X : $[1, 2, 3, 4, 5]$ (Hours spent studying)

Y : $[50, 60, 70, 80, 90]$ (Exam scores)

$y_1 = 500, 600, 700$

Calculate covariance and correlation of above.

$$cov = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{n}$$

$$cor = \frac{cov}{\sigma_x \sigma_y}$$

x	μ_x	$x - \mu_x$	y	μ_y	$y - \mu_y$
1	3	-2	50	70	-20
2	3	-1	60		-10
3	3	0	70		0
4	3	1	80		10
5	3	2	90		20

$$E(x - \mu_x)(y - \mu_y)$$

$$\frac{(-2)(-20) + (-1)(-10) + 0 + 1(10) + 2(20)}{5}$$

$$\Rightarrow 20$$

$$\sigma_x = \sqrt{\frac{(x - x_0)^2}{n}} \quad \Rightarrow \sigma_x^2 = \frac{\Sigma (x_i - x_0)^2}{n}$$

$$\text{Cov}_{xy'} > \text{Cov}_{xy}$$

$\Rightarrow$ x & y' are more related.

$$\gamma_{xy} = \frac{Cov(x,y)}{\sigma_x \sigma_y} \qquad -1 \le \gamma_{xy} \le 1$$

$$= \frac{\sigma_{xy}}{\sigma_x \sigma_y} \qquad \gamma_{xy} = \frac{E\left[(x-\mu_x)(y-\mu_y)\right]}{\sqrt{E(x-\mu_x)^2 \; E(y-\mu_y)^2}}$$

$$\sum_{i=1}^{n} \left[a(x_i-\bar{x}) + (y_i-\bar{y})\right]^2 \ge 0$$

a = real no. $\neq 0$

$$\sum_{i=1}^{n} \frac{a^2 (x_i-\bar{x})^2}{n} + \sum_{i=1}^{n} \frac{2a\,(x_i-\bar{x})(y_i-\bar{y})}{n} + \sum_{i=1}^{n} \frac{(y_i-\bar{y})^2}{n} \ge 0$$

$$a^2 \sigma_x^2 + 2a\,\sigma_{xy} + \sigma_y^2 \ge 0$$

$$a^2 \sigma_x^2 + 2a\sigma_{xy} + \left(\frac{\sigma_{xy}}{\sigma_x}\right)^2 + \sigma_y^2 - \left(\frac{\sigma_{xy}}{\sigma_x}\right)^2$$

$$\underbrace{\left(a\sigma_x + \frac{\sigma_{xy}}{\sigma_x}\right)^2}_{0} + \underbrace{\left[(\sigma_y)^2 - \left(\frac{\sigma_{xy}}{\sigma_x}\right)^2\right]}_{\ge 0} \ge 0$$

$$a = -\frac{a_{xy}}{a_x}$$

$$(a_y)^2 + \left(\frac{a_{xy}}{a_x}\right)^2 \geq 0$$

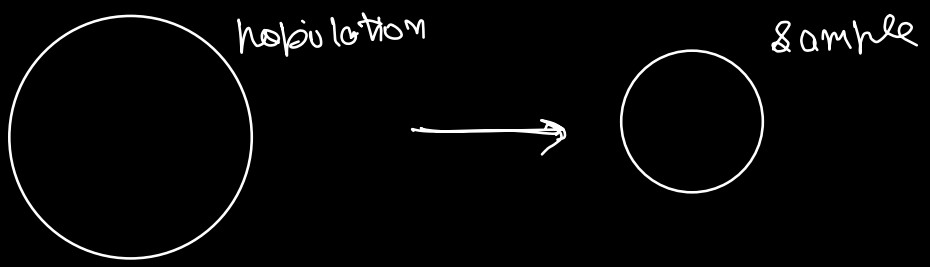
$$(a_y)^2 \geq \frac{(a_{xy})^2}{(a_x)^2}$$

$$1 \geq \frac{a_{xy}^2}{a_x^2\, a_y^2}$$

$$1 \geq r^2$$

$$\boxed{-1 \leq r \leq 1}$$

Sampling Techniques



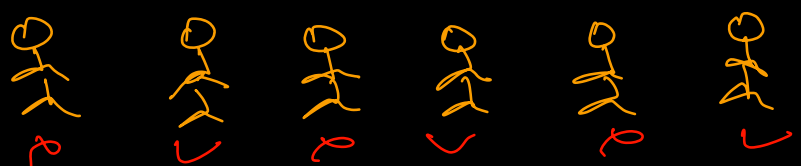
few ways to take the sample :-

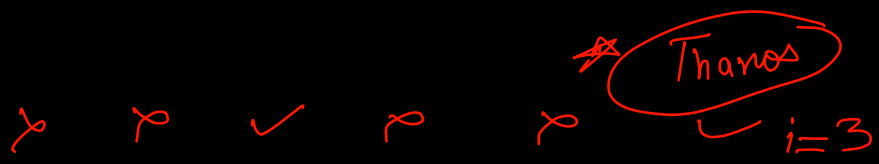
1. **Simple Random Sampling** :- Every member of the population (N) has a equal chance of getting selected in the sample (n).

 ⇒ 1/5

2. **Systematic Sampling**

We take every i^{th} individual / item

 i=2

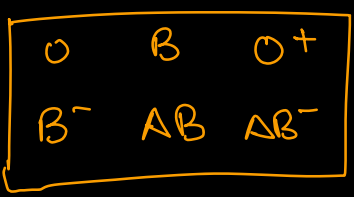


Airport security check ⟹ 10th size

3. Stratified sampling.

Strata
↓
layer
↓
non-overlapping region
↓
clusters.

we will collect some proportion from each strata.



B+

[5% + 3% + ...]

[Zero tax 0-5% 5-10% 10-20% 30% +]

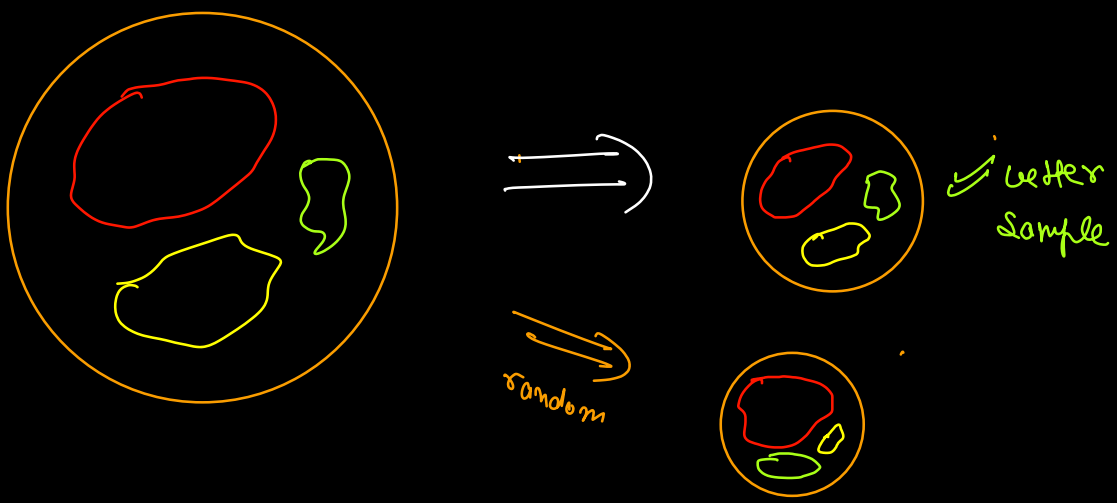
3%

5%

5%

7%

80%



Convenience Sampling ⇒ Only those who are interested will be participating

apple ecosystem ?

Few More Points on data

18, 20, 22, 25, 27, 28, 30, 32, 34,

35, 36, 40, 42, 45, 48, 50, 52, 55, 60, 75

80 90 100

Range = highest − lowest

= 75 − 18 = 57

Percentage = per cent

represents a portion of a whole

expressed as fraction of 100

Percentile :

percentile actually divides your data into 100 equal parts.

% 100 %ile

99 marks 99 % people behind you

Calculate p^{th} percentile ?

$$\text{rank} = \frac{p}{100} \times (n+1)$$

linear interpolation

hw ?

whole no. ⇒ take that rank

if not ⇒ nearest value / interpolation as well

$\frac{25}{4 \cancel{100}} \times (20+1) \Rightarrow \frac{21}{4} \Rightarrow 5.25$

homework

Why $n+1$

What is the 75^{th} percentile ?

Five number Summary

1. Min
2. First Quartile (25%ile) Q1
3. Median (50%ile) Q2
4. Third Quartile (75%ile) Q3
5. Maximum.

IQR [Inter Quartile range]

$= Q_3 - Q_1$

Outlier

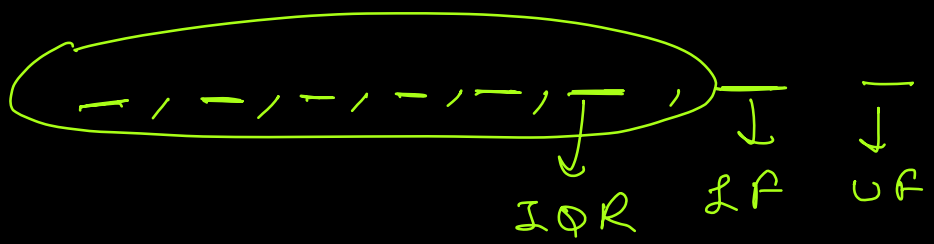
lower fence $= Q_1 - 1.5\,(IQR)$

upper fence $= Q_3 + 1.5\,(IQR)$

$Q_1 = \frac{25}{100} \times 13$

Q_1 Q_3

55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 105



IQR

55, 65, 80, 95, 105, 30, 20, 140

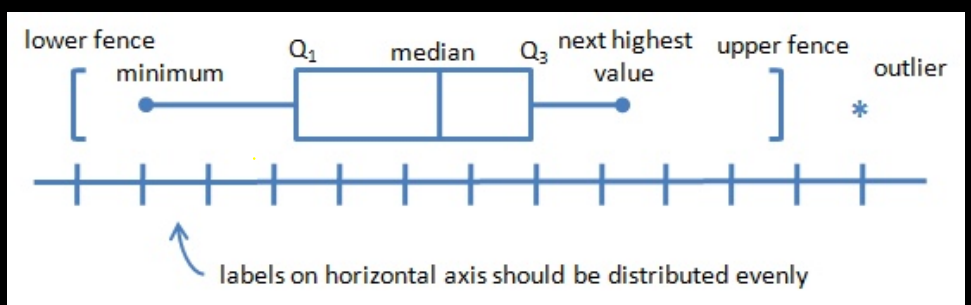
Q_3

$LF = 65 - 1.5(30)$

$= 65 - 45 = 20$

$UF = 95 + 1.5(30)$

$= 95 + 45 \Rightarrow 140$



Types Of Graphs

1] Normal / Gaussian Distribution

2] Standard Normal Distribution

3] Log Normal Distribution

4] Power law distribution

5] Bernaulis

Hypothesis

Rank Interpolation

60, 65, 70, 75, 80, 85, 90,
95, 100, 105, 110

$$\frac{30^{th}\ \%ile}{100} \times (12+1)$$

$$= 3.6$$

3.6 → 75.0 & 70

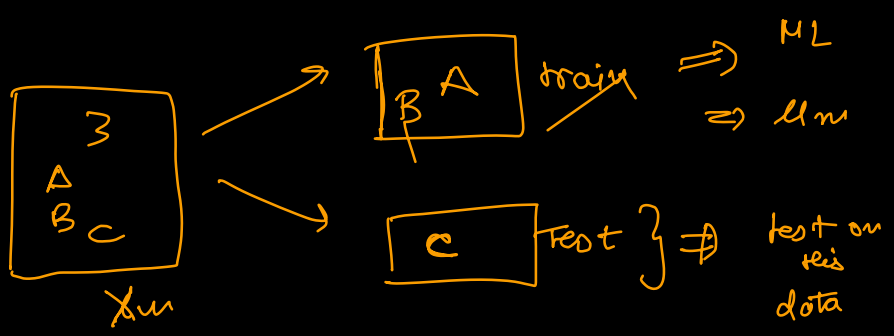
70 → 75

0.6 | 0.4

$0.4 \times 70 + 0.6 \times 75$

$= 28 + 45$

$= 73$ exact


3
A
B C
Xm
A
B
train
ML
Mn
C
Test
test on
this
data