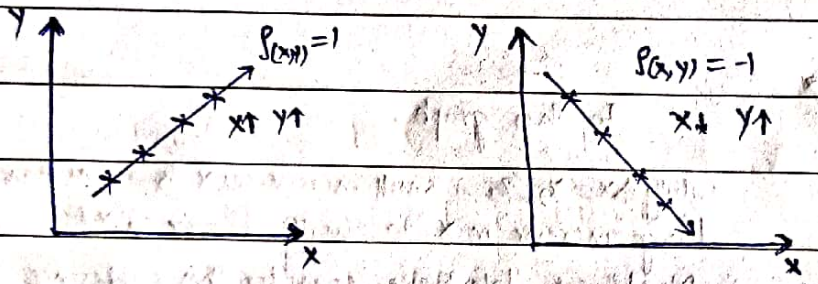
x: Height, y: Height

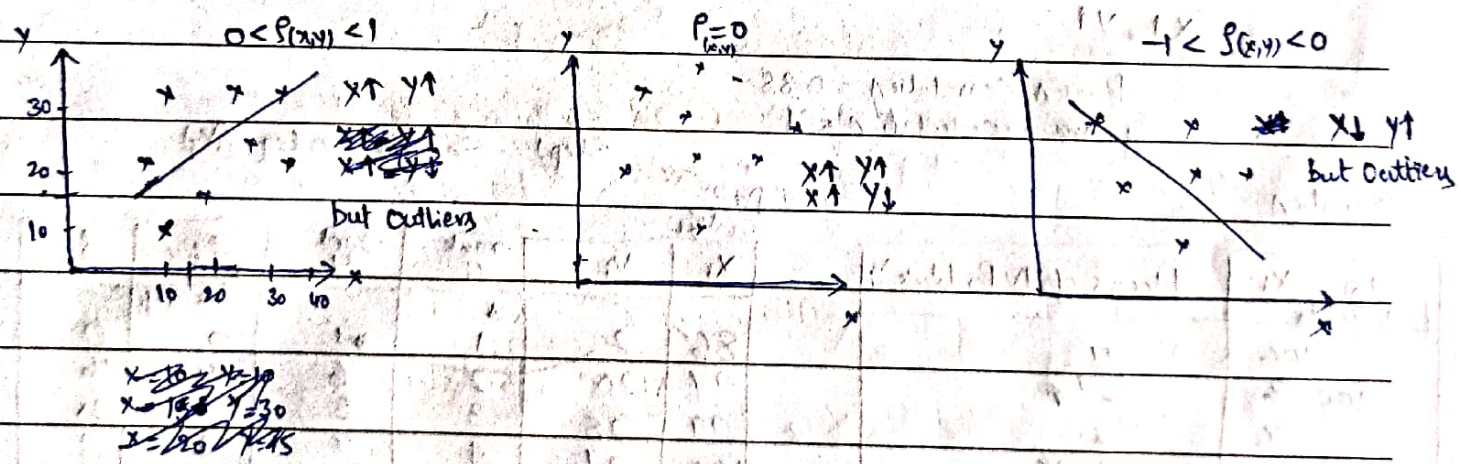
$x\uparrow \Rightarrow y\uparrow^{cm}$

$$\text{Pearson C.C} = \rho_{(x,y)} = \frac{\text{Cov}(X,Y)}{\sigma_x \cdot \sigma_y}$$

We get How much positivity/Negativity i.e. Strength.

Value Ranges : $\rho_{(x,y)}$: $\boxed{-1 \le \rho \le 1}$





Suppose $\rho_{(x,y)} = 1$, i.e. $x\uparrow\ y\uparrow$ we can drop One Variable either X or Y for feature selection in M.L. as both features are same.

Linear & Non Linear

B. Spearman's Rank Correlation Coef:- Wikipedia

* HeatMap use this Spearman's Correlation.

* $r_s = \rho_{rg_x, rg_y} = \frac{cov(rg_x, rg_y)}{\sigma_{rg_x} \cdot \sigma_{rg_y}}$ Here We considering Ranks

Rank of x × Rank of y

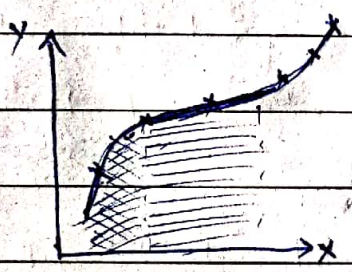
$\rho \rightarrow$ Pearson Correlation Coef., but applied to Rank Variables.

$cov(rg_x, rg_y) \rightarrow$ Covariance of Rank Variables.

$\sigma_{rg_x}, \sigma_{rg_y} \rightarrow$ Std. deviations of Rank Variables

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

$d_i = rg(x_i) - rg(y_i) \Rightarrow$ Diff. b/w 2 Ranks of Each Observation.



In this graph,

Initially a small increase in X ⇒ Small increase in Y.

Large increase in X ⇒ Small increase in Y.

So, Non Linear We have consider this difference also.

X↑ ▲ Y↑

Pearson Correlation = 0.88

Spearman Correlation = 1

Question

Students	IQ Xi	Hours of TV Per Week Yi
1	106	7
2	100	27
3	86	2
4	101	50
5	99	28
6	103	29
7	97	20
8	113	12
9	112	6
10	110	17

Answer:-

Step1 :- Sort the order (Xi)

Sorted Xi	Yi	rank Xi	rank Yi	d_i = rank(Xi) − rank(Yi)	d_i^2
86	2	1	1	0	0
97	20	2	6	−4	16
99	28	3	8	−5	25
100	27	4	7	−3	9
101	50	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

Pearson Correlation :- Cov(x,y) → diff. of X & diff of Y Independently (9)

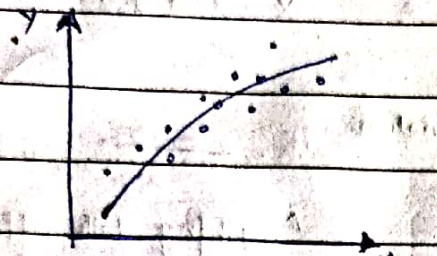
Spearman Correlation :- Rank(x,y) → diff. of X & Y dependent.

$$\rho = 1 - \frac{6 \times 194}{10(10^2-1)} = \frac{-29}{165} = -0.17575757....$$

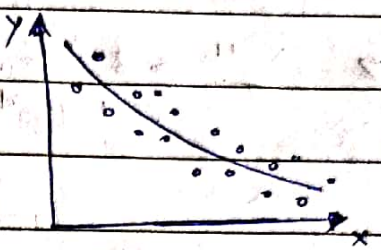
ρ Coef. is very low.

No Relation b/w IQ & Hours per T.V.

(p Value 0.627 t-distribution)

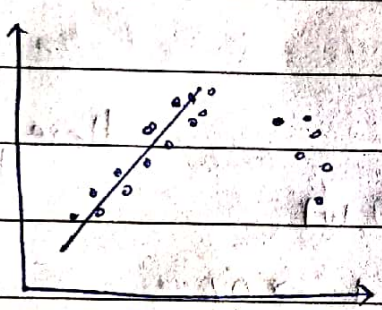


Spearman Correlation = 0.92



Spearman Correlation = −0.91

Outliers



Spearman Correlation = 0.84

Pearson Correlation = 0.67

Pearson Correlation focus on Linear

but Spearman helps on Non Linear also.

C&S electric

Finding Outliers in Dataset using Z score and IQR

Outlier:-

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

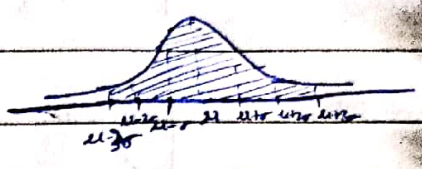
Z-score :

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

$$x = \mu + z\sigma$$

$$z = \frac{x-\mu}{\sigma} > 3 \quad \text{[Outlier]}$$

Out of 3rd Std. Deviation.



Anything falls after $\mu+3\sigma$ is an Outlier

$z \geq 3$.
Outlier.

I.Q.R Inter Quartile Range – Est

dataset = [5, 6, 7, 1, 2, 8, 10, 3, 4, 7, 9]

~~for suppose~~

10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

Here Calculation Wrong.

Ans: Sort the dataset = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]

1 → ~~Those one~~ 1 → 0 numbers are less than 1 → 0%

2 → 10% of Total Numbers are less than 2 ie 1
(before 2 = 1 Number only) i.e. $\frac{1\text{ Number}}{\text{Total Numbers}} \times 100\% = \frac{1}{10} \times 100 = 10\%$

10%

*** 3 → 1, 2, 3, 4 . . . 10

2 Numbers are less than 3. ie. $2/10 \times 100 = 20\%$

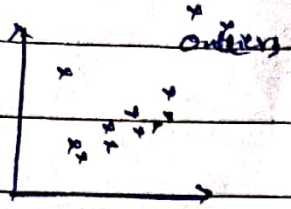
4 → 30%

In Inter Quartile Range, Most Numbers are Concentrated ~~at~~ at 25% 75%

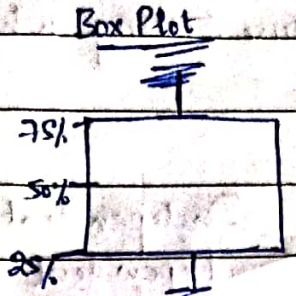
I.Q.R = 75% – 25% values

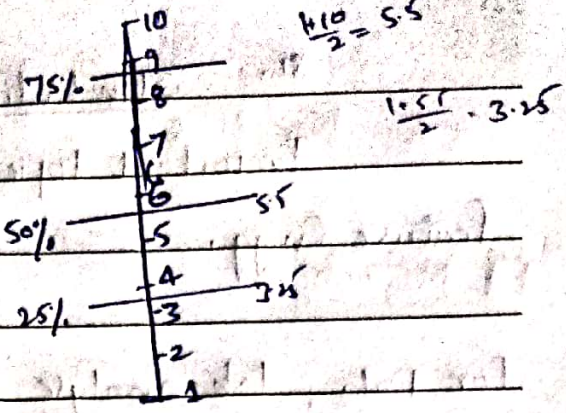
C&S electric

Scatter Plot



Box Plot





$\frac{4+10}{2} = 5.5$

$\frac{1+5}{2} = 3.25$

Below 25% & Above 75%

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
17 18 19 20 21 22 27 24 25 26 27 28 29 30 31 32 33

S = [10, 10, 10, 10, 10, 11, 11, 12, 12, 12, 12, 12, 12, 13, 13, 13, 13,
14, 14, 14, 14, 14, 14, 15, 15, 15, 15, 15, 17, 17, 102, 107, 108]

Sorted the List.

0% 2.9% 5.8% 8.8 11.7 14.7 17.6 20.5 23.5 26.4 29.4 32.5 35.2 38.2 41.1 44 47 50

S = [10, 10, 10, 10, 10, 11, 11, 12, 12, 12, 12, 12, 12, 12, 12, 13, 13, 13, 13,
14, 14, 14, 14, 14, 14, 15, 15, 15, 15, 15, 17, 19, 102, 107, 108]

52.9 55 58.8 61.7 64.7 67.6 70.5 73.5 76.4

Total Elements = 34

Percentile:

1st Element '0' Index: '10' - $\frac{0}{34} \times 100\% = 0\%$

2nd Element 1 Index: '10' - $\frac{1}{34} \times 100\% = 2.9\%$

6th Element 5 Index '11' - $\frac{5}{34} \times 100\% = 14.7\%$ (5 elements are there before 11)

Lower

$\frac{x}{34} \times 100 = 25\%$

$x = 8.5$

9

8

Upper

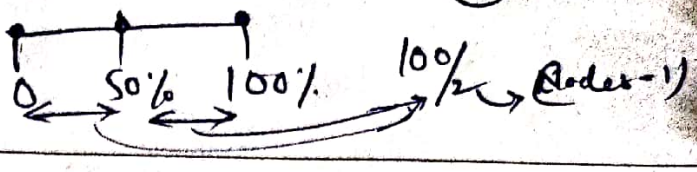
$\frac{y}{34} \times 100 = 75\%$

$y = 25.5$

25

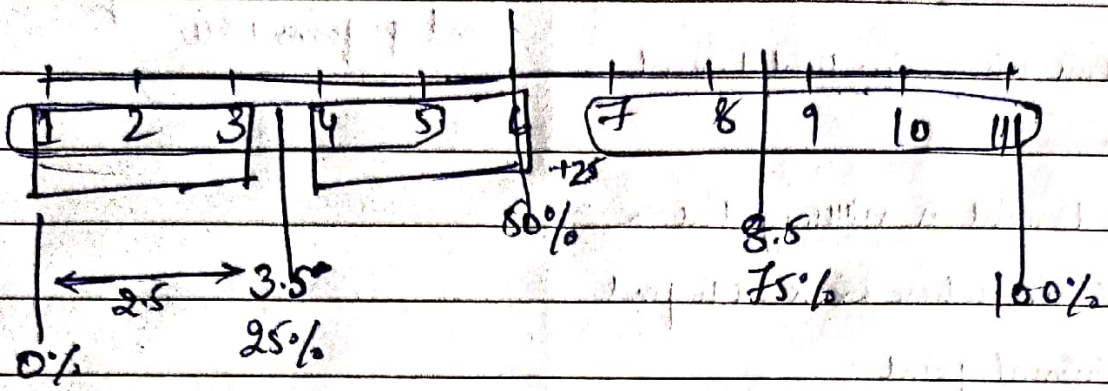
24

C&S electric



IQR Practice

11 Numbers



~~(Total No. Count + 1) P%~~
~~25% ((11+1) 25% = ((1+1) × 1/4 = 3~~ Value

~~1/ ×100 = 9.09%~~

11 nodes = Total No. Count

Each is Segment $100/11-1 = 10\%$

1 → 0%

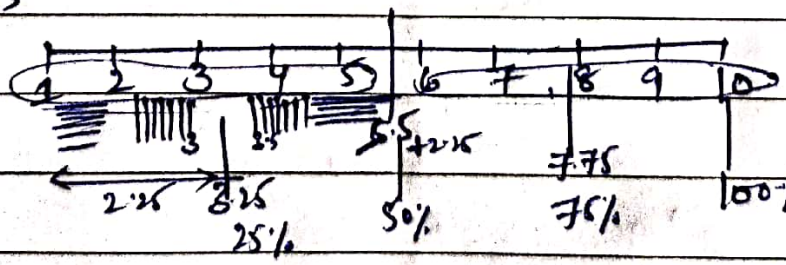
2 → $1/(11-1) \times 100\% = 10\%$

3 → $2/11-1 \times 100\% = 20\%$ Before Node 3 → There are 2 Numbers

$x = \frac{25\%}{100\%} \times 10 = 2.5$ Each 25% is.

~~75% Value = 25% (Total~~

10 Numbers



$25\% (10-1) = 9/4 = 2.25$

P% Value = P% × (Total Count - 1)

C&S electric

Interquartile range

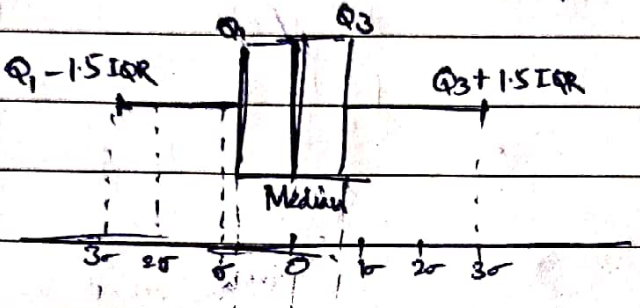
H-Spread, Q-Q, middle 50%

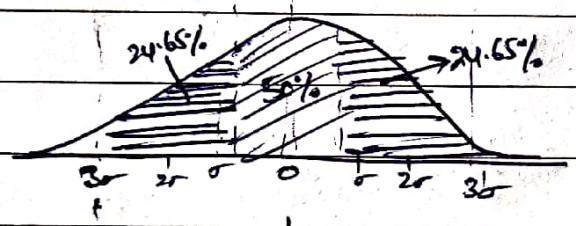
Measure of Statistical dispersion, diff. b/w 75th and 25th percentiles

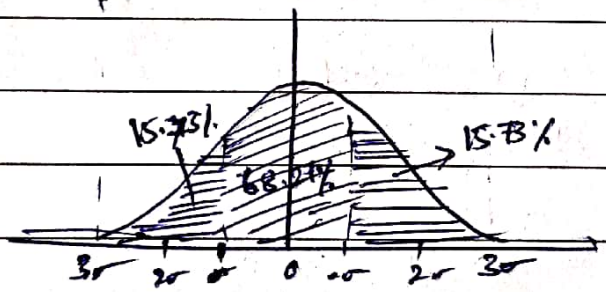
$$IQR = Q_3 - Q_1$$

Here Median is Corresponding Measure of Central Tendency

IQR - Used in Outliers Identifying







Order 7, 7, 31, 31, 47, 75, 87, 115, 116, 119, 119, 155, 177.

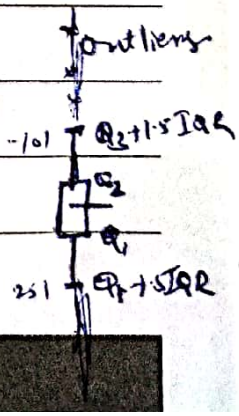
31 → Q_1 Median of First Half

Q_2 (Median of Whole Table)

119 → Q_3 Median of Last Half.

$$IQR = Q_3 - Q_1 = 119 - 31 = 88.$$

$$\text{Range (L.B, U.B)} = (Q_1 - 1.5IQR,\ Q_3 + 1.5IQR)$$
$$= (-101, 251)$$



C&S electric

11. Standardization Vs Normalization

Important topic for Feature Scaling which is an integral part of feature Engg.

For Data, We analyse features < Independent / Dependent

* Using Independent Variables, We analyse the Dependent Variables.

Features → We Consider ~~of~~ Magnitude & Units.

	Age	Weight (kgs)	Height (cms)
Units (No of items)	25	60	162
Magnitude ↓	25 No of Items	Kgs	Cms

Two scaling Techniques are Standardization & Normalization.

(i) Normalization :- (min-max Normalization)

Helps to scale down the feature b/w 0 to 1.

min

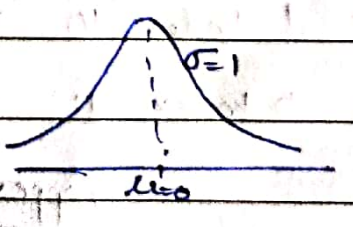
$$X_{Norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

(ii) Standardization :- (Z-Score Normalization)

Helps to scale down the feature ~~or~~ based on Std. Normal Distribution.

with Mean $\mu = 0$
Std. Deviation $\sigma = 1$

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



~~Mean is~~

C&S electric

Normalization (MinMax)

* M.L. Algorithm which involves Euclidean distance.
* D.L Algorithms where Gradient Descent is involved.

→ Gradient Descent is nothing but a Parabola Curve where we need to find. Global Minimal point.

* We use Min-max Normalization
* Algorithms like KNN
  K means Neighbour
  K means Clustering
  ANN Artificial Neural N/w
  CNN Convolutional Neural N/w } Scale down 0 to 1. Images [0 to 255]
  Linear Regression
  Logistic Regression.

Standardization. (StandardScaler)

* Many of the use cases Standardization Technique is used. and performs well.

* Some Algorithms, We wont perform Scaling. like Decision Tree, RandomForest, XGBoost and all boosting techniques.
  No use of Scaling Techniques here.

ANN - Tensor Flow, Keras require Inputs b/w 0 to 1 to analyse weights quickly.