1. The theory predicts the population of beans in the four groups A, B, C and D should be 9:3:3:1. In an experiment among 1600 beans, the numbers in the four groups were 882, 213, 287 and 118. Does the experimental result support the theory? (The table value for 3 degrees of freedom at the 5% level of significance is 7.81).

Ex. 2.7.2 : The theory predicts the proportion of beans, in the four groups A, B, C and D should be 9 : 3 : 3 : 1. In an experiment among 1,600 beans, the numbers in the four groups were 882, 313, 287 and 118. does the experimental result support the theory ? (The table value of χ^2 for 3 d.f. at 5% level of significance is 7.81).

Soln. :

- **Null Hypothesis :** H_0 : There is no significant difference between the experimental values and the theory; i.e. the theory supports the experiment.
- The proportion of beans in four groups A, B, C and D should be 9 : 3 : 3 : 1. Hence the theoretical (expected) frequencies are as shown.

Category	Expected frequency (E)
A	$\frac{9}{16} \times 1600 = 900$
B	$\frac{3}{16} \times 1600 = 300$
C	$\frac{3}{16} \times 1600 = 300$
D	$\frac{1}{16} \times 1600 = 100$

Computation of χ^2

Category	Observed frequency (O)	Expected frequency (E)	O – E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
A	882	900	– 18	324	0.360
B	313	300	+ 13	169	0.563
C	287	300	– 13	169	0.563
D	118	100	18	324	3.240
Total	1600	1600	0	986	4.726

$$\therefore \quad \chi^2 = \sum\left[\frac{(O-E)^2}{E}\right] = 4.726$$

Now, d.f. = 4 – 1 = 3 and tabulated χ^2 for 3 d.f.

$$\chi^2_{0.05} = 7.81$$

- **Conclusion :** Since calculated value of χ^2 is less than tabulated value, it is not significant. Hence we accept the null Hypothesis at 5% level of significance. Thus, the experimental results support the theory.

2. A die is rolled 100 times with the following distribution. (The table value for 5 degrees of freedom at the 1% level of significance is 15.086).

Number	1	2 3 4 5	6
Observed Frequency	17	14 20 17 17	15

Ex. 2.7.3 : A die is rolled 100 times with the following distribution.

Number	1	2	3	4	5	6
Observed frequency	17	14	20	17	17	15

At 0.01 level of significance, determine whether the die is true (or uniform).

☑ **Soln. :**

We have number of categories = 6

N = Total Frequency = 17 + 14 + 20 + 17 + 17 + 15 = 100

- **Null Hypothesis :** H_0 : The die is true (uniform)
- Under H_0, the probability of obtaining each of the six faces 1,2,....,6 is same, i.e. $P = \frac{1}{6}$

$\therefore$ Expected frequency for each face = N. P

$= 100 \cdot \frac{1}{6} = 16.67$

Computation of χ^2

Number	Observed frequency (O)	Expected frequency (E)	O – E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
1.	17	16.67	0.33	0.1089	0.0065
2.	14	16.67	– 2.67	7.1289	0.4276
3.	20	16.67	3.33	11.0889	0.6652
4.	17	16.67	0.33	0.1089	0.0065
5.	17	16.67	0.33	0.1089	0.0065
6.	15	16.67	– 1.67	2.7889	0.1673
total	-	-	0	-	1.2796

$$\therefore \quad \chi^2 = \sum\left[\frac{(O-E)^2}{E}\right] = 1.2796 \qquad \ldots(i)$$

The degrees of freedom (d.f.) = 6 – 1 = 5

The critical or tabulated value of Chi-square for $\gamma = 5$ and at 1% level of significance is :

$\chi^2_5(0.01) = 15.086$ ...(ii)

Since calculated valued of χ^2 is less than critical value, it is not significant.

Hence H_0 may be accepted at 1% level of significance; i.e. the die may be regarded as true or iform.

3. Tata Soaps manufacturing company was distributing a particular brand of soap through a large number of retail shops. Before a heavy advertisement campaign, the mean sales per week per shop was 140 dozens,. After the campaign, a sample of 26 shops was taken and the mean sales was found to be 147 dozens with a standard deviation 16. Can you consider the advertisement effective?. (The table value for 25 degrees of freedom at the 5% level of significance is 1.798).

Ex. 2.7.8 : Godrej soaps manufacturing company was distributing a particular brand of soap through a large number of retail shops. Before a heavy advertisement campaign, the mean sales per week per shop was 140 dozens. After the campaign, a sample of 26 shops was taken and the mean sales was found to be 147 dozens with standard deviation 16. Can you consider the advertisement effective?

☑ **Soln. :**

We have $n = 26, \bar{x} = 147,$

$s = 16$ dozens

Null Hypothesis : $H_0 : \mu = 140$ dozens; i.e. the deviation between $\bar{x}$ and μ is just due to fluctuations of sampling. In other words, advertisement is not effective.

Alternate Hypothesis, $H_1 : \mu > 140$ and

Under H_0, the test statistic is

$$t = \frac{147 - 140}{16/\sqrt{25}} = \frac{7 \times 5}{16} = 2.19 \qquad \left[\because t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{\bar{x} - \mu}{s/\sqrt{n-1}} \sim t_{n-1} = t_{25}\right]$$

Tabulated value of t for 25 d. f. At 5% level of significance for single tail test is 1.798, i.e. $t_{25}(0.05) = 1.708$.

Since calculated value of t is greater than the tabulated value, it is significant and we reject H_0 at 5% level of significance.

Hence, the increase in sales cannot be attributed to fluctuations of sampling and we conclude that advertisement is certainly effective in increasing the sales.

4. A machine is designed to produce insulating washers for electrical devices of average thickness 0.025cm. A random sample of 10 washers was found to have an average thickness of 0.024cm with a standard deviation of 0.002cm. Test the significance of deviation. Value 0f t for 9 degrees of freedom at 5% level of significance is 2.262).

Ex. 2.7.7 : A machine is designed to produce insulating washers for electrical devices of average thickness 0.025 cm. A random sample of 10 washers was found to have an average thickness of 0.024 cm. with a standard deviation of 0.002 cm. Test the significance of deviation. Value of t for 9 degrees of freedom at 5% level is 2.262.

☑ **Soln. :**

We have $n = 10$, $\bar{x} = 0.024$ cm.

$s = 0.002$ cm (sample standard deviation)

☞ **Null Hypothesis**

$H_0 : \mu = 0.025$ cm; i.e. there is no significant deviation between sample mean $\bar{x} = 0.024$ and population mean $\mu = 0.025$.

☞ **Alternative Hypothesis**

$H_1 : \mu \neq 0.025$ cm

Under H_0, the test statistic is

Under H_0, the test statistic is

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} = \frac{\bar{x} - \mu}{s/\sqrt{n-1}} \sim t_{10-1} = t_a$$

$$\therefore t = \frac{0.024 - 0.025}{0.002/\sqrt{9}} = \frac{-0.001 \times 3}{0.002} = -1.5$$

Now, tabulated valve for 9 d. f. is = 2.262 since $|t| < 2.262$, it is not significant, at 5% level of significance.

Hence the deviation $(\bar{x} - \mu)$ is not significant.

5. Find the least value of r in a sample of 18 pairs of observation from a bivariate normal population significant at 5% level of significance. (The table value for 16 degrees of freedom at the 5% level of significance for a two tailed test is 2.12).

We have : n = 1.8. The observed value of sample correlation coefficient r will be significant at level of significance if

$$|t| = \left|\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}\right| > t_{n-2}(0.025) = t_{16}(0.025)$$

But from the table, $t_{16}(0.025) = 2 \cdot 12$

$\therefore$ From, Equation (i), $\left|\frac{r\sqrt{18-2}}{\sqrt{1-r^2}}\right| > 2 \cdot 12$

$$\therefore \left|\frac{4r}{\sqrt{1-r^2}}\right| > 2 \cdot 12$$

Squaring and transposing,

$$16r^2 > (2 \cdot 12)^2 (1-r^2) = 4.4944 (1-r^2)$$

$$\therefore (16 + 4.4944) r^2 > 4.4944$$

$$\therefore r^2 > \frac{4.4944}{20.4944}$$

$$\therefore |r| > \sqrt{0.2193} = 0.4683$$

6. A random sample of 27 pairs of observations from a normal population gives a correlation coefficient of 0.42. Is it likely that the variables in the population are uncorrelated?. (The table value for 25 degrees of freedom at the 5% level of significance for a two tailed test is 2.06).

Ex. 2.8.1 : A random sample of 27 pairs of observations from a normal population gives a coefficient of 0.42. Is it likely that the variables in the population are uncorrelated?

☑ **Soln. : We have n = 27 and r = 0.42**

- **Null Hypothesis :** $H_0 : \rho = 0$; i.e. variables are uncorrelated in the population
- **Alternate Hypothesis :** $H_1 : \rho \neq 0$ (Two –tailed test)
- **Test Statistic :** Under H_0, the test statistic is :

$$t = r \cdot \frac{\sqrt{n-2}}{\sqrt{1-r^2}}$$

$$\therefore \quad t = 0.42 \frac{\sqrt{25}}{\sqrt{1-(0.42)^2}} = \frac{0.42 \times 5}{0.908} = 2.31$$

- The tabulated value of t for 25 d.f and 5% level of significance for a two- tailed test is 2.06.
- Since, calculated t > tabulated t, it is significant. Hence, null hypothesis ($\rho = 0$) is rejected at 5% level of significance and we conclude that variables are correlated in the population.

Ex. 2.8.2 : Find the least value of r in a sample of 10 [illegible] population [illegible]

7. A random sample of 20 daily workers of state. A was found to have an average daily earning of Rs. 44 with sample variance 900. Another sample of 20 daily workers from state B was found to earn on an average Rs. 30 per day with sample variance 400. Test whether the workers in state A are earring more than in state B.

Ex. 2.9.2 : A random sample of 20 daily workers of state A was found to have average daily earning of Rs. 44 with sample variance 900. Another sample of 20 daily workers from state B was found to earn on an average Rs. 30 per day with sample variance 400.

Test whether the workers in state A are earning more than in state B.

☑ **Soln. :**

Let the daily earnings (in Rs.) of the workers in states A and B be denoted by the variables X and Y respectively. Then we have

$n_1 = 20, \bar{x} = 44, S_x^2 = 900; \qquad n_2 = 20, \bar{y} = 30, S_y^2 = 400$

☞ **Null Hypothesis**

$H_0 : \mu_x = \mu_y$, i.e. there is no significant difference in the average daily earnings of the workers in states A and B.

Alternative Hypothesis : $H_1 : \mu_x > \mu_y$ (Right tailed)

Test Statistic : Under H_0,

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad \ldots(i)$$

where $S^2 = \left(\frac{n_1 S_x^2 + n_2 S_y^2}{n_1 + n_2 - 2}\right)$

$$= \left(\frac{20 \times 900 + 20 \times 400}{38}\right) = \frac{18000 + 8000}{38}$$

$$= 648.21 \qquad \ldots(ii)$$

$$\therefore \quad t = \frac{44 - 30}{\sqrt{648.21\left(\frac{1}{20} + \frac{1}{20}\right)}} = \frac{14}{\sqrt{64.821}}$$

$$= 1.7389 \qquad \ldots(iii)$$

Now, tabulated $t_{0.05}$ for d.f. $= n_1 + n_2 - 2 = 38$, for right tailed test is 1.645 (for d.f. > 30, since significant values of t are same as those of Z for the normal test).

Since calculated t is greater than tabulated t. It is significant at 5% level of significance.

Hence H_0 is rejected; i.e. H_1 is accepted at 5% level of significance.

Hence we conclude that the workers in state A are earning more than those in state B.

8. A school claimed that the students studying are more intelligent than the average school.On calculating the IQ scores of 50 students, with mean 11. The mean of the population IQ is 100 and the standard deviation is 15. State whether the claim of principal is right or not at a 5% level of significance.(Z-score at 5% level of significance is 1.645).

2.10.5 Solved Example

- A school claimed that the students study is more intelligent than the average school. On calculating the IQ scores of 50 students, the average turns out to be 11. The mean of the population IQ is 100 and standard deviation is 15. State whether the claim of principal is right or not at a 5% level of significance.

Soln. :

(i) We define the null hypothesis and the alternate hypothesis.

Null hypothesis : $H_0 : \mu = 100$

And our alternate hypothesis : $H_1 : \mu = 100$

(ii) The level of significance mentioned is $\propto = 0.05$

(iii) From the table, Z – score for $\propto = 0.05$ (i.e. for night - tailed test) is 1.645.

(iv) Using Z – text for the problem :

$$Z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$$

here, $\bar{x} = 110$

$\mu = 100$

$\sigma = 15$

$\propto = 0.05$

$n = 50$

$$\therefore z = \frac{110 - 100}{\frac{15}{\sqrt{50}}} = \frac{10}{2.12} = 4.71$$

(v) Since $Z = 4.71 > 1.645$, so we reject the null hypothesis.

9. Calculate the coefficient of correlation for the following data.

x	9	8	7 6 5	4 3 2	1
y	15	16	14 13 11	12 10 8	9

UEx. 2.12.1 (MU-20, 4 Marks)

Calculate the coefficient of correlation for the following data.

x	9	8	7	6	5	4	3	2	1
y	15	16	14	13	11	12	10	8	9

☑ **Soln. : Here, n = 9**

x	y	x^2	y^2	xy
9	15	81	225	135
8	16	64	256	128
7	14	49	196	98
6	13	36	169	78
5	11	25	121	55
4	12	16	144	48
3	10	9	100	30
2	8	4	64	16
1	9	1	81	9
$\sum x = 45$	$\sum y = 108$	$\sum x^2 = 285$	$\sum y^2 = 1356$	$\sum xy = 597$

The coefficient of correlation is

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}} = \frac{9(597) - (45)(108)}{\sqrt{9(285) - 45^2}\sqrt{9(1356) - 108^2}}$$

$$r = 0.95$$

10. From the following data which shows the ages X and systolic B.P. Y of 12 womens. Are the two variable ages X and B.P. Y correlated?

Age (X)	56	42	72	36 63 47 55 49 38	42	68	60
B.P. (Y)	147	125	160	118 149 128 150 145 115	140	152	155

Ex. 2.12.4 : From the following data which shows the ages X and systolic B.P. Y of 12 women. Are the two variable ages X and B.P. Y correlated ?

Age (X)	56	42	72	36	63	47	55	49	38	42	68	60
B.P. (Y)	147	125	160	118	149	128	150	145	115	140	152	155

☑ **Soln. :**

We determine correlation coefficient r to find the association between age and B.P.

Now,
$$r = \frac{N\sum XY - \sum X \cdot \sum Y}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum (Y^2) - (\sum Y)^2]}}$$

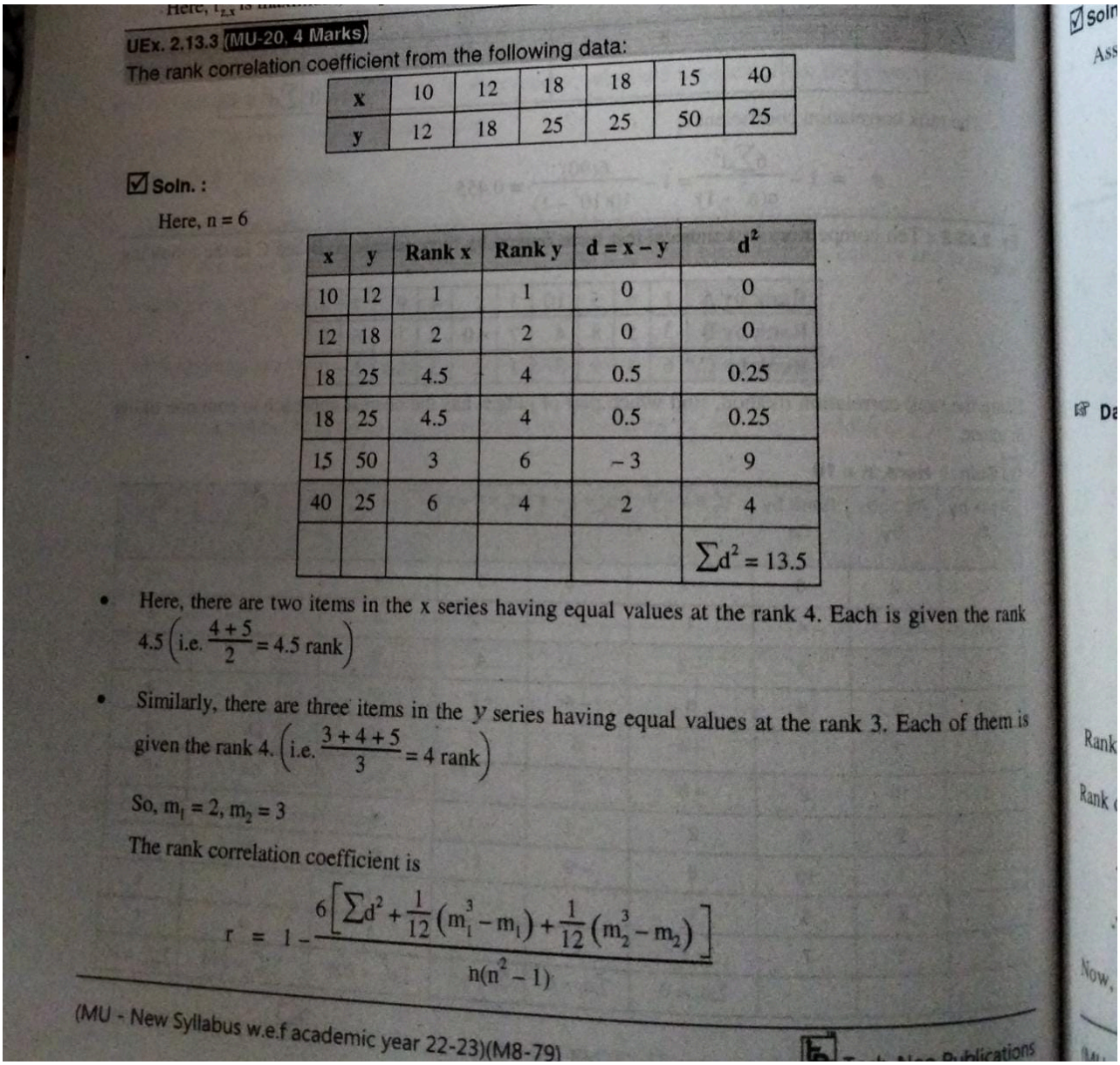
$$= \frac{12\,(89894) - (628)\,(1684)}{\sqrt{[(12)\,(34416) - (628)^2]\,[(12)\,(238822) - (1684)^2]}}$$

$$= 0.8961$$

∴ Age X and B.P. Y are strongly positively correlated.

11. Find the rank correlation coefficient from the following data.

x	10	12 18 18 15	40
y	12	18 25 25 50	25



Here, r_{xx} is ...

UEx. 2.13.3 (MU-20, 4 Marks)

The rank correlation coefficient from the following data:

x	10	12	18	18	15	40
y	12	18	25	25	50	25

Soln. :

Here, n = 6

x	y	Rank x	Rank y	d = x – y	d^2
10	12	1	1	0	0
12	18	2	2	0	0
18	25	4.5	4	0.5	0.25
18	25	4.5	4	0.5	0.25
15	50	3	6	– 3	9
40	25	6	4	2	4
					$\sum d^2 = 13.5$

- Here, there are two items in the x series having equal values at the rank 4. Each is given the rank 4.5 $\left(\text{i.e. } \frac{4+5}{2} = 4.5 \text{ rank}\right)$
- Similarly, there are three items in the y series having equal values at the rank 3. Each of them is given the rank 4. $\left(\text{i.e. } \frac{3+4+5}{3} = 4 \text{ rank}\right)$

So, $m_1 = 2$, $m_2 = 3$

The rank correlation coefficient is

$$r = 1 - \frac{6\left[\sum d^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2)\right]}{n(n^2 - 1)}$$

12. Determine the rank correlation for the following data which shows the marks obtained in two quizzes in mathematics.

Marks in the 1st quiz (X)	6	5	8 8 7 6 10 4	9	7
Marks in the 2nd quiz (Y)	8	7	7 10 5 8 10 6	8	6

Example On Rank

UEx. 2.13.4 **(MU-19, 4 Marks)**

Determine rank correlation for the following data which shows the marks obtained in two quizes in mathematics:

Marks in 1st quiz (X) :	6	5	8	8	7	6	10	4	9	7
Marks in 2nd quiz (Y) :	8	7	7	10	5	8	10	6	8	6

Soln. :

Assigning ranks to the data of X, we get

X :	4	5	6	6	7	7	8	8	9	10
Rank :	1	2	3	4	5	6	7	8	9	10
or :	1	2	3.5	3.5	5.5	5.5	7.5	7.5	9	10
Similarly Y :	5	6	6	7	7	8	8	8	10	10
Rank :	1	2	3	4	5	6	7	8	9	10
or :	1	2.5	2.5	4.5	4.5	7	7	7	9.5	9.5

Data assigned with ranks is

X :	3.5	2	7.5	7.5	5.5	3.5	10	1	9	5.5
Y :	7	4.5	4.5	9.5	1	7	9.5	2.5	7	2.5
D :	– 3.5	–2.5	3	–2	4.5	–3.5	0.5	–1.5	2	3
D^2 :	12.25	6.25	9	4	20.25	12.25	0.25	2.25	4	9

$$\sum D^2 = 79.5$$

$$\text{Rank correlation} = 1 - \left[\frac{6\sum D^2}{n(n^2-1)}\right] = 1 - \left[\frac{6(79.5)}{10(99)}\right] = 1 - 0.4818$$

Rank correlation = 0.5182

$$\therefore \frac{e^{-m} \cdot m^1}{1!} = 2\frac{e^{-m} \cdot m^2}{2!} \quad \therefore m = 1$$

$\therefore$ Mean = 1 and Variance = 1.

Now, $P(x = 3) = \frac{e^{-1} \cdot (1)^3}{3!} = \frac{1}{6e} = 0.0613$ ...Ans.

13. Find the Bowley Skewness for the following set of data:

No of pets	B No. of family Cumulative frequency
0	60 60

1	60 120
2	50 170
3	20 190
4	25 215
5	10 225
6 or more	5 230

☞ **Worked Example**

(1) Find the Bowley skewness for the following set of data:

#of pets	#of families	Cumulative freq.
0	60	60
1	60	120
2	50	170
/		
3	20	190
4	25	215
	10	225
5	5	230
6 or more		

☑ **Soln. :**

▶ **Step (i) :** To find the Quartiles for the data set.

$$Q_1 = \frac{(\text{total cum. freq.} + 1)}{4} = \frac{230 + 1}{4} = 57.75$$

$$Q_2 = \frac{(\text{total cum. freq.} + 1)^{th}}{2} \text{ observation}$$

$$= \frac{230 + 1}{2} = 115.5$$

$$Q_3 = 3\left(\frac{\text{total cum. freq.} + 1}{4}\right)^{th} \text{ observation}$$

$$= 3\left(\frac{230 + 1}{4}\right) = 173.25$$

- **Step (ii) :** From the table, we find the n^{th} observations, we calculated in step (i) :

$$Q_1 = 57.75^{th} \text{ observation} = 0$$

$$Q_2 = 115.5^{th} \text{ observation} = 1$$

$$Q_3 = 173.25^{th} \text{ observation} = 3$$

- **Step (iii) :** Using the formula,

$$\text{B. Coeff.} = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1} = \frac{(3 + 0 - 2)}{3 - 0} = \frac{1}{3}$$

$$\because \quad \text{Coeff.} = \frac{1}{3} > 0$$

$\therefore$ distribution is positively skewed.

14. Find out the fallacy. If any in the following statement: if X is a possession variate such that P(X=2) = 9P(X=4) + 90 P(X= 6), then mean of X=1.

UEx. 2.16.2 (MU : 2017)

Find out the fallacy if any in the following statement : If X is a Poisson Variate such that $P(X=2) = 9\,P(X=4) + 90\,P(X=6)$, then mean of X = 1.

Soln. : Let m be the mean of X,

$$\text{then } P(X=x) = \frac{e^{-m} \cdot m^x}{x!}$$

$$\because P(X=2) = 9\,P(X=4) + 90\,P(X=6)$$

$$\therefore e^{-m} \cdot \frac{m^2}{2!} = 9 \cdot \frac{e^{-m} \cdot m^4}{4!} + 90\, e^{-m} \cdot \frac{m^6}{6!}$$

$$\therefore \frac{1}{2} = \frac{3m^2}{8} + \frac{m^4}{8}$$

$$\therefore m^4 + 3m^2 - 4 = 0$$

$$\therefore (m^2+4)(m^2-1) = 0$$

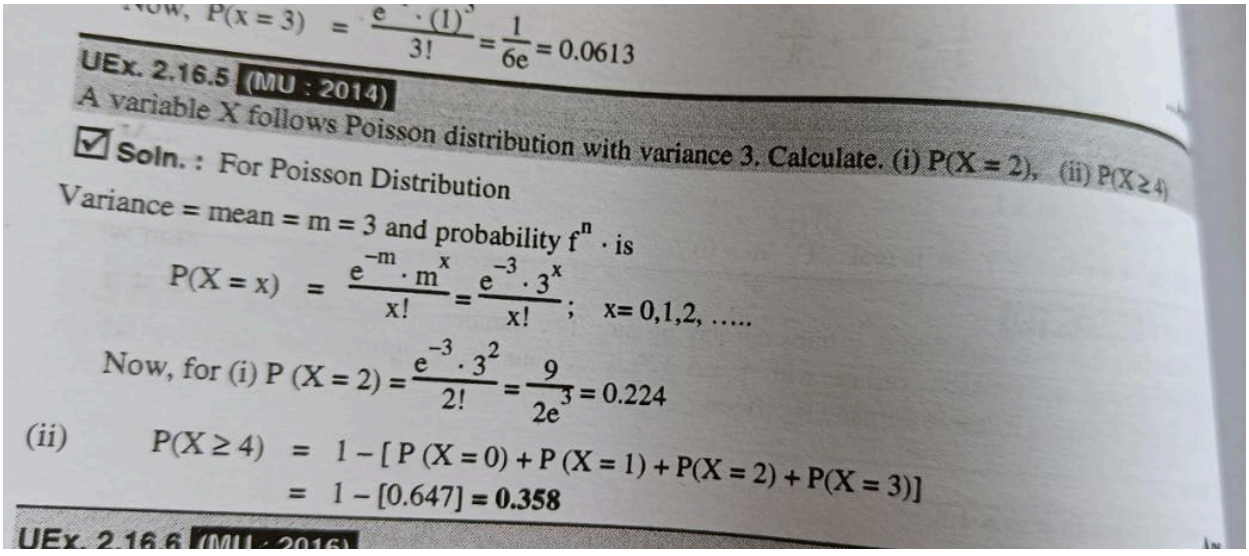
$$\therefore m^2 = 1 \quad \therefore m = 1$$

...Ans.

and $m^2 + 4 \neq 0$, $\because$ m is real $(\because m > 0)$

$\therefore$ Statement is correct

15. A variable X follows Poison distribution with variance 3. Calculate (i) P(X=2), (ii) P(X>=4).



Now, $P(x=3) = \frac{e^{-3} \cdot (1)^3}{3!} = \frac{1}{6e} = 0.0613$

UEx. 2.16.5 (MU : 2014)

A variable X follows Poisson distribution with variance 3. Calculate. (i) $P(X=2)$, (ii) $P(X \geq 4)$

Soln. : For Poisson Distribution

Variance = mean = m = 3 and probability f^n · is

$$P(X=x) = \frac{e^{-m} \cdot m^x}{x!} = \frac{e^{-3} \cdot 3^x}{x!}; \quad x = 0,1,2, \ldots..$$

Now, for (i) $P(X=2) = \frac{e^{-3} \cdot 3^2}{2!} = \frac{9}{2e^3} = 0.224$

(ii) $P(X \geq 4) = 1 - [P(X=0) + P(X=1) + P(X=2) + P(X=3)]$

$= 1 - [0.647] = \mathbf{0.358}$

UEx. 2.16.6 (MU : 2016)

16. For a certain normal distribution, the first moment of about 10 is 40 nd the fourth moment of about 50 is 48. What is the arithmetic mean and standard deviation of the distribution?

Ex. 2.19.1 : For a certain normal distribution, the first moment about 10 is 40 and the fourth moment about 50 is 48. What is the arithmetic mean and standard deviation of the distribution ?

☑ **Soln. :**

We know that if μ_1' is the first moment about the point X = a, then arithmetic mean is given by

mean = $a + \mu_1'$

Now, μ_1' (about the pointt X = 10) = 40

$\therefore$ Mean = $10 + 40 = 50$

Also, μ_4' (about the point X = 50) = 48

$$\therefore \mu_4 = 48 \left(\because \text{Mean} = 50 , \therefore \mu_4' = \mu_4 \right)$$

Now, for a normal distribution with standard deviation σ,

$\mu_4 = 3\sigma^4 \quad \therefore 3\sigma^4 = 48 \quad \therefore \sigma^4 = 16 \quad \therefore \sigma = 2$ ...Ans.

Ex. 2.19.2 : To illustrate the use of table, we evaluate the area A under the normal curve.

17. If X is a normal variate with a mean of 30 and a standard deviation is 5. Find the probabilities that (i) 26<=X<= 40 (II) X<=45

Ex. 2.19.4 : If X is a normal variate with mean 30 and S. D. 5. Find the probabilities that (i) $26 \le X \le 40$ (ii) $X \ge 45$ and (iii) $|X - 30| > 5$

☑ **Soln. :**

(i) We have mean $m = 30$, $\sigma = 5$

$$\text{Let} \quad Z = \frac{x - m}{\sigma} = \frac{x - 30}{5}$$

When $X = 26$, $Z = \frac{26 - 30}{5} = -0.8$

When $X = 40$ $Z = \frac{40 - 30}{5} = 2$

$$\therefore \quad P(26 \le X \le 40) = P(-0.8 \le Z \le 2) = P(-0.8 \le Z \le 0) + P(0 \le Z \le 2)$$
$$= P(0 \le Z \le 0.8) + P(0 \le Z \le 2) \quad \text{(by symmetry)}$$
$$= 0.2881 + 0.4772 = 0.7653 \quad \text{(from table)}$$

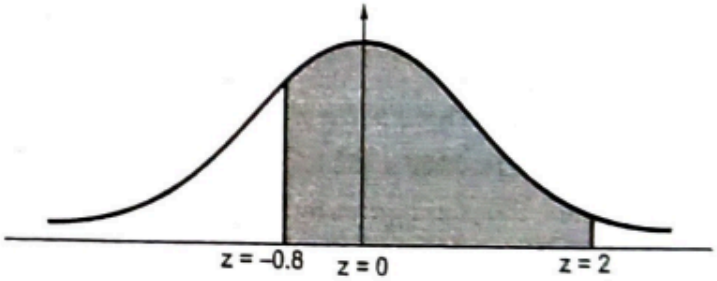


Fig. Ex. 2.19.4

(ii) When $x = 45$, $Z = \frac{45 - 30}{5} = 3$

$$\therefore \quad P(X \ge 45) = P(Z \ge 3) = 0.5 - P(0 \le Z \le 3) = 0.5 - 0.49865 = 0.00135$$

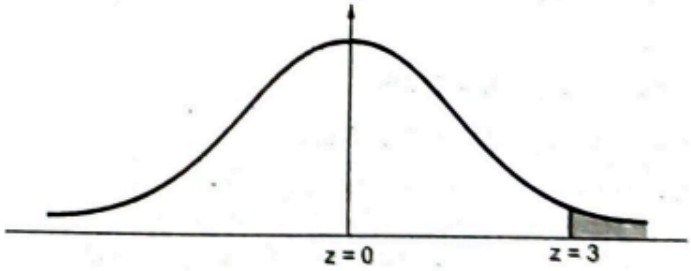


Fig. Ex. 2.19.4(a)

18. A trucking company wishes to test the average life of the four brands of tyres. The company uses all the brands on randomly selected trucks. The records showing the lives (Thousands of miles) of tyres are as given in the table.
Test the hypothesis that the average life for each band of tyres is the same. Assume standard deviation= 0.01. Apply suitable test

Brand 1	Brand 2 Brand 3 Brand 4
20	19 21 15
23	15 19 17
18	17 20 16
17	20 17 18

Ex. 2.22.1 : A trucking company wishes to test the average life of each of the four brands of tyres. The company uses all the brands on randomly selected trucks. The records showing the lives (thousands of miles) of tyres are as given in the table.

Test the hypothesis that the average life for each brand of tyres is the same. Assume $\alpha = 0.01$.

Table P.2.22.1

Brand 1	Brand 2	Brand 3	Brand 4
20	19	21	15
23	15	19	17
18	17	20	16
17	20	17	18
	16	16	

☑ **Soln. :**

Here, the factors of variation are brands of tyres.

Set up the hypothesis

- **Null Hypothesis :** $\mu_1 = \mu_2 = \mu_3 = \mu_4$,

  i.e. the mean life of the tyres of all the brands is same.
- **Alternative Hypothesis :** At least two means are different.

	Brand 1	Brand 2	Brand 3	Brand 4	
	20	19	21	15	
	23	15	19	17	
	18	17	20	16	
	17	20	17	18	
Total $T_i = \sum_j X_{ij}$	$T_1 = 78$	$T_2 = 87$	$T_3 = 93$	$T_4 = 66$	$G = \Sigma\Sigma X_{ij} = 324$
T_i^2	$T_1^2 = 6084$ $n_1 = 4$	$T_2^2 = 7569$ $n_2 = 5$	$T_3^2 = 8649$ $n_3 = 5$	$T_4^2 = 4356$ $n_4 = 4$	

$\Sigma n_i = 18$

G = Grand total $= 78 + 87 + 93 + 66 = 324$

$n = n_1 + n_2 + n_3 + n_4 = 4 + 5 + 5 + 4 = 18$

$\therefore$ Correction factor (C.F.) $= \frac{G^2}{n} = \frac{(324)^2}{18} = \frac{104976}{18} = 5832$

Raw sum of squares (RSS) $= \Sigma X_{ij}^2$

$= (400 + 529 + 324 + 289) + (361 + 225 + 289 + 400 + 256)$
$+ (441 + 361 + 400 + 289 + 256) + (225 + 289 + 256 + 324)$
$= 1542 + 1531 + 1747 + 1094 = 5914$

$\therefore$ Total sum of squares = (TSS)

$= RSS - CF = 5914 - 5832 = 82$

Between Samples (brands of tyres)

Sum of squares (BSS) is given by

$$BSS = \frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + \frac{T_4^2}{n_4} - C.F.$$

$$= \frac{6084}{4} + \frac{7569}{5} + \frac{8649}{5} + \frac{4356}{4} - 5832$$

$= (1521 + 1513.8 + 1729.8 + 1089) - 5832 = 21.6$

Within samples (Error) Sum of squares (WSS)

$= TSS - BSS = 82 - 21.6 = 60.4$

ANOVA Table

Sources of Variation	d.f.	Sum of Squares	Mean S.S.	Variance ratio
Between brands of tyres	$4 - 1 = 3$	21.6	$\frac{21.6}{3} = 7.2$	$\frac{7.2}{4.31} = 1.67$
Error	$17 - 3 = 14$	60.4	$\frac{60.4}{14} = 4.31$	–
Total	$18 - 1 = 17$	82	–	–

☞ Critical Value

The critical (tabulated) value of F ($r_1 = 3, r_2 = 14$) d.f. at $\alpha = 0.01$ is 5.56 (from table)

Since the calcualted value of the test statistic F = 1.67 is less than the critical value, it is not significant, i.e. it does not fall in the rejection region. Hence, the null hypothesis H_0 is to be rejected.

☞ Conclusion

There is no significant differnece between the average lives of the four brands of tyres 1, 2, 3 and 4.

19.

Example 5.2: The following is the information about the settlement of an industrial dispute in a factory. Comment on the gains and losses from the point of view of workers and that of management:

	Before	*After*
No. of Workers	3000	2900
Mean wages (₹)	2200	2300
Median wages (₹)	2500	2400
Standard deviation	300	260

Solution: The comments on gains and losses from both worker's and management's point of view are as follows:

Total Wages Bill

Before	*After*
$3000 \times 2200 = 66,00,000$	$2900 \times 2300 = 66,70,000$

The total wage bill has increased after the settlement of dispute, workers retained after the settlement are 50 workers less than the previous number.

After the settlement of dispute, the workers as a group are better off in terms of monetary gain. If the workers' efficiency remain same, then it is against the interest of management. But if the workers feel motivated, resulting in increased efficiency, then management can achieve higher productivity. This would be an indirect gain to management also.

Since workers retained after the settlement of dispute are less than the number employed before, it is against the interest of the workers.

Median Wages: The median wage after the settlement of dispute has come down from ₹ 2500 to ₹ 2400. This indicates that before the settlement, 50 per cent of the workers were getting wages above ₹ 2500 but after the settlement, they will be getting only ₹ 2400. It has certainly gone against the interest of the workers.

Uniformity in the Wage Structure: The extent of relative uniformity in the wage structure before and after the settlement can be determined by comparing the coefficient of variation as follows:

	Before	*After*
Coefficient of Variation (CV)	$\frac{300}{2200} \times 100 = 13.63$	$\frac{260}{2300} \times 100 = 11.30$

Since CV has decreased after the settlement from 13.63 to 11.30, the distribution of wages is more uniform after the settlement, that is, there is now comparatively less disparity in the wages received by the workers. Such a position is good for both the workers and the management in maintaining a cordial work environment.

Pattern of the Wage Structure: The nature and pattern of the wage structure before and after the settlement can be determined by comparing the coefficients of skewness.

	Before	*After*
Coefficient of skewness, Sk_p	$\frac{3(2200 - 2500)}{300} = -3$	$\frac{3(2300 - 2400)}{260} = -1.15$

Since coefficient of skewness is negative and has increased after the settlement, therefore it suggests that number of workers getting low wages has increased and that of workers getting high wages has decreased after the settlement.

Q20. Elucidate in what way measures of central tendency, variation, skewness and kurtosis are complementary to one another in understanding a frequency distribution table.

Ans. Measures of central tendency, variation, skewness, and kurtosis are complementary to each other in understanding a frequency distribution table because they provide different perspectives on the shape and characteristics of the distribution.

Central tendency measures, such as the mean, median, and mode, provide information about the typical or average value of the data. They give a sense of where the majority of the data lies and can be used to compare different datasets.

Variation measures, such as the range, variance, and standard deviation, provide information about the spread or dispersion of the data. They describe how the data is distributed around the central tendency and can be used to compare the variability of different datasets.

Skewness measures, such as the skewness coefficient, provide information about the symmetry of the distribution. A positively skewed distribution has a longer tail to the right, while a negatively skewed distribution has a longer tail to the left. Skewness can affect the interpretation of the mean as a measure of central tendency, and it can be important in certain applications such as finance and insurance.

Kurtosis measures, such as the kurtosis coefficient, provide information about the peakedness of the distribution. A distribution with high kurtosis has a sharp peak and heavy tails, while a distribution with low kurtosis has a flat peak and light tails. Kurtosis can affect the interpretation of the variance and standard deviation as measures of variation, and it can be important in certain applications such as risk management.

Taken together, these measures provide a comprehensive picture of the distribution of data and can be used to identify patterns, outliers, and other characteristics that may be relevant to understanding the data. For example, a distribution with a high mean and high standard deviation may indicate that there are both high and low values in the data, while a distribution with a negative skewness coefficient and low kurtosis may indicate that the data is concentrated in the lower range.

By using a combination of these measures, analysts can gain a deeper understanding of the underlying patterns and characteristics of the data and make more informed decisions based on that understanding.

Q.2. Find the standard deviation and kurtosis of the following set of data pertaining to kilowatt hours (kwh) of electricity consumed by 100 persons in a city.

Consumption (in kwh)	0-10	10-20	20-30	30-40	40-50
Number of users	10	20	40	20	10

Solution

Calculation for standard deviation

Consumption (in kwh)	Number of users (f)	Mid-value (m)	d= (m-A)/10 = (m-25)/10	-fd	fd^2
0-10	10	5	-2	-20	40
10-20	20	15	-1	-20	20
20-30	40	25	0	0	0
30-40	20	35	1	20	20
40-50	10	45	2	20	40
	100			0	120

Mean, $\bar{x} = A + \frac{\sum fd}{N} \times h = 25 + \frac{0}{100} \times 10 = 25$

Since $\bar{x} = 25$ is an integer value, therefore we may calculate moments about the actual mean

$\mu_r = \frac{1}{N}\sum f(-\bar{X})^r = \frac{1}{N}\sum f(m-\bar{x})^r$

Let d $= \frac{m-\bar{x}}{h}$ or (m- $\bar{x}$)=hd. Therefore

$\mu_r = h^r\frac{1}{N}\sum fd^r$; h= width of class intervals

Mid-value	Frequency	$d=\frac{m-25}{10}$	fd	fd^2	fd^3	fd^4
5	10	-2	-20	40	-80	160
15	20	-1	-20	20	-20	20
25	40	0	0	0	0	0
35	20	1	20	20	20	20
45	10	2	20	40	80	160
	100			120	0	360

Moments about the origin A=25

$$\mu_1 = h\frac{1}{N}\sum fd) = 10 \times \frac{1}{100} = 0$$

$$\mu_2 = h^2\frac{1}{N}\sum fd^2 = (10)^2\frac{1}{100} \times 120 = 120$$

$$\mu_3 = h^3\frac{1}{N}\sum fd^3 = (10)^3 \times \frac{1}{100} \times 0 = 0$$

$$\mu_4 = h^4\frac{1}{N}\sum fd^4 = (10)^4 \times \frac{1}{100} \times 360 = 36{,}000.$$

$$\text{S.D.} = \sqrt{\mu_2} = \sqrt{120} = 10.95$$

Karl Pearson's measure of kurtosis is given by

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{36{,}000}{(120)^2} = 2.5$$

and therefore $\gamma_2 = \beta_2 - 3 = 2.5-3 = -0.50$

Since $\beta_2 < 3$(or $\gamma_2 < 0$), distribution curve is platykurtic.

21. List down the cases when one should apply bowleys coefficient of skewness over karl Pearsons PNSCS

Below are the cases when one should apply Bowley's coefficient of skewness over Karl Pearson's:

Presence of extreme outliers: Bowley's coefficient is less sensitive to extreme outliers compared to Pearson's coefficient. Hence, when there are extreme outliers present in the data, Bowley's coefficient is more appropriate.

Non-normal distribution: Pearson's coefficient assumes a normal distribution of data, whereas Bowley's coefficient can be used for non-normal distributions.

Small sample size: Pearson's coefficient may not be appropriate for small sample sizes, whereas Bowley's coefficient can be used for small sample sizes.

Categorical or ordinal data: Pearson's coefficient requires interval or ratio data, whereas Bowley's coefficient can be used for ordinal or categorical data.

Skewed data: When the data is skewed, Bowley's coefficient is more robust compared to Pearson's coefficient.

23) Explain data science process in detail PCPEMED

Data science process refers to the systematic approach used to extract insights and knowledge from data using various techniques and tools. The process typically involves the following steps:

Problem identification
Data collection
Data preparation
Exploratory data analysis
Modeling
Evaluation
Deployment

24) Terms AI, Machine learning , deep learning and data science are used interchangeably. Mention your views about the relations among them

AI - simulation of human intelligence in machines
Machine learning - learn patterns and make decisions
Deep learning - uses artificial neural networks to model and solve complex problems
Data science - use of statistical and computational methods to extract insights and knowledge

Therefore, we can say that data science is a broader field that encompasses machine learning, deep learning, and AI. While these fields are related, they have distinct differences in terms of their techniques, methodologies, and applications.

26) Discuss types of data with suitable example

Nominal data: categorical data that cannot be ordered or ranked. It consists of labels or names that do not have any inherent numerical value. Examples include gender (male, female), marital status (married, single, divorced), or favorite color (red, blue, green).

Ordinal data: categorical data, but it can be ranked or ordered based on some criterion. Examples include education level (elementary school, high school, college, postgraduate), socioeconomic status (low, middle, high), or customer satisfaction ratings (poor, fair, good, excellent).

Interval data: numerical data that has consistent intervals between each value, but there is no true zero point. Examples include temperature measured in Celsius or Fahrenheit, IQ scores, or dates.

Ratio data: numerical data that has consistent intervals between each value and also has a true zero point. Examples include height, weight, time, or income.

27) Discuss in brief different types of probability distributions NPBEGBU

Normal Distribution(Gaussian) - continuous data that is symmetric and bell-shaped

Poisson Distribution: count data, such as the number of events that occur in a certain time period

Binomial Distribution: binary data, which can take on only two values, such as success or failure

Exponential Distribution: data that represents the time between events

Gamma Distribution: continuous data that is positively skewed

Beta Distribution: continuous data that is bounded between 0 and 1

Uniform Distribution: data that is equally likely to take on any value within a specified range

28) under which circumstances u will apply i) z test ii) T test iii) Anova test

Z-Test:
The z-test is used when the population standard deviation is known and the sample size is large

(typically n > 30). The z-test is appropriate for testing hypotheses about the mean of a normally distributed population. For example, if we want to determine if the mean weight of a population of fish is significantly different from a known value, we can use a z-test.

T-Test:
The t-test is used when the population standard deviation is unknown and the sample size is small (typically n < 30). The t-test is also appropriate for testing hypotheses about the mean of a normally distributed population. For example, if we want to determine if the mean test scores of two groups of students are significantly different, we can use a t-test.
There are two types of t-tests:
One-sample t-test: Used to compare the mean of a single sample to a known population mean.
Two-sample t-test: Used to compare the means of two independent samples.

ANOVA Test:
The ANOVA (Analysis of Variance) test is used to compare the means of three or more groups. It tests the null hypothesis that all group means are equal. If the null hypothesis is rejected, it means that at least one group mean is different from the others. For example, if we want to determine if the mean salary of employees is different across different job levels (junior, senior, manager), we can use an ANOVA test.
There are two types of ANOVA tests:
One-way ANOVA: Used when there is only one independent variable.
Two-way ANOVA: Used when there are two independent variables.

29) What is the central limit theorem?

The central limit theorem is a fundamental concept in probability theory and statistics. It states that if a large number of independent and identically distributed (i.i.d.) random variables are summed or averaged, their sum or average will tend towards a normal distribution, regardless of the underlying distribution of the original random variables.

In other words, the central limit theorem allows us to use the normal distribution to approximate the distribution of many different types of random variables, as long as the sample size is large enough. This is particularly useful in statistical inference, where we often want to make inferences about a population based on a sample of data.

For example, let's say we want to estimate the average height of all people in a city. We can take a random sample of people from the city, calculate their average height, and use the central limit theorem to approximate the distribution of the sample mean as a normal distribution. This allows us to calculate confidence intervals and perform hypothesis tests to make inferences about the population mean.

30) what is the P value? What is the level of significance?

In statistics, the p-value is the probability of obtaining a result equal to or more extreme than the observed result, assuming the null hypothesis is true. It measures the strength of evidence against the null hypothesis.

The level of significance, also known as alpha level, is the probability of rejecting the null hypothesis when it is actually true. It is usually set before conducting a statistical test and is denoted by the Greek letter alpha (α). The most common level of significance is 0.05, which means that there is a 5% chance of rejecting the null hypothesis when it is actually true.

In simpler terms, the p-value tells you whether your results are statistically significant, while the level of significance tells you how much evidence you require to reject the null hypothesis. If the p-value is less than or equal to the level of significance, then you can reject the null hypothesis and conclude that there is a significant difference between the groups being compared

31) What is the confidence interval? Explain with an example.

A confidence interval is a statistical measure that represents a range of values that is likely to contain an unknown population parameter, such as the mean or the proportion. It is calculated from a sample of data and provides a level of confidence about the true value of the parameter.

For example, suppose we want to estimate the average height of students in a particular school. We take a random sample of 100 students and measure their heights. The sample mean height is 170 cm, and the standard deviation is 5 cm. We can use this information to calculate a 95% confidence interval for the population mean height.

Next, we use a formula to calculate the confidence interval. For a 95% confidence level and a sample size of 100, the formula is:

Confidence interval = sample mean ± (1.96 x standard error)

where the standard error is the standard deviation of the sample divided by the square root of the sample size.

Plugging in the numbers from our example, we get:

Confidence interval = 170 ± (1.96 x 5/√100) = 170 ± 0.98

So our 95% confidence interval for the true population mean height is 169.02 cm to 170.98 cm. This means we are 95% confident that the true population mean height falls within this range, based on our sample data

32) What is hypothesis testing? Why and when it is conducted?

Hypothesis testing is a statistical method used to determine if there is enough evidence to support a certain hypothesis or claim about a population. It involves testing a sample of data to draw conclusions about the entire population.

Hypothesis testing is conducted in situations where there is a claim or hypothesis about a population parameter (such as mean, proportion, or standard deviation) that needs to be tested. The hypothesis can either be a null hypothesis (denoted as H0) or an alternative hypothesis (denoted as Ha or H1).

The null hypothesis is the hypothesis that there is no significant difference between the sample and population, while the alternative hypothesis is the opposite of the null hypothesis. The hypothesis testing process involves setting a level of significance (alpha) and using statistical tests to determine the probability of obtaining the sample results if the null hypothesis is true. If the probability is less than the chosen level of significance, we reject the null hypothesis and accept the alternative hypothesis.

Hypothesis testing is conducted to make statistical inferences about the population parameters based on a sample. It is used in various fields, including medical research, social sciences, and business analysis, to draw conclusions about the effectiveness of treatments, the impact of policies, and the validity of theories, among others.