Statistics

Agenda

1. Sampling Techniques ✓
2. Covariance And Correlation
3. Probability distribution function
4. Probability density function
5. Probability Mass function
6. Cummulative Density Function
7. Types of Distribution

1. Sampling Techniques.

1. Random Sampling : Simple Random Sampling gives each member of the population an equal chance of being chosen for the sample.

Eg: Vaccination Test → 100 → Randomly Select Couple

Accidental Test → 1000 → A → 2 vehicle

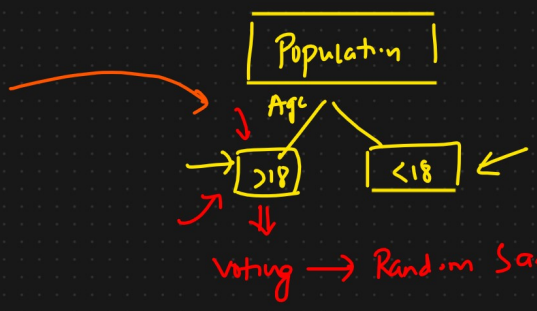
Exit Poll

Average IQ of the school → Select 10 people ⇒

2. Stratified Sampling [Stratified → layers]

Stratified Sampling Involves dividing the population into sub population that may differ in important ways

Eg: Exit Pool

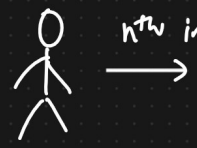


2. Salary

③ Systematic Sampling : Systematic Sampling is a statistical method involving the Selection of elements from an ordered sampling frame.

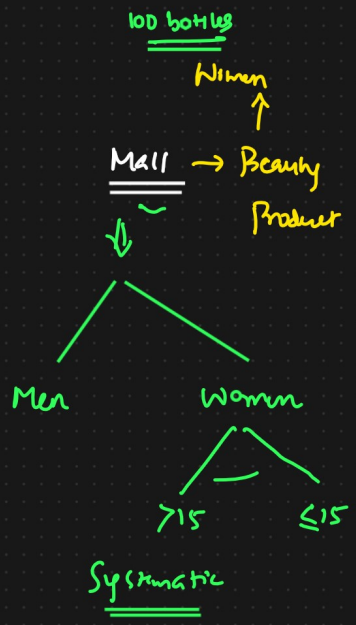
Eg: Airport

A Credit Card Company



nth individual → 7th

Mall → Fill up a form for a cause.

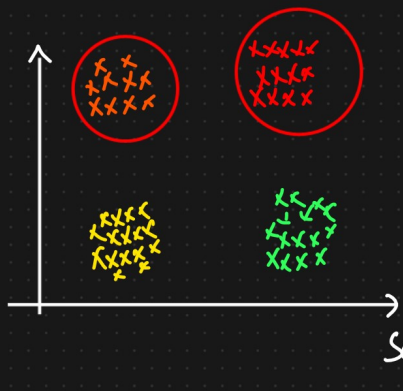


④ Convenience Sampling Assignment

⑤ Purposive Sampling → Judgemental Sampling is a method where researchers decides which members of the target population will be sampled.

⑥ Cluster Sampling

Mall iphone 28 → 100 → 20%.

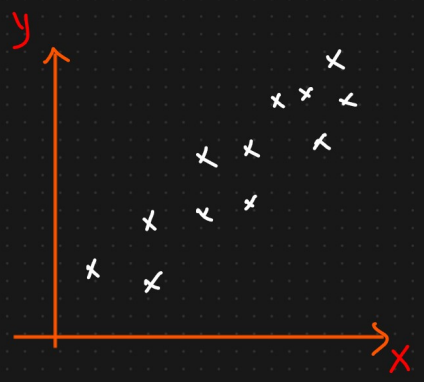


② Covariance And Correlation

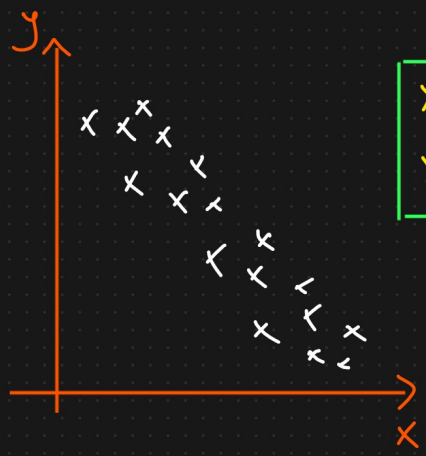
Quantify
[Relationship between X And y].
⇓
Covariance And Correlation.

X	y
2	3
4	5
6	7
8	9

X↑	y↑
X↓	y↓



X↑	y↑
X↓	y↓



X↑	y↓
X↓	y↑

Covariance

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

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

x_i → Data point of x
$\bar{x}$ → Sample mean of x
y_i → Datapoints of y
$\bar{y}$ → Sample mean of y.

$Var(x) = Cov(x,x)$ ⇒ Spread

$Cov(x,y)$

X↑	y↑
X↓	y↓

+ve Covariance

X↑	y↓
X↓	y↑

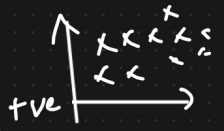
-ve Covariance

X	y
→ 2	3
→ 4	5
6	7
$\bar{x}=4$	$\bar{y}=5$

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

$$= \frac{[(2-4)(3-5) + (4-4)(5-5) + (6-4)(7-5)]}{2}$$

$$= \frac{4+0+4}{2} = 4 //.$$



X	y
8	6
7	5
6	4
3	1
2	0
$\bar{x}=5.2$	$\bar{y}=3.2$

+ve Covariance

X	y
8	6
9	5
10	4
12	2
11	1
10	3.6

−ve Covariance

−3

6.7

Advantages

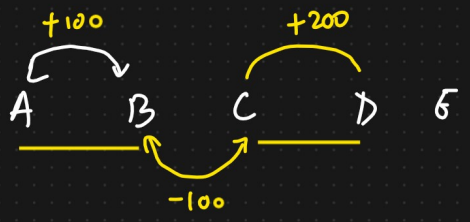
① Relationship between X and y

+ve or −ve value

A ⟷ B + infinity, − infinity.

Disadvantages

① Covariance does not have a specific limit value



② Pearson Correlation Coefficient [-1 to 1]

$$\rho_{x,y} = \frac{Cov(X,Y)}{\sigma_x \cdot \sigma_y} \Rightarrow -1 \text{ to } 1$$

① The more the value towards +1, the more +ve correlated X, Y is.

② The more the value towards -1, the more -ve " X, Y is

X	Y
8	6
9	5
10	4
12	2
11	1
$\bar{X} = 10$	$\bar{Y} = 3.6$

$\sigma_x = 1.58$ $\sigma_y = 2.073$

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

$\sigma = \sqrt{\text{Variance}}$

$Cov(x,y) = -3$

$$\rho_{(x,y)} = \frac{-3}{1.58 * 2.073} = \frac{-3}{3.27} = -0.917 \Rightarrow \text{Negative Correlated} \Rightarrow 91.7\%.$$

$$\frac{[4+1+0+4+1]}{4} = \frac{10}{4} = 2.5$$

$\sigma = \sqrt{2.5}$

X↑ Y↓ => 91.7%

1 unit X↑ Y↓ 0.917.

③ Spearman Rank Correlation [-1 to 1].

$$r_s = \frac{Cov(R(x), R(y))}{\sigma(R(x)) * \sigma(R(y))}$$

X	y	R(x)	R(y)
1	2	2	2
3	4	3	3
5	6	4	4
7	8	5	6
0	7	1	5
8	1	6	1

Ascending Order.

{ Non linear Relationship }

Spearman Rank Correlation

⇓

Non linear Relationship

DATA SCIENCE

Feature Selection

Housing Price DATASET

Size of House ↑ — No. of Rooms ↑

95%

-95%

+ve or -ve Correlation

-ve correlation

Haunted

≈0

No. of people staying

Price ↑

O/P