19UCC023 Mohit Akhouri

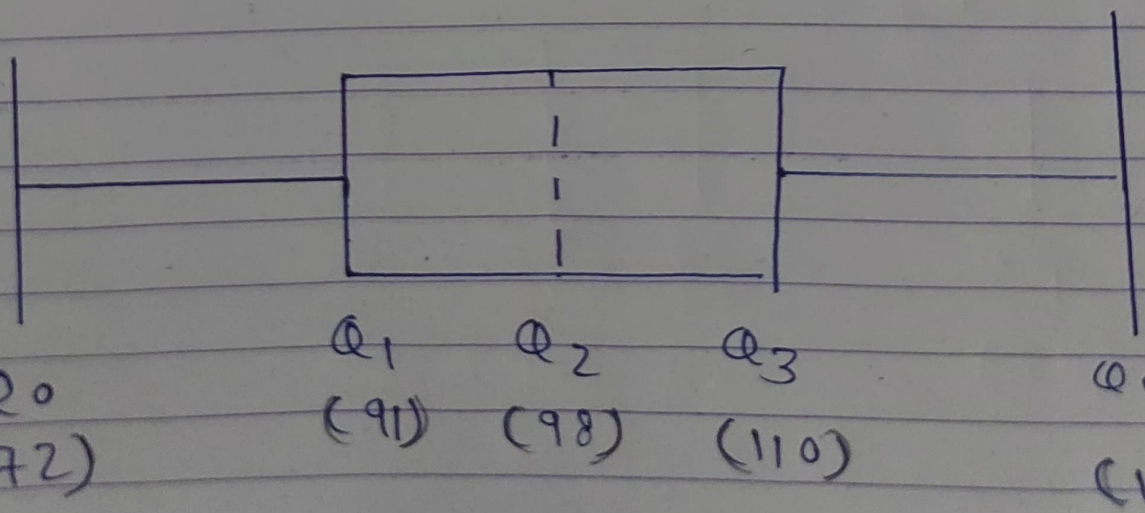
IDS Assignment- 1

Q1) - - - - - - - - - - - - - - -

(a) - - - - - - - - - - - - - - -

Soln: minimum (Q_0) = 72
Lower Quartile (Q_1) = 91
Median (Q_2) = 98
Upper Quartile (Q_3) = 110
Maximum (Q_4) = 130

Box & Whisker Plot is:



19UCC023 – Mohit Akhouri

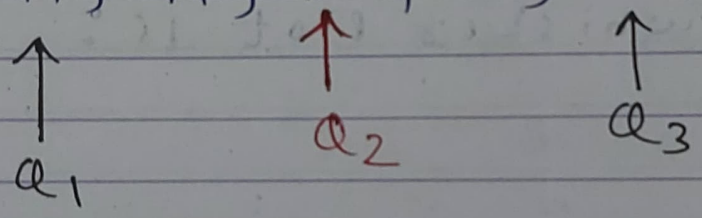
(b.) – – – – – – – – – – – – – –

Soln:

Sample = 25, 15, 11, 90, 14, 20, 10

In Ascending order:

10, 11, 14, 15, 20, 25, 90



$Q_2 = 15$ (median)

$Q_1 = 11$ ~~(First 25%)~~

$Q_3 = 25$

$I.Q.R = Q_3 - Q_1 = 25 - 11 = 14$

Calculating outlyers:

$Q_1 - 1.5\ IQR = 11 - 1.5 \times 14 = -10$

$Q_3 + 1.5\ IQR = 25 + 1.5 \times 14 = 46$

$= [-10, 46]$ = outlyers

Igucc023 – Mohit Akhouri

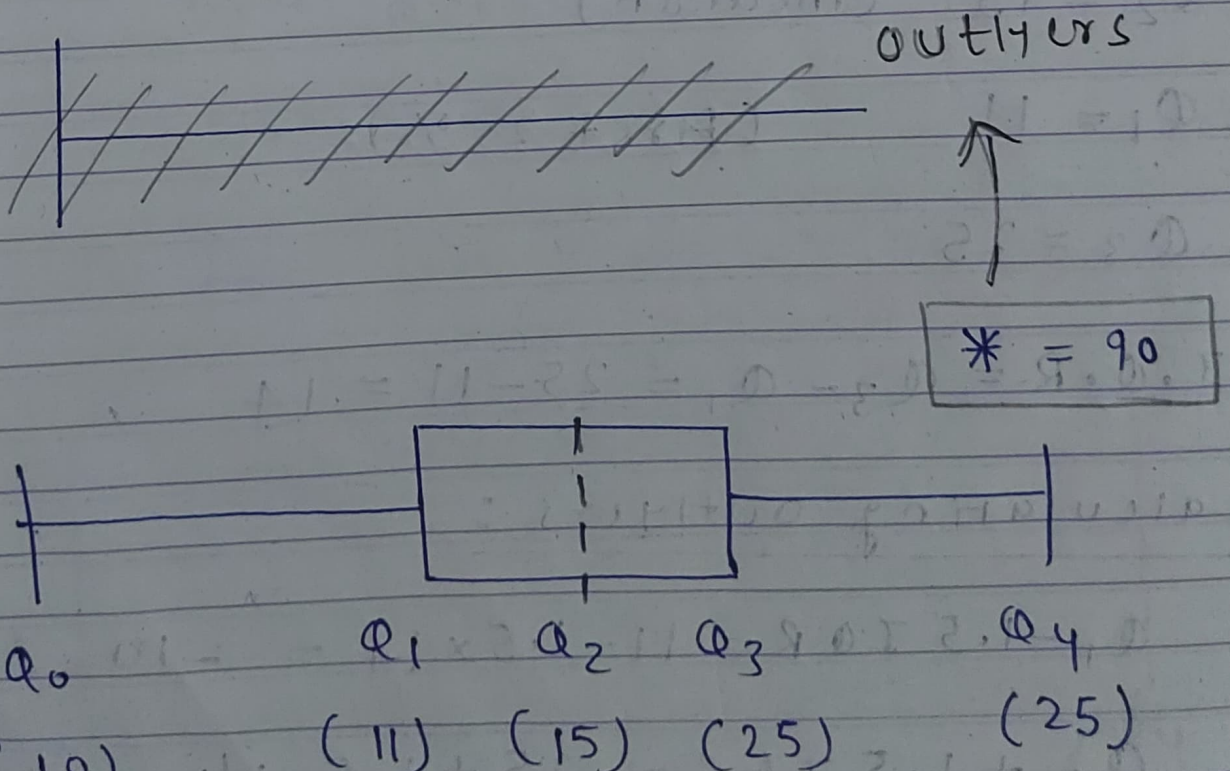
minimum (Q_0) after removing outliers

$Q_0 = 10$

maximum (Q_4) after removing outliers

$Q_4 = 25$

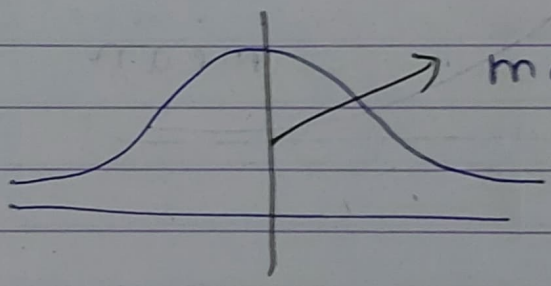
So the box-whisker plot is:



19UCC023- Mohit Akhouri

Q2) — — — — — — — — — —

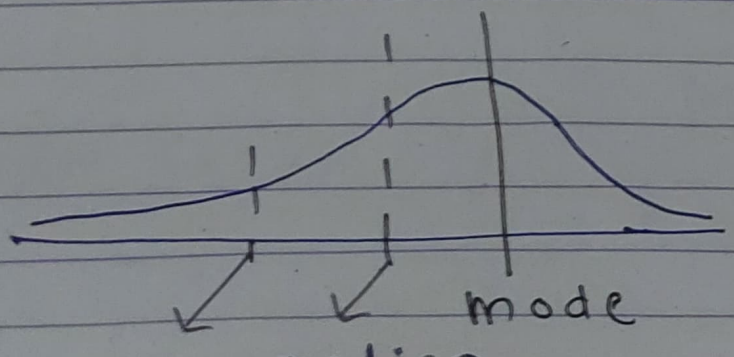
Soln: If the distribution is symmetrical, all three parameters mean, median & mode are the same. (If the distribution has only 1 mode, then it is equal to mean and median, but not in case of multi-mode distribution)



LEFT-SKEWED:

* If the distribution is negatively skewed (left skewed), then ~~mean~~ relationship between mean, median and mode is:

mean < median < mode

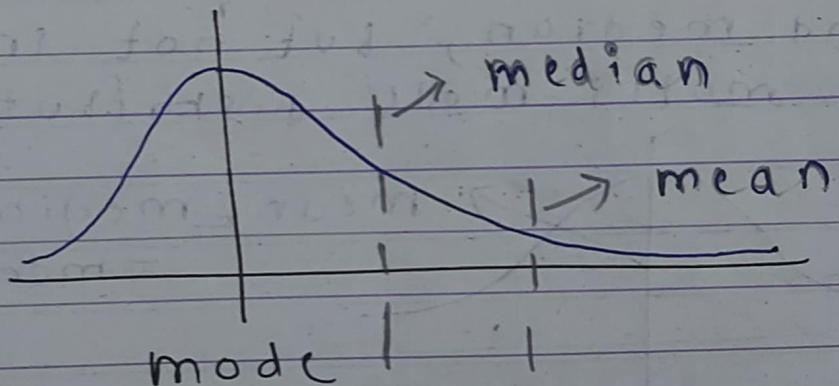


19UCC023 – Mohit Akhoumi

RIGHT-SKEWED:

If the distribution is right-skewed (positively skewed), then relationship between mean, median and mode is:

mode < median < mean



19UCC023- Mohit Akhouri

Q3) — + — + — — — + — + — + — -

Soln:

* Probability of head = P(H)

$$P(H) = \frac{40}{100} = \left(\frac{2}{5}\right) = p$$

* no. of times coin flipped (50)

n

Coin follows binomial distribution $\sim B(n, p)$

$\sim B\left(50, \frac{2}{5}\right)$

Now, mean of binomial random variable is:

if $X \sim B(n, p)$

then $E(X) = np$

19UCC023- Mohit Akhouri

X: tossing the coin and getting head

$n = 50$

$p = 2/5$

$X \sim B(n, p)$

~~$X \sim B(50, 2/5)$~~ $X \sim B(50, 2/5)$

Here, $a = 27$

By Markov's Inequality,

$$P(X \geq a) \leq \frac{E(X)}{a}, \quad a > 0, \quad a \in R$$

$$\Rightarrow P(X \geq 27) \leq \frac{E(X)}{27}$$

$$E(X) = np = 50 \times 2/5 = 20$$

$$\Rightarrow P(X \geq 27) \leq \frac{20}{27}$$

So, the upper bound on probability $= \frac{20}{27}$

19UCC023 - Mohit Alchouri

Q4) - - - - - - - - - - - -

Soln:

$n = 50$

$p = 1/2$ (since it is a fair coin)

~~X = coin is tossed~~

coin follows binomial distribution

Let X = random variable

~~X = coin is tossed & we get head~~

X = we get head on toss of coin

$X \sim B(50, 1/2)$

We have to find probability:

$$P(X < 20 \cup X > 30)$$

$(X < 20) \cup (X > 30)$ can also be written as: $|X - 25| \geq 5$

19UCC023 – Mohit Akhouri

$X \sim B(50, \frac{1}{2})$ (n, p)

$E(X) = np = 50 \times \frac{1}{2} = 25$

$Var(X) = np^2 = 50 \times \frac{1}{2} \times \frac{1}{2} = 12.5$

~~$P(X \geq 2)$~~

$P(X < 20 \cup X > 30) = P(|X - 25| \geq 5)$

So by Chebyshev's Inequality,

$$P(|X - M| \geq k) \leq \frac{Var(X)}{k^2}$$

M = mean, $k \in R$, $k > 0$

So, $P(|X - 25| \geq 5) \leq \frac{Var(X)}{5^2}$

$\leq \frac{12.5}{25}$

$$P(|X - 25| \geq 5) \leq 0.5$$

So, the upper bound on probability = 0.5

19UCC023- Mohit Akhouri

Q5) - - - - - - - - - - - - -

Soln:

$\mu = 7$

(a) - - - - - - - - - - - - -

Soln: ~~X = student gets marks~~

X = marks of students

~~X~~ X follows normal distribution with mean μ & variance σ^2

$X \sim N(\mu, \sigma^2)$ (By central limit theorem)

Here $\mu = 7$

Now, we have to find probability that $X \geq 8$

$P(X \geq 8) \leq \frac{E(X)}{8}$ (By Markov's Inequality

$E(X) = \mu = 7$

$P(X \geq 8) \leq 7/8$

Upper bound on Probability $= 7/8 = 0.875$

.5

19UCC023- Mohit Akhouri

(b) - - - - - - - - - - - -

Soln:

Given $\sigma = 2, \mu = 7$

Now, we have to find lower bound on probability

$P(5 \leq X \leq 9)$

$P(5 < X < 9) = P(|X-7| \leq 2)$

$P(|X-7| \geq 2) \leq \frac{Var(X)}{2^2}$ (By chebyshev's inequality)

replacing $P(|X-7| \geq 2)$ with $P(|X-7| \leq 2)$

$1 - P(|X-7| \leq 2) \leq \frac{Var(X)}{2^2}$

$P(|X-7| \leq 2) \geq 1 - \frac{Var(X)}{2^2}$

$P(|X-7| \leq 2) \geq 1 -$

$Var(X) = \sigma^2 = 4$

19UCC023 – Mohit Akhouri

So,

$$P(|x-7| \leq 2) \geq 1 - \frac{4}{2^2}$$

$$P(|x-7| \leq 2) \geq 1 - 0$$

$$P(|x-7| \leq 2) \geq 1$$

So, the lower bound on Probability is 1

PTO ⟹

19UCC023 - Mohit Akhouri

Q6) - - - - - - - - - - - -

Soln:

Sample x is drawn from population with mean 'μ' & SD 'σ'

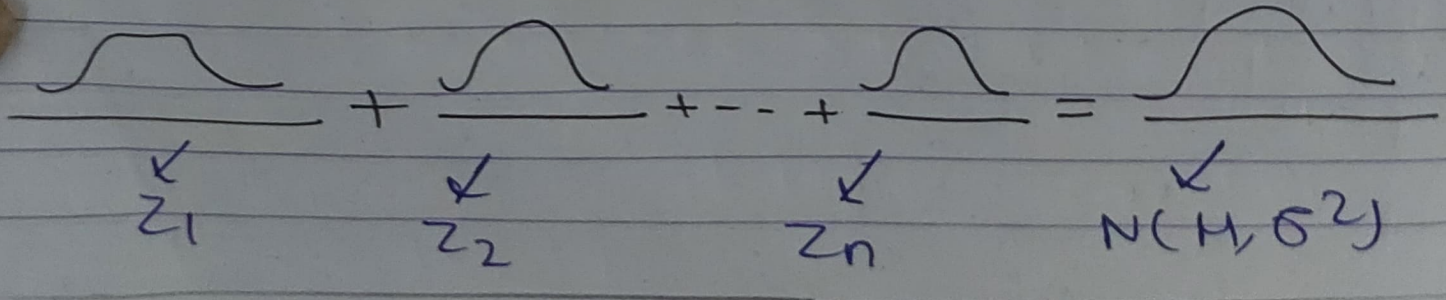
mean of the sample $\bar{X}$

$$\bar{X} = \frac{x_1 + x_2 -- x_n}{n}$$

$$\bar{X} = \frac{\mu n}{n} = \mu$$

Variance of the sample $\bar{X} = \sigma/\sqrt{n}$

Now, if we take all such samples with mean μ & SD σ, and we combine them all, we get again a distribution which follows normal distribution due to "central limit theorem"

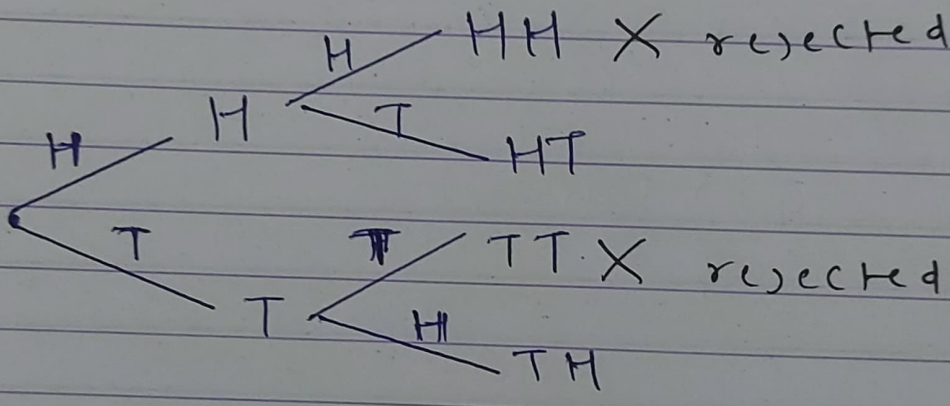


19ucc023 - Mohit Akhouri

Q7) - - - - - - - - - - - -

Soln: To simulate a fair coin using a biased coin, we can follow the steps given below:

(i) Toss the coin twice, we get four possible results



(ii) Now, if we get HH or TT, reject the move and try again.

(iii) Now if we get ~~HT, TH~~ HT or TH, we keep (acknowledge) the first result and reject the second result.

This is how, we ~~simulate~~ simulate a fair coin using biased coin.

19UCC023 - Mohit Akhouri

Q8) - - - - - - - - - - - -

Soln:

X_1 = samples of plant 1

X_2 = samples of plant 2

X_3 = samples of plant 3

n_1 = no. of samples in $X_1 = 10$

n_2 = no. of samples in $X_2 = 10$

n_3 = no. of samples in $X_3 = 10$

$\bar{X}_1 = \sum_{j=1}^{10} X_{1j} / 10 = 10.4$

$\bar{X}_2 =$

$\bar{X}_1 = 10.4$

$\bar{X}_2 = \frac{93.3}{10} =$

19UCC023 - Mohit Akhouri

$$\bar{X}_1 = \frac{\Sigma x_1}{n_1} = \frac{104}{10} = 10.4$$

$$\bar{X}_2 = \frac{\Sigma x_2}{n_2} = \frac{93.3}{10} = 9.33$$

$$\bar{X}_3 = \frac{\Sigma x_3}{n_3} = \frac{103.49}{10} = 10.35$$

$$\bar{\bar{X}} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3}{3} = \frac{30.08}{3} = 10.02$$

null hypothesis : H_0

$H_0 : M_1 = M_2 = M_3$

$\alpha = 0.05$

$$SSC = \sum_{i=1}^{3} (\bar{X}_i - \bar{\bar{X}})^2 * n_i$$

$$= (10.4 - 10.02)^2 \times 10 + (9.33 - 10.02)^2 \times 10 + (10.35 - 10.02)^2 \times 10$$

$$= 7.294$$

19UCC023

~~19UCC023~~ - Mohit Akhouri

$$SSE = \sum_{i=1}^{3} (var(X_i) * (n_i - 1))$$

$$= [var(x_1) + var(x_2) + var(x_3)] \times 9$$

$$= [3.11 + 1.68 + 1.06] \times 9$$

$$= 52.65$$

$df_{sample} = n - 1 = 10 - 1 = 9$

$df_{error} = N - (k)$ → Total ~~SSE~~ no. of samples

$= (10 + 10 + 10) - 3$

$= 27$

$$MSC = \frac{SSC}{df_{sample}} = \frac{7.294}{9} = 0.8104$$

$$MSE = \frac{SSE}{df_{error}} = \frac{52.65}{27} = 1.95$$

$$F_{stat} = \frac{MSC}{MSE} = \frac{0.8104}{1.95} = 0.4155$$

$F_{\alpha, df_{col}, df_{within}}$

$df_{col} = 9$, $df_{within} = 27$

$F_{critical} = F_{0.05, 9, 27} = 2.2501$

$F_{stat} = 0.4155$

$F_{critical} = 2.2501$

Since $F_{stat} < F_{critical}$,

null hypothesis H_0 "failed to reject"

So, there are not enough evidences to reject the claim that all three manufacturing plants are similar.

19UCC023- Mohit Akhoumi

Q9) - - - - - - - - - - - - -

Soln:

$n = 325$

$\bar{X} = 225.27$

$\sigma = 0.765$

$\alpha = 0.01$

$\mu = 222$

Step 1: Claim: $\mu > 222$

Opposite: $\mu \leq 222$

Step 2:

$H_0 : \mu = 222$

$H_1 : \mu > 222$

Step 3: $\alpha = 0.01$

Step 4:

Calculate Z-statistics since sample size (n) ≥ 30 and population SD is given

$$Z_{stat} = \frac{\bar{X} - M}{\sigma/\sqrt{n}}$$

$\bar{X} = 225.27$

$M = 222$

$\sigma = 0.765$

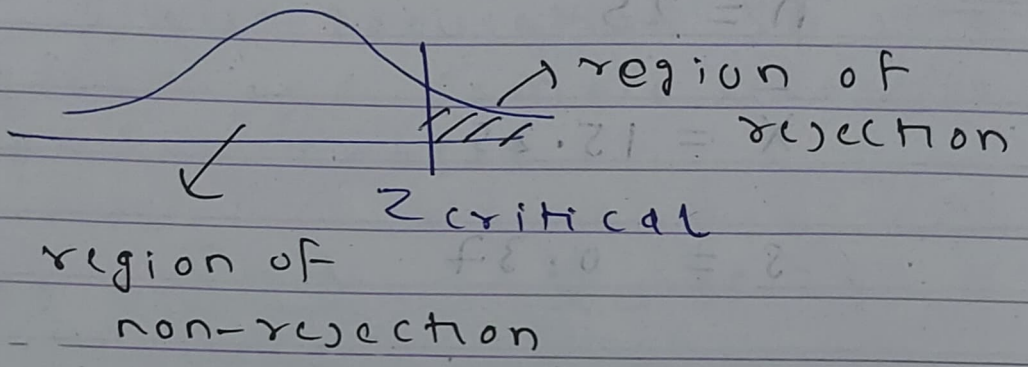
$n = 325$

$$= \frac{225.27 - 222}{0.765/\sqrt{325}}$$

$$= \frac{3.27}{0.042} = 77.85$$

$$Z_{stat} = 77.85$$

Step 5: Since H_1 has '>' sign, go for right tailed test



$Z_{critical} = 2.33$

Since $Z_{stat} > Z_{critical}$,

$\Rightarrow$ "Ho is rejected"

$\Rightarrow$ "alternate hypothesis failed to reject"

Step 6: Interpretation

There is enough evidences to support the claim that mean volume is greater than 222 mL.

19UCC023- Mohit Akhoumi

Q10) _ _ _ _ _ _ _ _ _ _

Soln:

$n = 25$

$\bar{X} = 15.25$

$s = 0.37$

$\mu = 15$

$\alpha = 0.01$

Step 1:

Claim: $\mu > 15$

opposite: $\mu \leq 15$

Step 2:

$H_0: \mu = 15$

$H_1: \mu > 15$

19UCC023 - Mohit Akhouri

STEP 3: $\alpha = 0.01$

STEP 4: Since population SD is not given and sample size < 30, go for T-test.

$$T_{stat} = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

$\bar{X} = 15.25$

$\mu = 15$

$S = 0.37$

$n = 25$

$$= \frac{15.25 - 15}{0.37/\sqrt{25}}$$

$$= \frac{0.25}{0.074} = 3.378$$

STEP 5: Since H_1 has '>' sign, go for one-tailed test (right tailed test)

$df = n - 1 = 25 - 1 = 24$

$\alpha = 0.01$

$T_{0.01, 24} = 2.4922 = T_{critical}$

one-tailed test

Step 6: $T_{stat} = 3.378$, $T_{critical} = 2.492$

Since $T_{stat} > T_{critical}$

$\Rightarrow$ "H_0 is rejected"

$\Rightarrow$ H_1 failed to reject (accepted)

Step 7:

There are enough evidences to support that claim that mean mean weight > 15 kg

19UCC023 – Mohit Akhoumi

Q11) – – – – – – – – – –

Soln:

Different types of probabialistic sampling techniques are:

(i) Simple random sampling
(ii) Systematic sampling
(iii) Stratified sampling
(iv) Cluster sampling

(i) Simple random sampling: It can be of two types:

(a) Sampling with replacement
(b) Sampling without replacement

In sampling with replacement, same object may be picked multiple times but probability of picking all samples remains same.

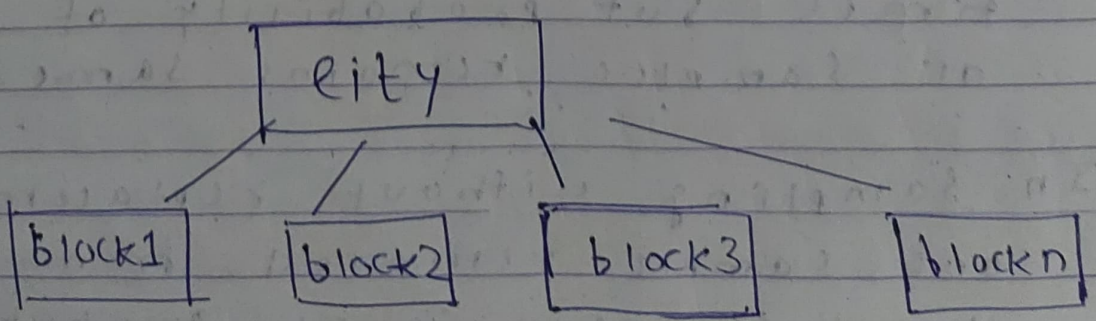
In sampling without replacement, one sample removed from the total samples, so the probability

of choosing samples gets changed reduced by 1. each time.

(iii)

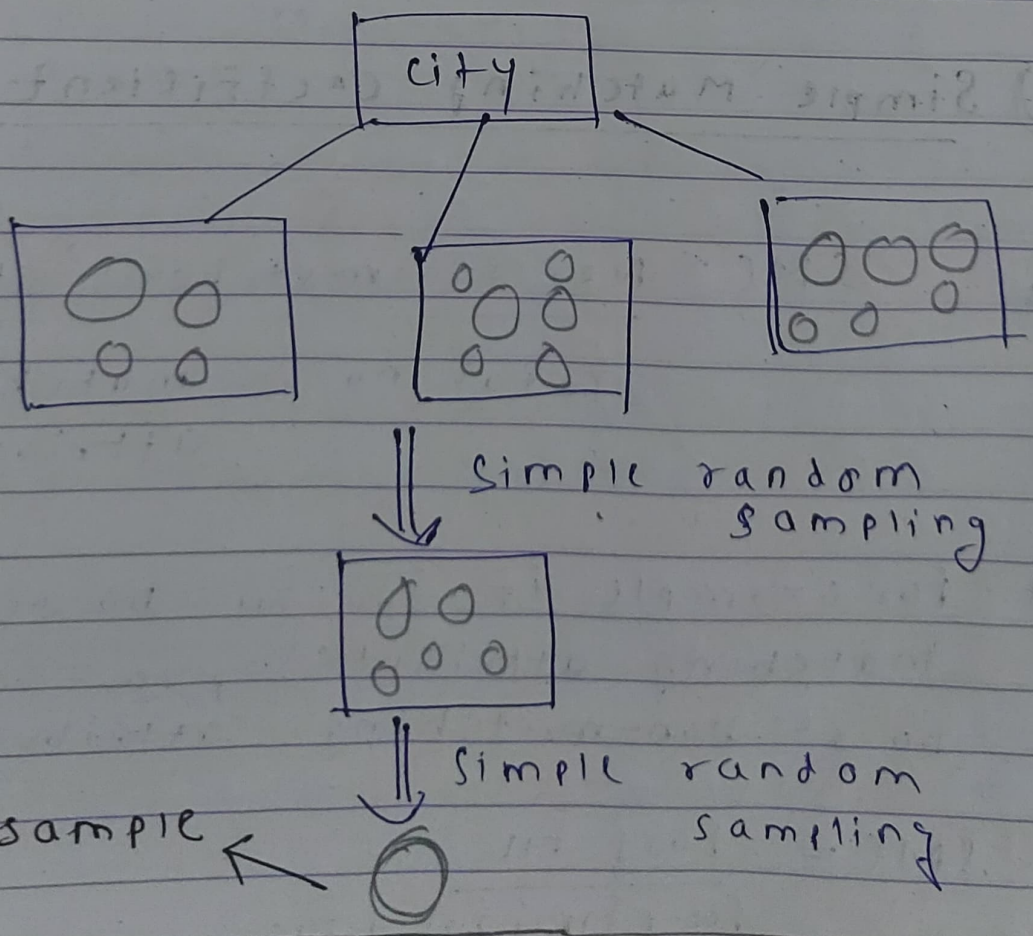
(ii) Systematic Sampling: In this sampling technique, we generate a uniformly distributed random number $R = U[0,1]$, then we find $i = \lceil R \times N \rceil$. Then we choose samples as X_{ki} where $ki = N$ and N = size of population

(iii) Stratified Sampling: In this sampling technique, the entire population is divided into homogenous subgroups called stratum, then we apply simple random sampling on each group.



In the above example, city is divided into blocks, then we apply simple random sampling on each block.

(iv) **Cluster sampling:** In this sampling technique, the entire population is divided into clusters, then we select a cluster according to simple random sampling, then we apply simple random sampling within that cluster also to select samples.



19UCC023 - Mohit Akhouri

Q12) - - - - - - - - - - - - -

Soln:

SIMILARITY:

Similarity is defined as the measure of the degree to which two objects are alike.

Similarity measure value range is between 0 and 1.

$$S \in [0, 1]$$

* MEASURES OF SIMILARITY:

(i) Simple Matching coefficient (SMC)

* $SMC = \frac{\text{No. of matching attributes}}{\text{Total no. of matching attributes}}$

* For example if f_{00}, f_{11} = no. of matching attributes, $f_{01} = f_{10}$ = no. of non-matching attributes

$$SMC = \frac{f_{00} + f_{11}}{f_{00} + f_{11} + f_{01} + f_{10}}$$

19UCC023 - Mohit Akhouri

(ii) Jaccard similarity coefficient :

In this, the term f_{00} is rejected since "it has no effect on the measure of similarity between two objects"

$$JSC = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

JSC = Jaccard similarity coefficient

(iii) Cosine similarity

⇒ it is a measure of similarity between 2 vectors of an inner product space.

⇒ it measures cosine of the angle between two vectors

$$\Rightarrow \cos\theta = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

19UCC023 - Mohit Akhouri

DISSIMILARITY:

It is the numerical measure of degree to which two objects are different

It also has a value between 0 and 1 and is related to the similarity coefficient by the following formula:

$$d = 1 - S \quad , d \in [0,1]$$

Dissimilarity is measured using distance between 2 objects (feature values) which are:

Minkowski distance

$$= \left(\sum_{i=1}^{n} |x_i - y_i|^r \right)^{1/r}$$

r = parameter

- $r = 1$ → Manhattan distance: $\sum_{i=1}^{n} (x_i - y_i)$
- $r = 2$ → Euclidean distance: $\sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$
- $r = \infty$ → Supremum distance (max. diff between any attributes of 2 objects)

19UCC023
~~19UCC023~~ - Mohit Akhouri

Uses of these measures:

These similarity & dissimilarity measures are used in various domains of data science like:

1. TEXT-MATCHING
2. CLUSTERING
3. NEAREST NEIGHBOUR

PTO ⟹

19UCC023 - Mohit Akhouri

Q13) — — — — — — — — — —

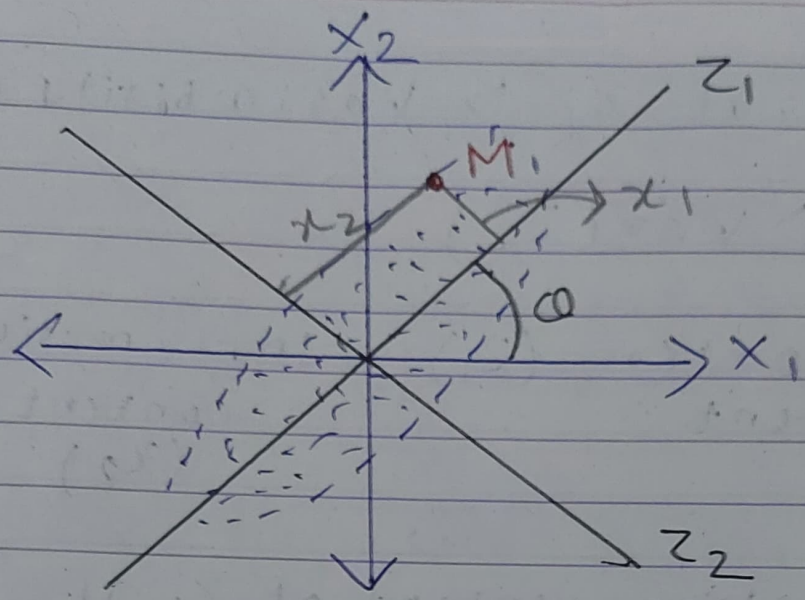
Soln: In dimensionality reduction, we reduce the number of dimensions to less no. of ~~dimension~~ dimensions for:

1. reducing decision taking cost
2. increasing information content
3. reducing complexity of algorithm used in data science.

PCA (Principal component Analysis)

This is the most common algorithm of Dimensionality reduction. In this we transform the values from $x_1 - x_2$ plane to $z_1 - z_2$ plane.

19UCC023 – Mohit Akhouri

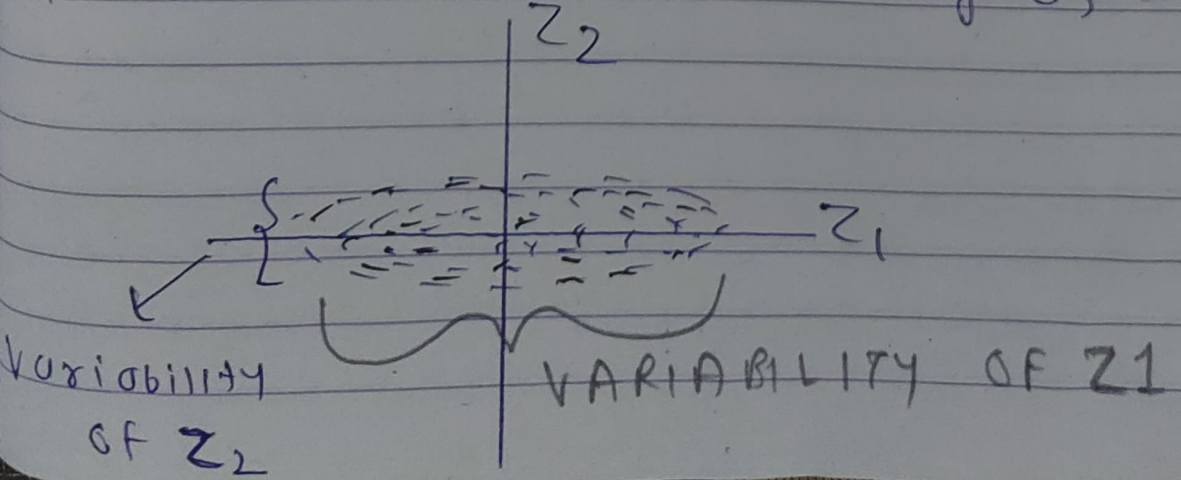


Here X_1 & X_2 are positively correlated and their projection can be found on $Z_1 - Z_2$ plane as follows:

$$\begin{cases} Z_1 = x_1 \cos\theta + x_2 \sin\theta \\ Z_2 = -x_1 \sin\theta + x_2 \cos\theta \end{cases}$$

PROJECTION OF point 'm' on $Z_1 - Z_2$ plane

If we rotate the axis by θ, we get:



variability of Z_1 > variability of Z_2

So, $var(Z_1) > var(Z_2)$

So, information (Z_1) content > Information content (Z_2)

Since information content of Z_1 is more, we need only variable Z_1 to take decision.

So, we reduced the dimension from 2 to 1.

This is known as "DIMENSIONALITY REDUCTION"