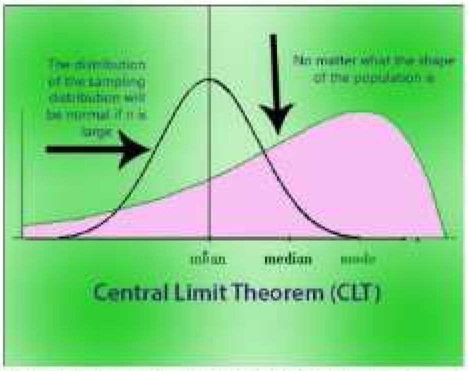
UT 1 Solution

1 a) Illustrate in detail the Central limit theorem.

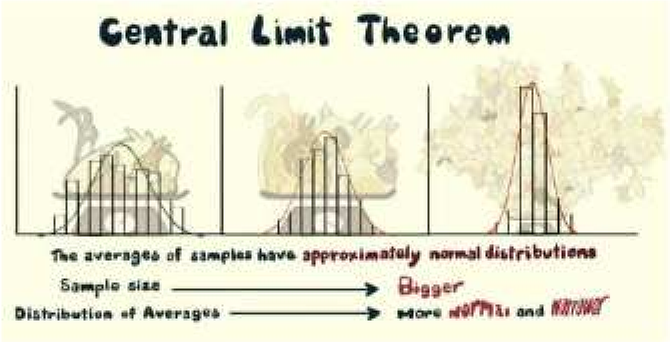
CENTRAL LIMIT THEOREM:

The central limit theorem (CLT) states that the distribution of a sample variable approximates a normal distribution (i.e., a "bell curve") as the sample size becomes larger, assuming that all samples are identical in size, and regardless of the population's actual distribution shape.

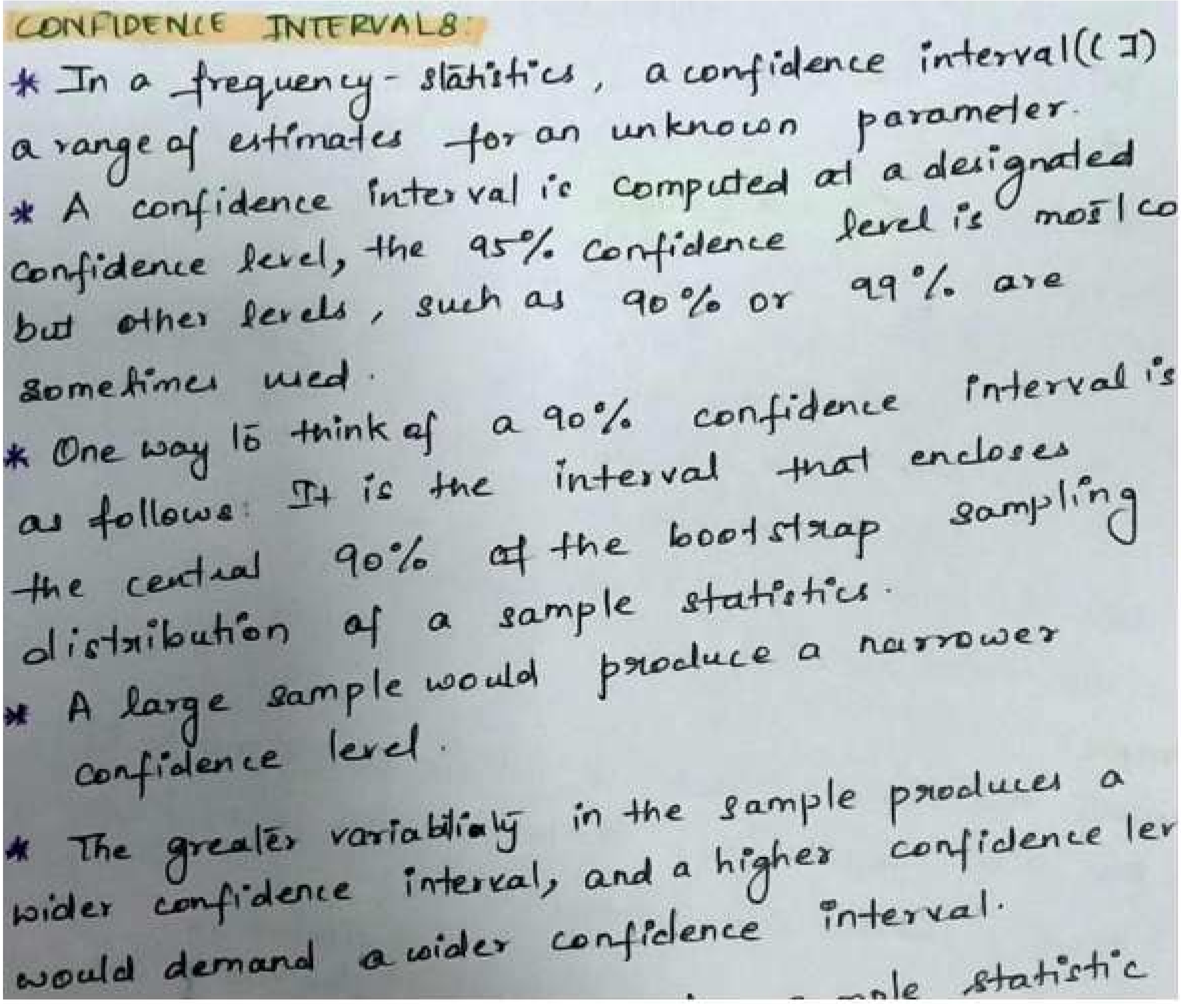

The distribution of the sampling distribution will be normal if n is large
No matter what the shape of the population is
mean
median
mode
Central Limit Theorem (CLT)

CLT is a statistical premise that, given a sufficiently large sample size from a population with a finite level of variance, the mean of all sampled variables from the same population will be approximately equal to the mean of the whole population. Furthermore, these samples approximate a normal distribution, with their variances being approximately equal to the variance of the population as the sample size gets larger, according to the law of large numbers.

- The central limit theorem (CLT) states that the distribution of sample means approximates a normal distribution as the sample size gets larger, regardless of the population's distribution.
- Sample sizes equal to or greater than 30 are often considered sufficient for the CLT to hold.
- A key aspect of CLT is that the average of the sample means and standard deviations will equal the population mean and standard deviation.
- A sufficiently large sample size can predict the characteristics of a population more accurately.


Central Limit Theorem
The averages of samples have approximately normal distributions
Sample size → Bigger
Distribution of Averages → More normal and narrower

b) Define Confidence Interval and Calculate the range of heights (95% confidence level) for the given population. The mean = 175cm, SD = 20cm, sample size = 40 and z=1.960.



CONFIDENCE INTERVALS:

* In a frequency- statistics, a confidence interval(CI) a range of estimates for an unknown parameter.

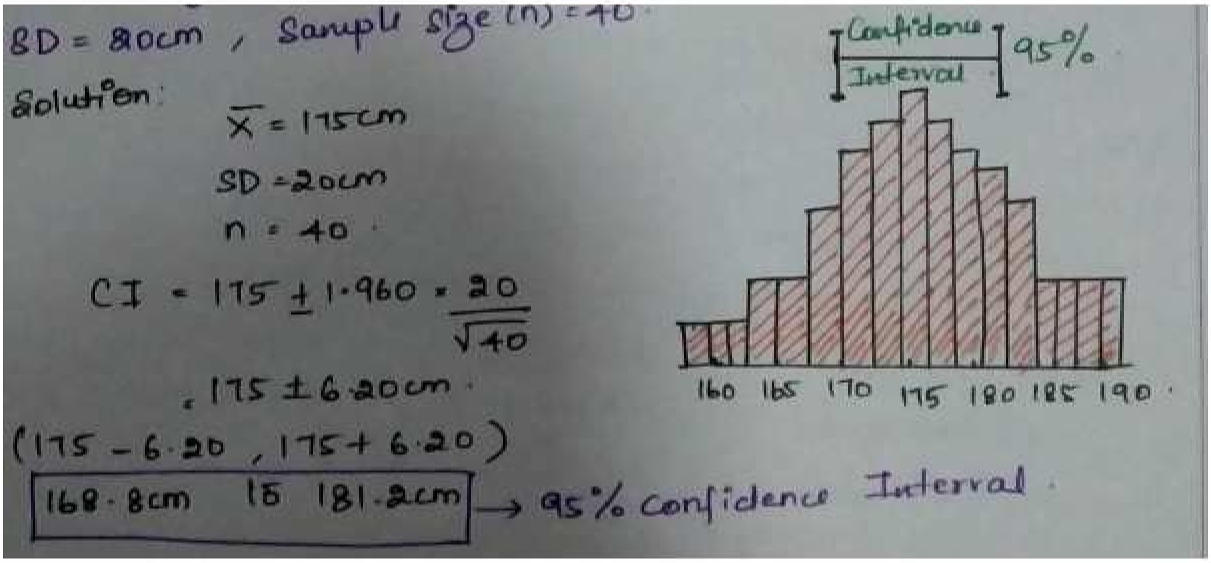
* A confidence interval is computed at a designated confidence level, the 95% confidence level is most co but other levels, such as 90% or 99% are sometimes used.

* One way to think of a 90% confidence interval is as follows: It is the interval that encloses the central 90% of the bootstrap sampling distribution of a sample statistics.

* A large sample would produce a narrower confidence level.

* The greater variability in the sample produces a wider confidence interval, and a higher confidence lev would demand a wider confidence interval.

mple statistic



SD = 80cm , Sample size (n) = 40.

Solution:

$\bar{x} = 175\text{cm}$

SD = 20cm

n = 40.

$$CI = 175 \pm 1.960 \times \frac{20}{\sqrt{40}}$$

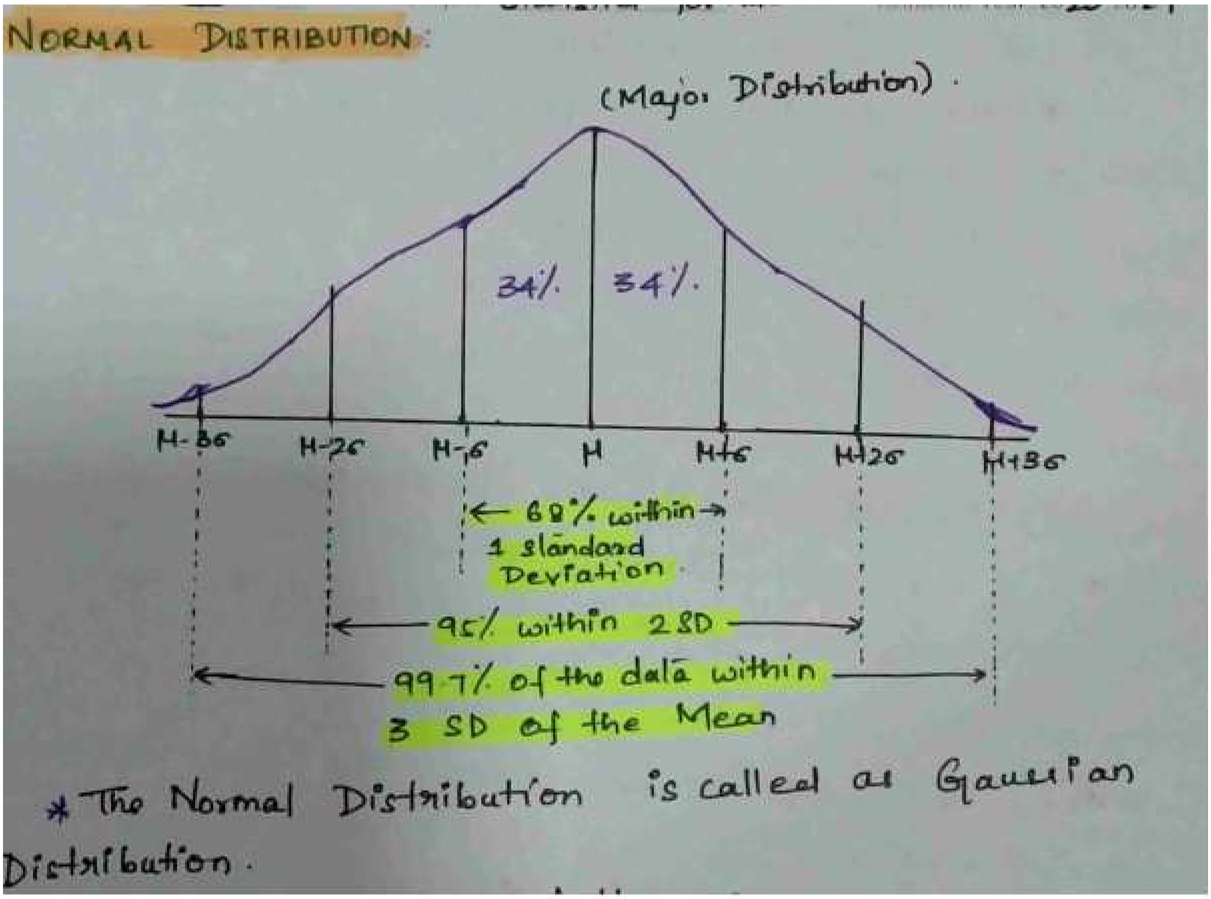
$$= 175 \pm 6.20\text{cm}.$$

(175 − 6.20 , 175 + 6.20)

$\boxed{168.8\text{cm} \text{ to } 181.2\text{cm}}$ → 95% confidence Interval.

Confidence Interval
95%
160
165
170
175
180
185
190

c) Describe Normal Distribution.



NORMAL DISTRIBUTION:

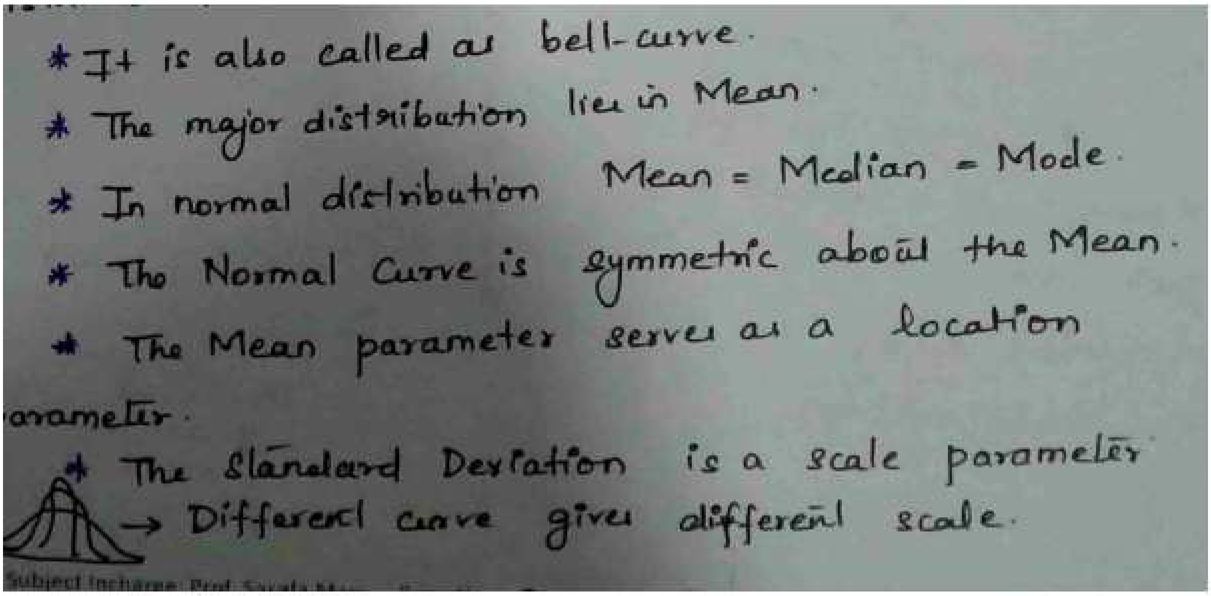
(Major Distribution).

← 68% within → 1 standard Deviation.

← 95% within 2SD →

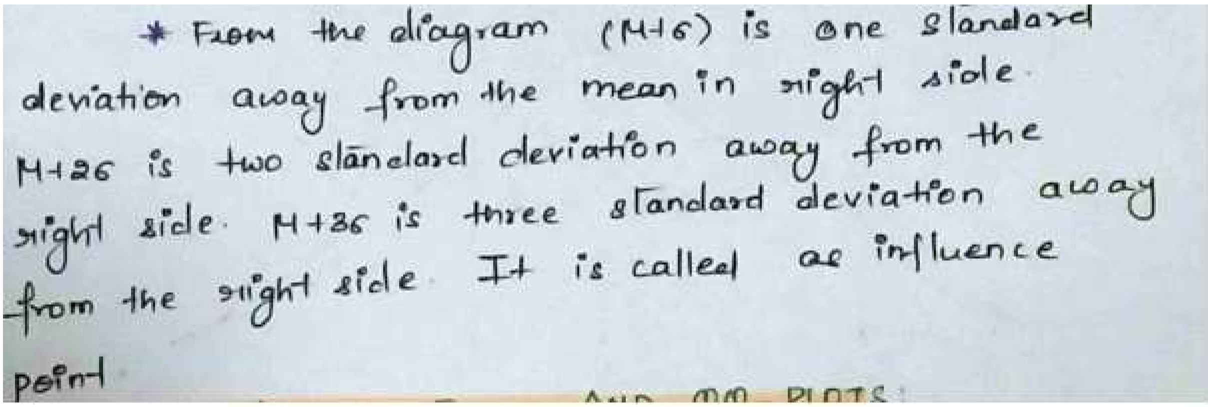
← 99.7% of the data within 3 SD of the Mean →

- * The Normal Distribution is called as Gaussian Distribution.



- * It is also called as bell-curve.
- * The major distribution lies in Mean.
- * In normal distribution Mean = Median = Mode.
- * The Normal Curve is symmetric about the Mean.
- * The Mean parameter serves as a location parameter.
- * The Standard Deviation is a scale parameter
  → Different curve gives different scale.

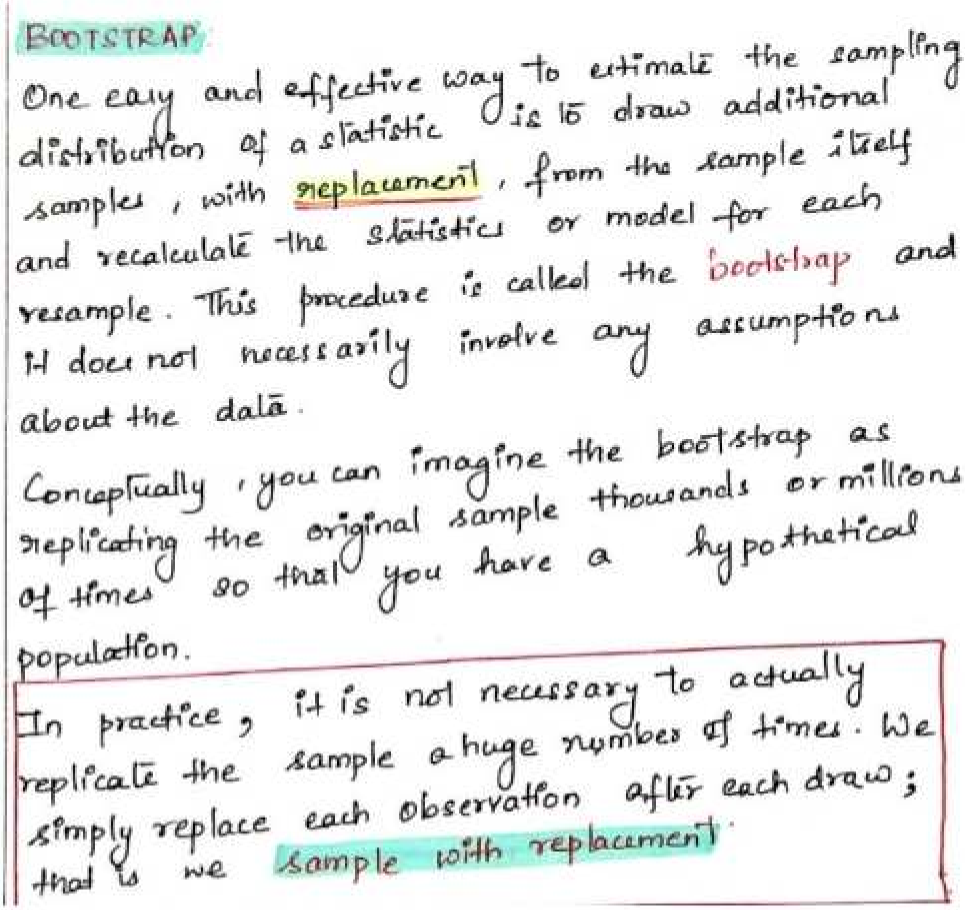
Subject Incharge: Prof. Sarala M



* From the diagram (M+6) is one standard deviation away from the mean in right side. M+26 is two standard deviation away from the right side. M+36 is three standard deviation away from the right side. It is called as influence point

AND QQ PLOTS:

d) Discuss Bootstrapping algorithm.

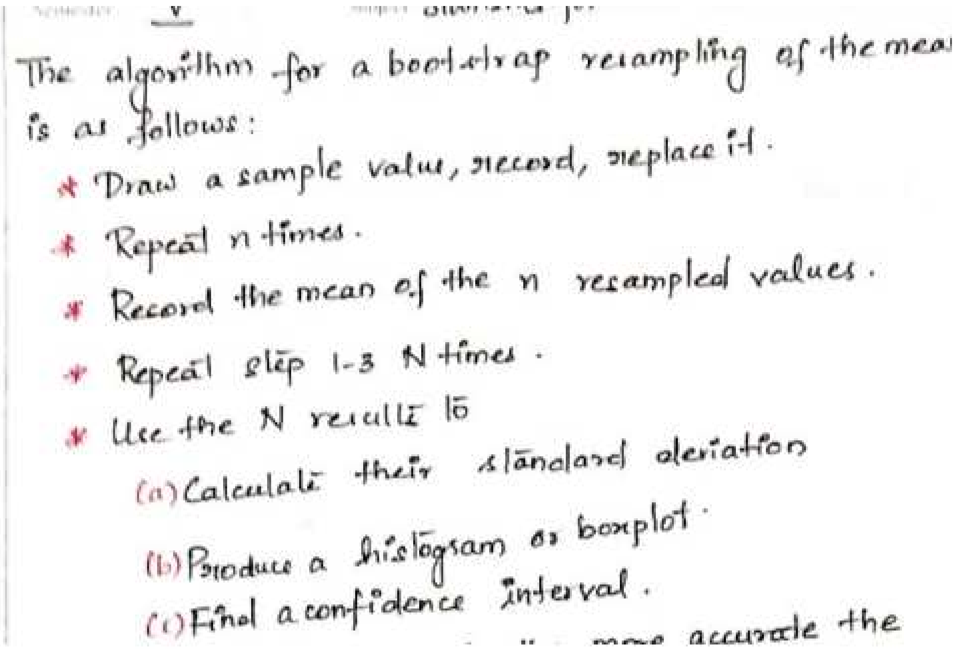


BOOTSTRAP:

One easy and effective way to estimate the sampling distribution of a statistic is to draw additional samples, with replacement, from the sample itself and recalculate the statistics or model for each resample. This procedure is called the bootstrap and it does not necessarily involve any assumptions about the data.

Conceptually, you can imagine the bootstrap as replicating the original sample thousands or millions of times so that you have a hypothetical population.

In practice, it is not necessary to actually replicate the sample a huge numbers of times. We simply replace each observation after each draw; that is we sample with replacement.



The algorithm for a bootstrap resampling of the mea

is as follows:

- Draw a sample value, record, replace it.
- Repeat n times.
- Record the mean of the n resampled values.
- Repeat step 1-3 N times.
- Use the N results to

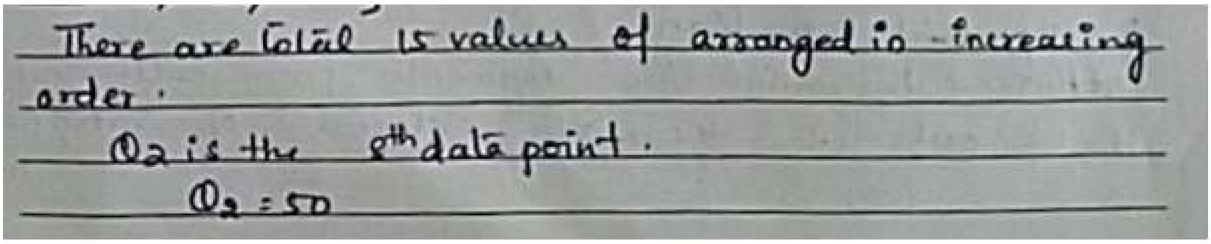
  (a) Calculate their standard deviation

  (b) Produce a histogram or boxplot.

  (c) Find a confidence interval.

more accurate the

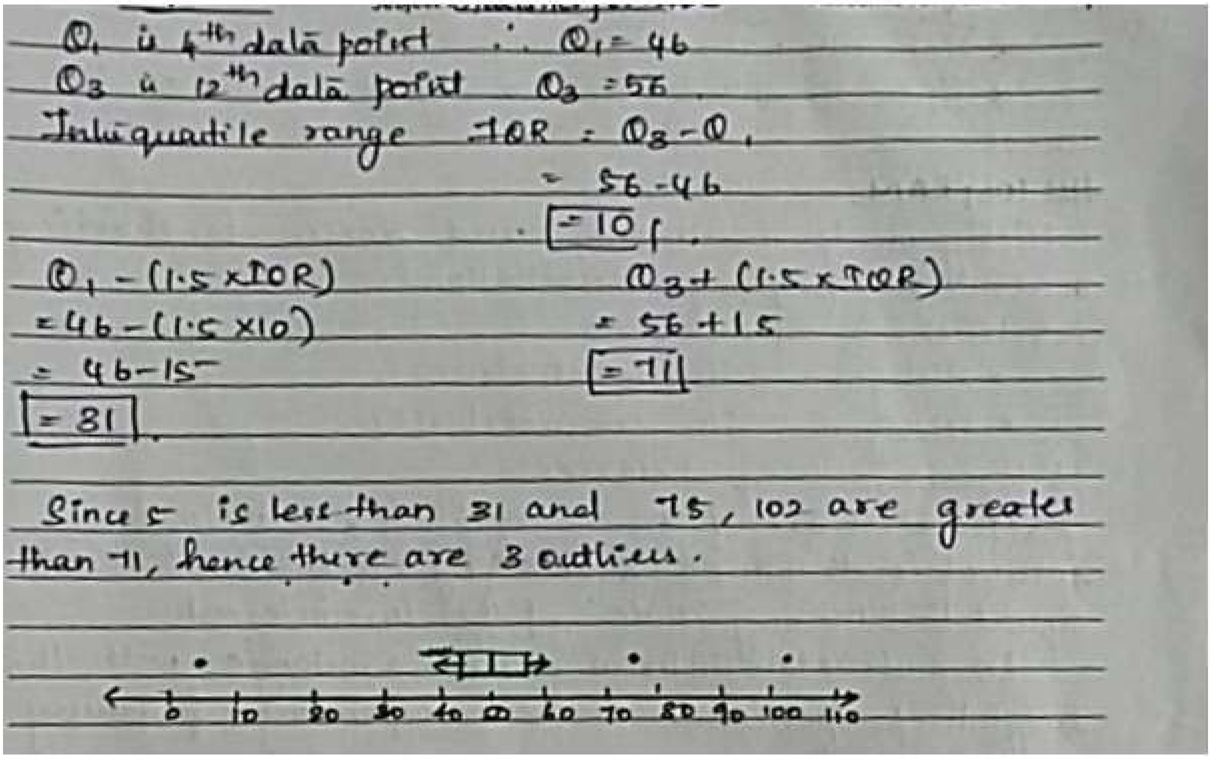
Q2 a) Find Q1, Q2 and Q3 for the following dataset. Identify the outliers and draw a box and whisker plot. {5, 40, 42, 46, 48, 49, 50,50, 52, 53, 55,56, 58, 75,102}.



There are total 15 values of arranged in increasing order.

Q_2 is the 8th data point.

$Q_2 = 50$



Q_1 is 4th data point ∴ $Q_1 = 46$

Q_3 is 12th data point $Q_3 = 56$

Interquartile range $IQR = Q_3 - Q_1$

$= 56 - 46$

$\boxed{= 10}$

$Q_1 - (1.5 \times IQR)$

$= 46 - (1.5 \times 10)$

$= 46 - 15$

$\boxed{= 31}$

$Q_3 + (1.5 \times IQR)$

$= 56 + 15$

$\boxed{= 71}$

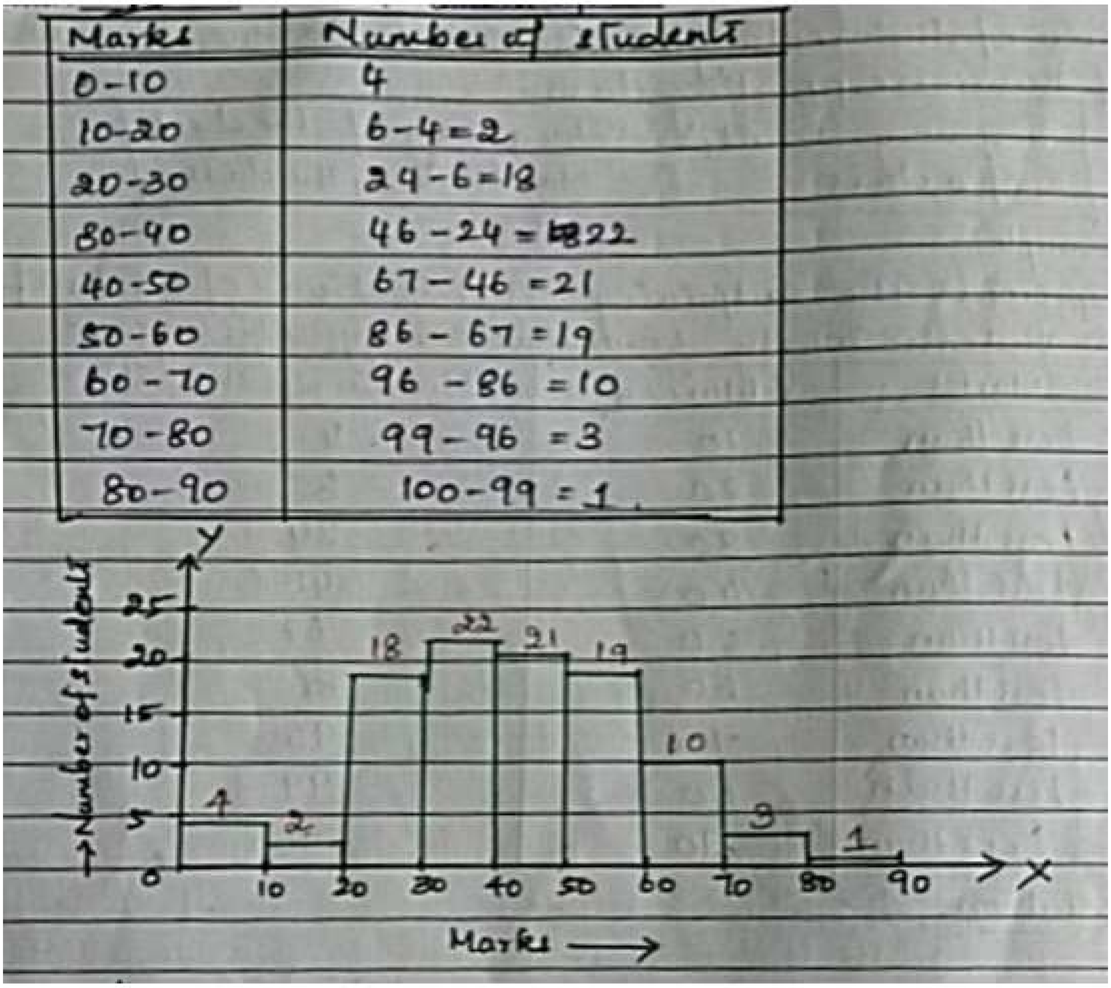
Since 5 is less than 31 and 75, 102 are greater than 71, hence there are 3 outliers.

0
10
20
30
40
50
60
70
80
90
100
110

b) b) Calculate the adjoining distribution of marks of 100 students in the examination by a histogram.

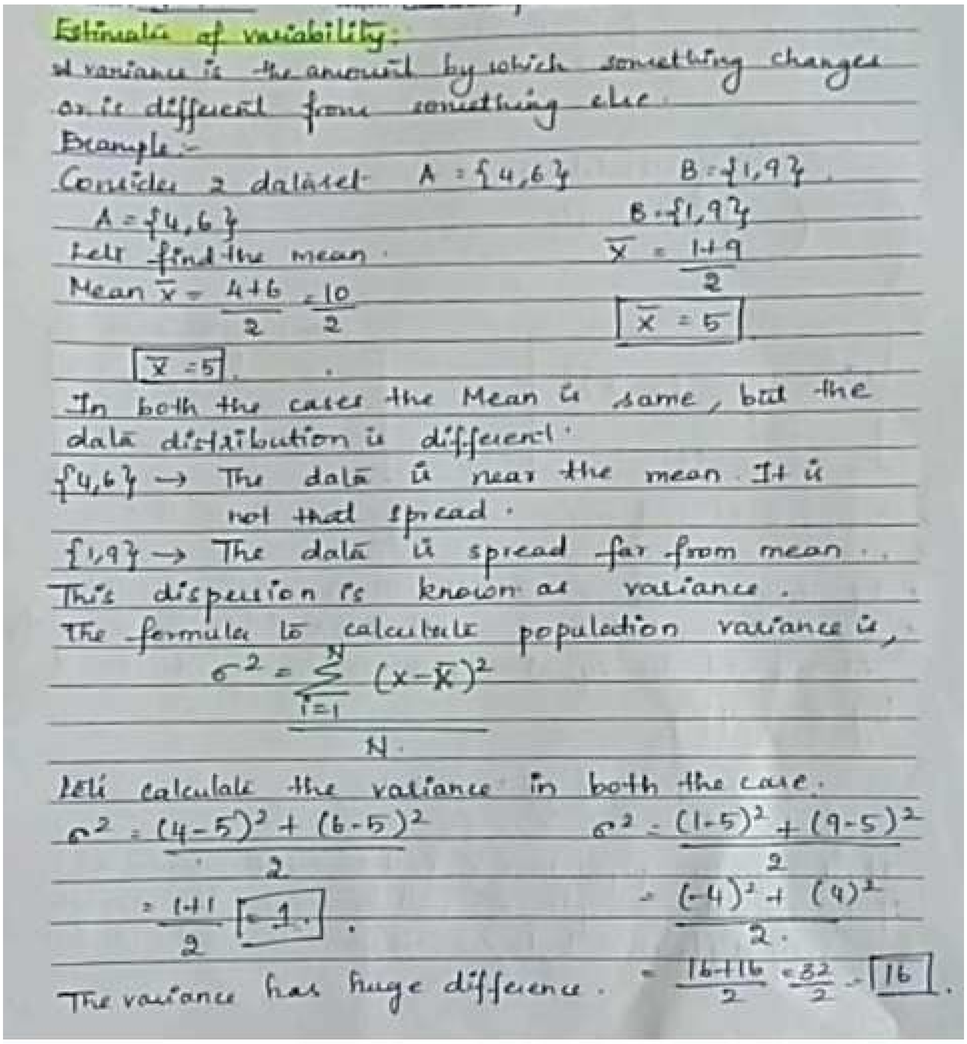
histogram.

Marks	Obtained	Number of students
Less than	10	4
Less than	20	6
Less than	30	24
Less than	40	46
Less than	50	67
Less than	60	86
Less than	70	96
Less than	80	99
Less than	90	100



Marks	Number of students
0-10	4
10-20	6-4=2
20-30	24-6=18
30-40	46-24=22
40-50	67-46=21
50-60	86-67=19
60-70	96-86=10
70-80	99-96=3
80-90	100-99=1.

c) Consider two data set A={4,6} and B={1,9}. Calculate the variance and justify the need of variance.



Estimate of variability:

A variance is the amount by which something changes or is different from something else.

Example:-

Consider 2 dataset A = {4,6} B = {1,9}.

A = {4,6}

Lets find the mean.

Mean $\bar{x} = \frac{4+6}{2} = \frac{10}{2}$

$\boxed{\bar{x} = 5}$.

B = {1,9}

$\bar{x} = \frac{1+9}{2}$

$\boxed{\bar{x} = 5}$

In both the cases the Mean is same, but the data distribution is different.

{4,6} → The data is near the mean. It is not that spread.

{1,9} → The data is spread far from mean.

This dispersion is known as variance.

The formula to calculate population variance is,

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x-\bar{x})^2}{N}.$$

Lets calculate the variance in both the case.

$\sigma^2 = \frac{(4-5)^2 + (6-5)^2}{2}$

$= \frac{1+1}{2}$ $\boxed{= 1.}$

$\sigma^2 = \frac{(1-5)^2 + (9-5)^2}{2}$

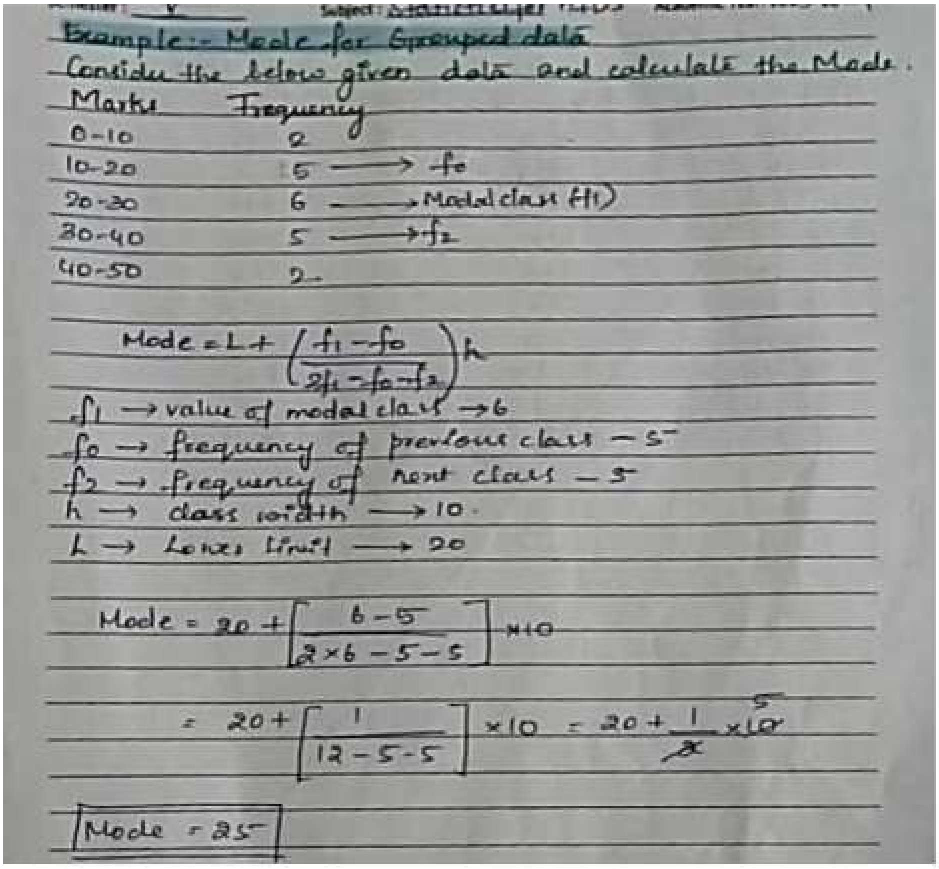
$= \frac{(-4)^2 + (4)^2}{2}.$

$= \frac{16+16}{2} = \frac{32}{2} = \boxed{16}.$

The variance has huge difference.

d) Consider the below given data and calculate the mode.

Marks	Frequency
0-10	2
10-20	5
20-30	6
30-40	5
40-50	2



Subject: [illegible]

Example:- Mode for Grouped data

Consider the below given data and calculate the Mode.

Marks	Frequency	
0-10	2	
10-20	5	$\longrightarrow f_0$
20-30	6	$\longrightarrow$ Modal class (f1)
30-40	5	$\longrightarrow f_2$
40-50	2	

$$\text{Mode} = L + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right)h$$

$f_1 \rightarrow$ value of modal class $\rightarrow 6$

$f_0 \rightarrow$ frequency of previous class $- 5$

$f_2 \rightarrow$ Frequency of next class $- 5$

$h \rightarrow$ class width $\longrightarrow 10$.

$L \rightarrow$ Lower limit $\longrightarrow 20$

$$\text{Mode} = 20 + \left[\frac{6-5}{2\times6-5-5}\right]\times10$$

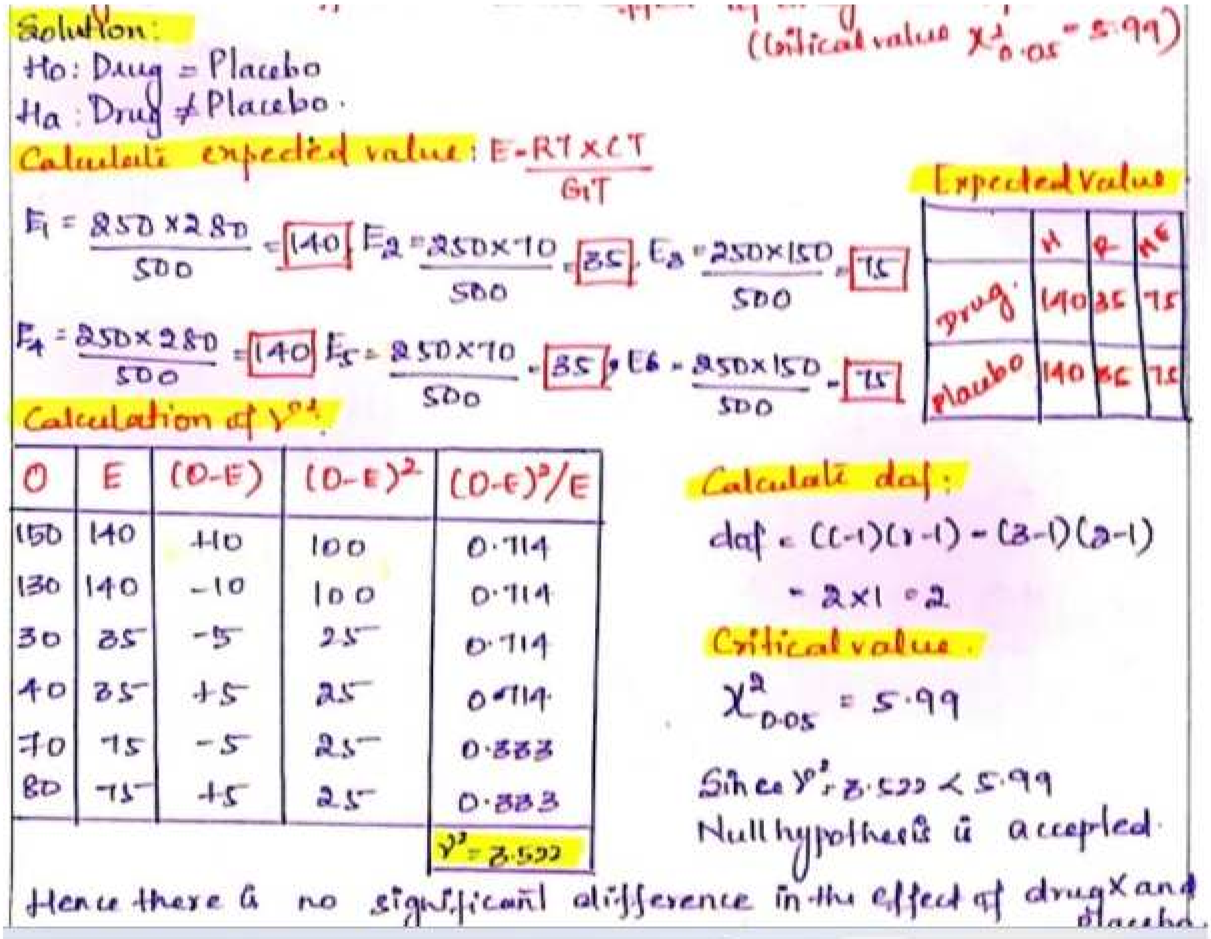
$$= 20 + \left[\frac{1}{12-5-5}\right]\times10 = 20 + \frac{1}{\cancel{2}}\times\overset{5}{\cancel{10}}$$

$$\boxed{\text{Mode} = 25}$$

3 a) A drug X claimed to be effective in curing colds. In an experiment on 500 persons with cold, half of them where given placebo (sugar pills). The patients' reactions to the treatment are recorded in the following table:

Treatment	Helped	Reaction	No Effect
Drug	150	30	70
Placebo	130	40	80

(Critical Value: 3.84)



(Critical value $\chi^2_{0.05} = 5.99$)

Solution:

Ho: Drug = Placebo

Ha : Drug ≠ Placebo.

Calculate expected value: $E = \frac{RT \times CT}{GT}$

$E_1 = \frac{250 \times 280}{500} = \boxed{140}$, $E_2 = \frac{250 \times 70}{500} = \boxed{35}$, $E_3 = \frac{250 \times 150}{500} = \boxed{75}$

$E_4 = \frac{250 \times 280}{500} = \boxed{140}$ $E_5 = \frac{250 \times 70}{500} = \boxed{35}$, $E_6 = \frac{250 \times 150}{500} = \boxed{75}$

Expected Value

	H	R	NE
Drug	140	35	75
Placebo	140	35	75

Calculation of χ^2:

O	E	(O-E)	$(O-E)^2$	$(O-E)^2/E$
150	140	+10	100	0.714
130	140	-10	100	0.714
30	35	-5	25	0.714
40	35	+5	25	0.714
70	75	-5	25	0.333
80	75	+5	25	0.333
				$\chi^2 = 3.522$

Calculate dof:

dof = $(c-1)(r-1) = (3-1)(2-1)$

$= 2 \times 1 = 2$

Critical value.

$\chi^2_{0.05} = 5.99$

Since $\chi^2 = 3.522 < 5.99$

Null hypothesis is accepted.

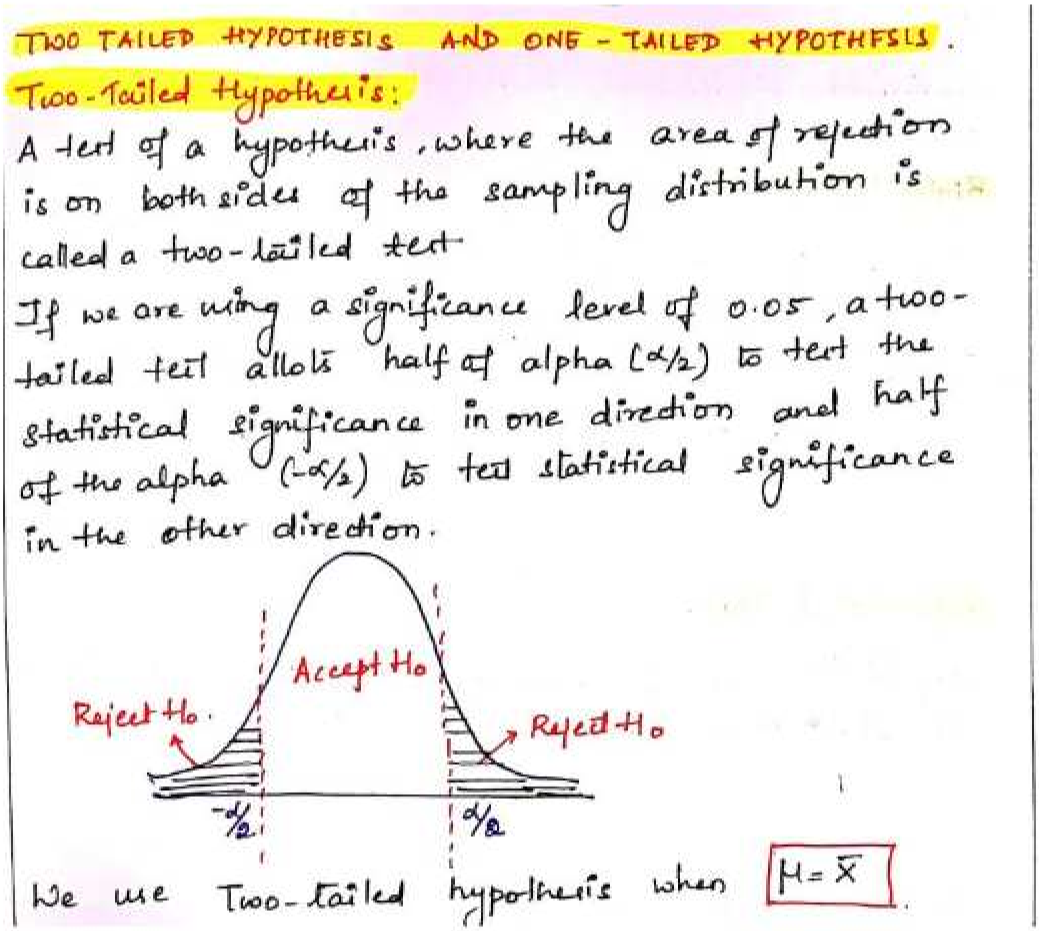
Hence there is no significant difference in the effect of drug X and placebo.

b) Illustrate in detail about Two – way Hypothesis and Test the following: Two random samples were drawn from two normal populations and their values are:

A: 16,17,25,26,32,34,38, 40,42

B: 14,16,24,28,32, 35, 37, 42,43, 45,47

Test whether the two populations have the same variance at 5% level of significance. (Critical Value = 3.35)



TWO TAILED HYPOTHESIS AND ONE - TAILED HYPOTHESIS .

Two-Tailed Hypothesis:

A test of a hypothesis, where the area of rejection is on both sides of the sampling distribution is called a two-tailed test

If we are using a significance level of 0.05, a two-tailed test allots half of alpha ($\alpha/2$) to test the statistical significance in one direction and half of the alpha ($-\alpha/2$) to test statistical significance in the other direction.

We use Two-tailed hypothesis when $\boxed{\mu = \bar{X}}$.

c) The length of life X of certain computers is approximately normally distributed with mean 800 hours and standard deviation 40 hours. If a random sample of 30 computers has an average life of 788 hours, test the null hypothesis that µ =800 hours against the alternate that µ != 800 hours at 15% level of significance.(Critical Value : 1.44)

Solution:

Null Hypothesis = 800 hours

Alternate Hypothesis != 800

It is two tailed

Calculation: z = (788 – 800)/(40/squareroot 30)

= -1.643

Reject the null hypothesis (-1.643 > -1.44)

d) Write the difference between null hypothesis and alternate hypothesis. A researcher wants to know if the height of students at school differs from the national average of 5.5 feet. State null and alternative hypothesis

Null Hypothesis (Ho):

A null hypothesis is astatement in which there is no relation between two variables.

- Researchers try to reject or disprove it.
- The testing process is always Indirect and Implicit.
- Null hypothesis is rejected if the p-value is less than the alpha-value ; otherwise it is rejected.
- It is denoted by Ho.
- The symbol used are (=, >=, <=) Equality.

ALTERNATIVE HYPOTHESIS:

- An alternative hypothesis is a statement in which there is some statistical relationship between the two variables.
- Researchers always try to accept or prove it.
- The testing process is always direct and explicit
- An alternative hypothesis is accepted if the p-value is less than the alpha - value otherwise, it is rejected.
- It is denoted by H1.
- The symbol used are Inequality (!=, <, >).

feet. State null and alternate hypo

Here, researchers are interested in determining whether the heights of students is either less than or greater than the national average height.

Null Hypothesis $H_0 = 5.5$ feet, Alternate Hypothesis (H_1) $\neq 5.5$ feet