UNIT IV

Inferences concerning Proportions

Estimations of Proportions:

The estimation of a proportion is the number of times, X, that an appropriate event occurs in n trails, occasions, or observations. The point estimator of the population proportion, itself, is usually the sample proportion $p = \frac{X}{n}$, namely, the proportion of the time that the event actually occurs.

Large sample confidence interval for proportion 'p' is

$$\frac{x}{n} - z_{\alpha/2}\sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}} < p < \frac{x}{n} + z_{\alpha/2}\sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}}$$

Where the degree of confidence is (1-α)100%.

The proportion of success $p = \frac{x}{n}$

Problem 1: If x = 36 of n = 100 persons interviewed are familiar with the tax incentives for installing certain energy-saving devices, construct a 95% confidence interval for the corresponding true proportion.

Solution: Given that x =36 and n = 100 and confidence is 95%

Then α=0.05 and $z_{\alpha/2} = z_{0.025} = 1.96$

Using above values in the confidence interval for p,

$$\frac{x}{n} - z_{\alpha/2}\sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}} < p < \frac{x}{n} + z_{\alpha/2}\sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}}$$

$$\frac{36}{100}-1.96\sqrt{\frac{\frac{36}{100}\left(1-\frac{36}{100}\right)}{100}}<p<\frac{36}{100}+1.96\sqrt{\frac{\frac{36}{100}\left(1-\frac{36}{100}\right)}{100}}$$

$$0.266<p<0.454$$

Note that the maximum error of estimate $E=z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$

Here $p=\frac{x}{n}$

<u>Problem 2</u>: In a sample survey conducted in a large city, 136 of 400 persons answered yes to the question of whether their city's public transportation is adequate. With 99% confidence, what can we say about the maximum error, if $\frac{x}{n}=\frac{136}{400}=0.34$ is used as an estimate of the corresponding true proportion?

<u>Solution:</u> Since $\frac{x}{n}=\frac{136}{400}=0.34$ and confidence is 99%, then α=0.01 and $z_{\alpha/2}=z_{0.005}=2.575$.

Then maximum error of estimate $E=z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$

$$E=2.575\sqrt{\frac{(0.34)(0.66)}{400}}=0.061$$

<u>Note:</u> To find the sample size when sample proportion is known is

$$n=p(1-p)\left[\frac{z_{\alpha/2}}{E}\right]^2 \text{ and}$$

When sample proportion p is unknown, $n=\frac{1}{4}\left[\frac{z_{\alpha/2}}{E}\right]^2$ (take p = 1/ 2)

Problem 3: What is the size of the smallest sample required to estimate an unknown proportion of customers who would pay for an additional service, to within a maximum error of 0.06 with at least 95% confidence?

Solution: From the given data maximum error of estimate E = 0.06,

Confidence is 95%, α= 0.05 and hence $z_{\alpha/2} = z_{0.025} = 1.96$

Then sample size $n = \frac{1}{4}\left[\frac{z_{\alpha/2}}{E}\right]^2$

$$= \frac{1}{4}\left[\frac{1.96}{0.06}\right]^2 = 266.77 \approx 267$$

Problem 4: In a random sample of 200 claims filed against an insurance company writing collision insurance on cars, 84 exceeded $3,500. Construct a 95% confidence interval for the true proportion of claims filed against this insurance company that exceed $3,500, using the large sample confidence interval formula.

Solution: Given that x =84 and n = 200 and confidence is 95%

Then α=0.05 and $z_{\alpha/2} = z_{0.025} = 1.96$

Using above values in the confidence interval for p,

$$\frac{x}{n} - z_{\alpha/2}\sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}} < p < \frac{x}{n} + z_{\alpha/2}\sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}}$$

$$\frac{84}{200} - 1.96\sqrt{\frac{\frac{84}{200}\left(1-\frac{84}{200}\right)}{200}} < p < \frac{84}{200} + 1.96\sqrt{\frac{\frac{84}{200}\left(1-\frac{84}{200}\right)}{200}}$$

$$0.3516 < p < 0.4200$$

Problem 5: In a random sample of 200 claims filed against an insurance company writing collision insurance on cars, 84 exceeded $3,500. What we say with 99% confidence about the maximum error if we use the sample proportion as an estimate of the true proportion of claim field against this insurance company.

Solution: we know that maximum error of estimate $E = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$

Here sample size n = 200, x = 84 and confidence is 99%.

Hence α = 1- 0.99 = 0.01

Then $z_{\alpha/2} = z_{0.005} = 2.575$

The maximum error of estimate is $E = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$

$$= 2.575\sqrt{\frac{\frac{84}{200}\left(1-\frac{84}{200}\right)}{200}}$$
$$= 0.08987$$

Problem 6: In a random sample of 400 industrial accidents, it was found that 231 were due at least partially to unsafe working conditions. Construct a 99% confidence interval for the corresponding true proportion using the large sample confidence interval formula.

Solution: Large sample confidence interval for p is

$$\frac{x}{n} - z_{\alpha/2}\sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}} < p < \frac{x}{n} + z_{\alpha/2}\sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}}$$

Where the degree of confidence is (1-α)100%.

From the given data n = 400, x = 231, α = 1-0.99= 0.01, $z_{\alpha/2} = z_{0.005} = 2.575$

Then 99% confidence interval is

$$\frac{231}{400} - 2.575\sqrt{\frac{\frac{231}{400}\left(1-\frac{231}{400}\right)}{400}} < p < \frac{231}{400} + 2.575\sqrt{\frac{\frac{231}{400}\left(1-\frac{231}{400}\right)}{400}}$$

$$0.5775 - 0.0636 < p < 0.5775 + 0.0636$$
$$0.5139 < p < 0.6411$$

<u>Problem 7:</u> In a random sample of 90 sections of pipe in a chemical plant, 15 showed signs of serious corrosion. Construct a 95% confidence interval for the true proportion of pipe sections showing signs of serious corrosion, using the large sample confidence interval forumula.

<u>Solution:</u>Large sample confidence interval for p is

$$\frac{x}{n} - z_{\alpha/2}\sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}} < p < \frac{x}{n} + z_{\alpha/2}\sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}}$$

Where the degree of confidence is (1-α)100%.

From the given data n = 90, x = 15, α = 1-0.95= 0.05, $z_{\alpha/2} = z_{0.025} = 1.96$

Then 99% confidence interval is

$$\frac{15}{90} - 1.96\sqrt{\frac{\frac{15}{90}\left(1-\frac{15}{90}\right)}{90}} < p < \frac{15}{90} + 1.96\sqrt{\frac{\frac{51}{90}\left(1-\frac{15}{90}\right)}{90}}$$
$$0.1667 - 0.0770 < p < 0.1667 + 0.0770$$
$$0.0897 < p < 0.2437$$

<u>Hypothesis concerning One Proportion</u>

Here we test the null hypothesis $p = p_0$ against one of the alternative hypothesis $p < p_0, p > p_0, or\ p \neq p_0$ with the use of the statistic

$$Z = \frac{X - np_0}{\sqrt{n\,p_0\left(1 - p_0\right)}}$$

Which is a random variable having approximately the standard normal distribution.

Critical region for Testing $p = p_0$	
Alternative Hypothesis	Reject null hypothesis if:
$p < p_0$	$Z < -z_\alpha$
$p > p_0$	$Z > z_\alpha$
$p \neq p_0$	$Z < -z_{\alpha/2}$ or $Z > z_{\alpha/2}$

<u>Problem 8:</u> Transceivers provide wireless communication among electronic components of consumer products. Responding to a need for a fast, low-cost test of Bluetooth-capable transceivers, engineers developed a product test at the wafer level. In one set of trails with 60 devices selected from different wafer lots, 48 devices passed. Test the null hypothesis p = 0.70 against the alternative hypothesis p > 0.70 at the 0.05 level of significance.

<u>Solution:</u> Null hypothesis: p =0.70

Alternative hypothesis : p > 0.70

Level of significance α = 0.05

Then $z_{0.05}$=1.645

The null hypothesis is rejected if Z > 1.645 where the test statistic

$$Z = \frac{X - np_0}{\sqrt{np_0\left(1 - p_0\right)}}$$

From the given x =48, n =60, and p_0=0.70

Then $Z = \frac{X - np_0}{\sqrt{np_0(1-p_0)}} = \frac{48 - 60(0.70)}{\sqrt{60(0.70)(0.30)}} = 1.69$

Since z = 1.69 is greater than 1.645, the null hypothesis is rejected. So, we accept alternative hypothesis. That is p > 0.70 is accepted.

<u>Problem 9:</u> A manufacturer of submersible pumps claims that at most 30% of the pumps require repairs within the first 5 years of operation. If a random sample of 120 of these pumps includes 47 which required repairs within the first 5 year, test the null hypothesis p = 0.30against the alternative hypothesis p > 0.30 at the 0.05 level of significance.

<u>Solution:</u> Null hypothesis: p =0.30

Alternative hypothesis : p > 0.30

Level of significance α = 0.05

Then $z_{0.05}$=1.645

The null hypothesis is rejected if Z > 1.645 where the test statistic $Z = \frac{X - np_0}{\sqrt{np_0(1-p_0)}}$

From the given x =47, n =120, and p_0=0.30

Then $Z = \frac{X - np_0}{\sqrt{np_0(1-p_0)}} = \frac{47 - 120(0.30)}{\sqrt{120(0.30)(0.70)}} = 2.1913$

Since z = 2.1913 is greater than 1.645, the null hypothesis is rejected. So, we accept alternative hypothesis. That is p > 0.30 is accepted.

<u>Problem 10:</u> The performance of a computer is observed over a period of 2 years to check the claim that the probability is 0.20 that its downtime will exceed 5 hours in any given week. Testing the null hypothesis p = 0.20 against the alternative hypothesis p ≠ 0.20, what can we conclude at the level of

significance α= 0.05, if there we only 11 weeks in which the downtime of the computer exceeded 5 hours?

Solution: Null hypothesis: p =0.20

Alternative hypothesis : p ≠ 0.20

Level of significance α = 0.05

Then $z_{0.025}$=1.96

The null hypothesis is rejected if Z <-1.96 or Z>1.96 where the test statistic

$$Z = \frac{X - np_0}{\sqrt{np_0(1-p_0)}}$$

From the given x =11, n =104, and p_0=0.20

Then $Z = \frac{X - np_0}{\sqrt{np_0(1-p_0)}} = \frac{11-104(0.20)}{\sqrt{105(0.20)(0.80)}} = -2.391$

Since z = -2.391 is less than -1.96, the null hypothesis is rejected. So, we accept alternative hypothesis. That is p ≠ 0.30 is accepted.

Problem 11: To check on an ambulance service's claim that at least 40% of its calls are life-threatening emergencies, a random sample was taken from its files, and it was found that only 49 of 150 calls were life-threatening emergencies. Can the null hypothesis p = 0.40 be rejected against the alternative hypothesis

P < 0.40 if the probability of a Type-I error is to be at most 0.01?

Solution: Null hypothesis: $p \geq 0.40$

Alternative hypothesis : p < 0.40

Level of significance α ≤0.05

Here we test the null hypothesis $p = 0.40$ against the alternative hypothesis p < 040 at the level of significance α= 0.05

Then $z_{0.05}$=1.645

The null hypothesis is rejected if Z <-1.645 where the test statistic $Z = \frac{X - np_0}{\sqrt{np_0(1-p_0)}}$

From the given x =49, n =150, and p_0=0.40

Then $Z = \frac{X - np_0}{\sqrt{np_0(1-p_0)}} = \frac{49 - 150(0.40)}{\sqrt{150(0.40)(0.60)}} = -1.8333$

Since z = -1.8333 is less than -1.645, the null hypothesis is rejected. So, we accept alternative hypothesis. That is p <0.40 is accepted.

<u>Problem 12:</u> In a random sample of 600 cars making a right turn at a certain intersections, 157 pulled into the wrong lane. Test the null hypothesis that actually 30% of all drivers make this mistake at the given intersection, using the alternative hypothesis p ≠ 0.30 and the level of significance

(a) α = 0.05 (b) α = 0.01

<u>Solution:</u>Null hypothesis: $p = 0.30$

Alternative hypothesis : $p \neq 0.30$

Here we test the null hypothesis $p = 0.30$ against the alternative hypothesis p ≠ 0.30 The null hypothesis is rejected if Z < $-Z_{\alpha/2}$ or Z > $Z_{\alpha/2}$where the test statistic $Z = \frac{X - np_0}{\sqrt{np_0(1-p_0)}}$

From the given x =600, n =157, and $p_o = 0.30$

(i) at the level of significance α= 0.05

Then $z_{0.025}$=1.96

Then $Z = \frac{X - np_0}{\sqrt{np_0(1-p_0)}} = \frac{157 - 600(0.30)}{\sqrt{600(0.30)(0.70)}} = -2.049$

Since z = -2.0491 is less than -1.96, the null hypothesis is rejected. So, we accept alternative hypothesis. That is p ≠0.30 is accepted.

(ii) at the level of significance α= 0.01

Then $z_{0.005}$=2.575

Then $Z = \frac{X - np_0}{\sqrt{np_0(1-p_0)}} = \frac{157 - 600(0.30)}{\sqrt{600(0.30)(0.70)}} = -2.049$

Since z = -2.0491 is between -2.575 and 2.575l, the null hypothesis is accepted. That is p = 0.30 is accepted.

Conclusion: The null hypothesis is rejected with 95% confidence and accepted with 99% confidence.

HYPOTHESIS CONCERING SEVERAL PROPOPRTIONS

Many engineering problems concern a random variable that follows the binomial distribution.For example, consider a production process that manufactures items that are classifiedas either acceptable or defective. Modeling the occurrence of defectives with the binomialdistribution is usually reasonable when the binomial parameter *p* represents the proportionof defective items produced. Consequently, many engineering decision problems involvehypothesis testing about *p*.

Suppose that we are interested in testing whether two or more binomial populations have the same parameter p. Let us consider k different binomial populations whose parameters are respectively p_1, p_2,…,p_k. Now we are interested in testing the null hypothesis p_1= p_2 =…= p_k = p against the

alternative hypothesis that these population proportions are not all equal. To perform a suitable large sample test of this hypothesis, we require independent random samples of size $n_1, n_2, \ldots, n_k$ from k different populations. The number of successes and failures in each of these k samples are given by the following table:

	Sample 1	Sample 2		Sample k	Total
Successes	x_1	x_2	…	x_k	x
Failures	$n_1 - x_1$	$n_2 - x_2$	…	$n_k - x_k$	n - x
Total	n_1	n_2	…	n_k	n

In the above table x represents the total number of successes, n – x represents the total number of failures and n the total number of trails. The entry in the cell belonging to the ith row and j^{th} column is called the observed frequency o_{ij} with i= 1, 2 and j = 1, 2, …, k. Let us denote the observed proportion of success by $\overline{p}$. So, the value of $\overline{p}$ is given by $\overline{p} = \frac{x}{n}$.

Hence the expected number of successes and failures for the jth sample are estimated by the following formulae:

$$e_{1j} = n_j \overline{p} == n_j \frac{x}{n}$$

$$e_{2j} = n_j(1-\overline{p}) = n_j(1-\frac{x}{n}) = n_j(\frac{n-x}{n})$$

The quantities e_{1j} and e_{2j} are called the expected cell frequencies for j =1, 2, …, k.

The test statistic for test concerning difference among proportions is given by

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{k}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$$

Decision: Reject the null hypothesis if the value of χ^2 exceeds χ^2_α with k – 1 degrees of freedom.

<u>Problem 13</u>: Samples of three kinds of materials subjected t extreme temperature changes, produced the results shown in the following table:

	Material A	Material B	Material C	Total
Crumbled	41	27	22	90
Remained intact	79	53	78	210
Total	120	80	100	300

Use the 0.05 level of significance to test whether, under the stated conditions, the probability of crumbing is the same for the three kinds of materials.

<u>Solution</u>:

Null Hypothesis, H_0: $p_1 = p_2 = p_3$

Alternative Hypothesis, H_1: p_1 , p_2 and p_3 are not all equal.

Level of significance, $\alpha = 0.05$

$$\bar{p} = \frac{90}{300}$$

The expected frequencies for the cells are given as follows:

e_{11} = 120x90/300 = 36

e_{12} = 80x90/300 = 24

e_{13} = 90 – (36 + 24) = 30 (since, 36 + 24 + e_{13} =90)

e_{21} = 120 – 36 = 84 (since, 36 + e_{21} = 120)

e_{22} = 80 – 24 = 56

e_{23} = 100 – 30 = 70

Test statistic, $\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{3}\frac{(o_{ij} - e_{ij})^2}{e_{ij}}$

$= (41 - 36)^2/36 + (27 - 24)^2/24 + (22 - 30)^2/30 + (79 - 84)^2/84$

$+ (53 - 56)^2/56 + (78 - 70)^2/70$

$= 4.575$

$\chi^2_{0.05}(3-1)\,d.f = 5.991$

Since $\chi^2 = 4.575$ does notexceeds $\chi^2_{0.05}$ (3 – 1) d.f = 5.991, we can't reject the null hypothesis at the 0.05 level of significance. Hence the probability of crumbling is the same for the three kinds of material.

Problem 14: Four methods are under development for making disks of a superconductivity material. Fifty disks are made by each method and they are checked for superconductivity when cooled with liquid nitrogen.

	Method 1	Method 2	Method 3	Method 4	Total
Superconductors	31	42	22	25	120
Failures	19	8	28	25	80
Total	50	50	50	50	200

Perform a chi square test with α = 0.05 to test whether the probability of superconductivity is the same for the four kinds of methods.

Solution:

Null Hypothesis, H_0: $p_1 = p_2 = p_3 = p_4$

Alternative Hypothesis, H_1: p_1 , p_2 , p_3 and pare not all equal.

Level of significance, α = 0.05

$$\bar{p} = \frac{120}{200}$$

The expected frequencies for the cells are given as follows:

e_{11} = 50x120/200 = 30

e_{12} = 50x120/200 = 30

e_{13} = 50x120/200 = 30

e_{14} = 50x120/200 = 30

e_{21} = 50 – 30 = 20

e_{22} = 50 – 30 = 20

e_{23} = 50 – 30 = 20

e_{24} = 50 – 30 = 20

Test statistic, $\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{4}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$

$= (31 - 30)^2/30 + (42 - 30)^2/30 + (22 - 30)^2/30 + (25 - 30)^2/30$

$+ (19 - 20)^2/20 + (8 - 20)^2/20 + (28 - 20)^2/20 + (25 - 20)^2/20$

$= 19.50$

$\chi^2_{0.05}(4-1)\, d.f = 7.815$

Since $\chi^2 = 19.50$ exceeds $\chi^2_{0.05}$ (4 – 1) d.f = 7.815, we reject the null hypothesis at the 0.05 level of significance. Hence the probability of superconductivity is not the same for four methods.

<u>Problem 15</u>: The following data come from a study in which random samples of the employees of three government agencies were asked questions about their pension plan:

	Agency 1	Agency 2	Agency 3	Total
For the pension plan	67	84	109	260
Against the pension plan	33	66	41	140
Total	100	150	150	400

Use the 0.01 level of significance to test the null hypothesis that the actual proportions of employees favoring the pension plan are the same.

<u>Solution</u>:

Null Hypothesis, H_0: $p_1 = p_2 = p_3$

Alternative Hypothesis, H_1: p_1, p_2 and p_3 are not all equal.

Level of significance, $\alpha = 0.01$

$\bar{p} = \frac{260}{400}$

The expected frequencies for the cells are given as follows:

$e_{11} = 100x260/400 = 65$

$e_{12} = 150x260/400 = 97.5$

$e_{13} = 260 – (65 + 97.5) = 97.5$

$e_{21} = 100 – 65 = 35$

$e_{22} = 150 – 97.5 = 52.5$

$e_{23} = 150 – 97.5 = 52.5$

Test statistic, $\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{3}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$

$= (67 - 65)^2/65 + (84 – 97.5)^2/97.5 + (109 – 97.5)^2/97.5$

$+ (33 - 35)^2/35 + (66 – 52.5)^2/52.5 + (41 – 52.5)^2/52.5$

$= 9.39$

$\chi^2_{0.01}(3-1)\, d.f = 9.210$

Since $\chi^2 = 9.39$ exceeds $\chi^2_{0.01}$ (3 – 1) d.f = 9.210, we have to reject the null hypothesis at the 0.01 level of significance. Hence the probability for favoring the pension plan by the three agencies is not the same.

Hypothesis concerning two proportions

This is a particular case of several proportions with k = 2. In this case we proceed as per the given below procedure:

Null Hypothesis, H_0: $p_1 = p_2$

Alternative Hypothesis, H_1 : $p_1 < p_2$ (or) $p_1 > p_2$ (or) $p_1 \neq p_2$

Given x_1, x_2, n_1 and n_2

Level of significance = α

Test Statistic, $Z = \dfrac{\frac{X_1}{n_1} - \frac{X_2}{n_2}}{\sqrt{\bar{p}(1-\bar{p})}\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})}}$ $with\ \bar{p} = \dfrac{X_1 + X_2}{n_1 + n_2}$

Critical Regions for testing the Null Hypothesis, H_0: $p_1 = p_2$

Alternate Hypothesis	Reject null hypothesis if
$p_1 < p_2$	$Z < -z_\alpha$

$p_1 > p_2$	$Z > z_\alpha$
$p_1 \neq p_2$	$Z < -z_{\alpha/2}$ or $Z > z_{\alpha/2}$

Finally, we have to write the decision that either accepting the null hypothesis or rejecting the null hypothesis.

Also the (1 – α)100% large sample confidence interval for the difference of two proportions is given by

$$\frac{x_1}{n_1} - \frac{x_2}{n_2} \pm z_{\alpha/2}\sqrt{\frac{\frac{x_1}{n_1}\left(1-\frac{x_1}{n_1}\right)}{n_1} + \frac{\frac{x_2}{n_2}\left(1-\frac{x_2}{n_2}\right)}{n_2}}$$

<u>Problem 16</u>: A study shows that 16 of 200 tractors produced on one assembly line required extensively adjustments before they could be shipped. While the same was true for 14 of 400 tractors produced on another assembly line. At the 0.01 level of significance, does this support the claim that the second production line does superior work? Also construct a 95% confidence interval for $p_1 - p_2$.

<u>Solution</u>:

Null Hypothesis, H_0: $p_1 = p_2$

Alternative Hypothesis, H_1 : $p_1 > p_2$

Given $x_1 = 16$, $x_2 = 14$, $n_1 = 200$ and $n_2 = 400$

Level of significance, $\alpha = 0.01$

Test Statistic, $Z = \dfrac{\frac{X_1}{n_1} - \frac{X_2}{n_2}}{\sqrt{\bar{p}(1-\bar{p})}\sqrt{\left(\frac{1}{200}+\frac{1}{400}\right)}}$ $with\ \bar{p} = \dfrac{X_1 + X_2}{n_1 + n_2}$

$$= \frac{\frac{16}{200} - \frac{14}{400}}{\sqrt{0.05(1-0.05)}\sqrt{\left(\frac{1}{200}+\frac{1}{400}\right)}} \quad with\ \bar{p} = \frac{16+14}{200+400} = 0.05$$

$$= \frac{0.045}{\sqrt{(0.0475)(0.0075)}}$$

$$= 2.384$$

From table 3, $Z_{0.01} = 2.33$

Critical Regions for testing the Null Hypothesis, H_0: $p_1 = p_2$

Alternate Hypothesis	Reject null hypothesis if
$p_1 > p_2$	$Z > z_\alpha$

Decision: Since Z = 2.384 exceeds $Z_{0.01}$ = 2.33, we have to reject the null hypothesis. So, accept the alternative hypothesis. That is the true proportion of tractors requiring extensive adjustments is greater for first assembly line than for the second.

A 95% confidence interval for the difference of two proportions is given by

$$\left(\frac{x_1}{n_1} - \frac{x_2}{n_2}\right) \pm z_{\alpha/2}\sqrt{\frac{\frac{x_1}{n_1}\left(1-\frac{x_1}{n_1}\right)}{n_1} + \frac{\frac{x_2}{n_2}\left(1-\frac{x_2}{n_2}\right)}{n_2}}$$

$$\Rightarrow \quad ((0.08-0.035) \pm z_{0.025}\sqrt{\frac{0.08(1-0.08)}{200} + \frac{0.035(1-0.035)}{400}}$$

$$\Rightarrow (0.08-0.035) \pm 1.96\sqrt{\frac{0.08(1-0.08)}{200} + \frac{0.035(1-0.035)}{400}}$$

$$\Rightarrow \quad 0.045 \pm 1.96\sqrt{0.000368 + 0.00008443}$$

$$\Rightarrow \quad 0.003 < p_1 - p_2 < 0.087$$

<u>Problem 17</u>: Photolithography plays a central role in manufacturing integrated circuits made on thin disks of silicon. Prior to a quality improvement program, too many rework operations were required. In a sample of 200 units, 26 required reworking of the photolithographic step. Following training in the use of pareto charts and other approaches to identify significant problems, improvements were made. A new sample of size 200 had only 12 that needed rework. Is this sufficient evidence at the 0.01 level of significance that the improvements have been effective in reducing rework?

Solution:

Null Hypothesis, H_0: $p_1 = p_2$

Alternative Hypothesis, H_1 : $p_1 > p_2$

Given $x_1 = 26$, $x_2 = 12$, $n_1 = 200$ and $n_2 = 200$

Level of significance, $\alpha = 0.01$

Test Statistic, $Z = \dfrac{\dfrac{X_1}{n_1} - \dfrac{X_2}{n_2}}{\sqrt{\bar{p}(1-\bar{p})}\sqrt{\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$ $with\ \bar{p} = \dfrac{X_1 + X_2}{n_1 + n_2}$

$$= \frac{\dfrac{26}{200} - \dfrac{12}{200}}{\sqrt{0.095(1-0.095)}\sqrt{\left(\dfrac{1}{200} + \dfrac{1}{200}\right)}} \quad with\ \bar{p} = \frac{26+12}{200+200} = 0.095$$

$$= \frac{0.07}{(0.2932)(0.1)}$$

$$= 2.3873$$

$Z_{0.01} = 2.33$

Critical Regions for testing the Null Hypothesis, H_0: $p_1 = p_2$

Alternate Hypothesis	Reject null hypothesis if
$p_1 > p_2$	$Z > z_\alpha$

Decision: Since $Z = 2.3873$ exceeds $Z_{0.01} = 2.33$, we have to reject the null hypothesis. So, accept the alternative hypothesis.

Problem 18:The owner of a machine shop must decide which of two snack-vending machines to install in his shop. If each machine is tested for 250 times and the first machine fails to work(neither delivers the snack nor returns the money) 13 times and the second machine fails to work 7 times, test at the 0.05 level of significance whether the difference between the corresponding sample proportions is significant.

Solution:

Null Hypothesis, H_0: $p_1 = p_2$

Alternative Hypothesis, H_1 : $p_1 \neq p_2$

Given $x_1 = 13$, $x_2 = 7$, $n_1 = 250$ and $n_2 = 250$

Level of significance, $\alpha = 0.05$

Test Statistic, $Z = \dfrac{\dfrac{X_1}{n_1} - \dfrac{X_2}{n_2}}{\sqrt{\bar{p}(1-\bar{p})}\sqrt{(\dfrac{1}{n_1} + \dfrac{1}{n_2})}}$ $with\ \bar{p} = \dfrac{X_1 + X_2}{n_1 + n_2}$

$$= \frac{\dfrac{13}{250} - \dfrac{7}{250}}{\sqrt{0.04(1-0.04)}\sqrt{\left(\dfrac{1}{250} + \dfrac{1}{250}\right)}} \quad with\ \bar{p} = \frac{13+7}{250+250} = 0.04$$

$$= \frac{0.024}{(0.1959)(0.0.0844)}$$

$= 1.369$

$Z_{0.025} = 1.96$

Critical Regions for testing the Null Hypothesis, H_0: $p_1 = p_2$

Alternate Hypothesis	Reject null hypothesis if
$p_1 \neq p_2$	$Z < -z_{\alpha/2}$ or $Z > z_{\alpha/2}$

Decision: Since Z = 1.369 does not exceed $Z_{0.025} = 1.96$, we can't reject the null hypothesis. So, accept the null hypothesis.

Home work: A study showed that 64 of 180 persons who saw a photocopying machine advertised during the telecast of a baseball game and 75 of 180 other persons who saw it advertised on a variety show remembered the brand name 2 hours later. Use the 0.05 level of significance whether the difference between the corresponding sample proportions is significant?

Completely Randomized Designs

One way ANOVA

Suppose that there are k independent random samples from k different populations and we are interested to test whether the means of these k populations are all equal. Let us denoted the jth observation in the ith sample by y_{ij} and the data for one way classification is as follows:

	Observations					
Sample 1	y_{11}	y_{12}	…	y_{1j}	…	y_{1n_1}
Sample 2	y_{21}	y_{22}	…	y_{2j}	…	y_{2n_2}
…						
Sample i	y_{i1}	y_{i2}	…	y_{ij}	…	y_{in_i}
…						
Sample k	y_{k1}	y_{k2}	…	y_{kj}	…	y_{kn_k}

From the above data compute the following quantities:

$$T_{\bullet} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} \qquad N = \sum_{i=1}^{k} n_i \qquad \bar{y} = \frac{T_{\bullet}}{N}$$

$$C = \frac{T_{\bullet}^{2}}{N} \qquad T_i = \sum_{j=1}^{n_i} y_{ij}$$

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}^2 - C$$

$$SS(Tr) = \sum_{i=1}^{k} \frac{T_i^2}{n_i} - C$$

SSE = SST – SS(Tr)

Null Hypothesis, $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$

Alternative Hypothesis, H_1: The k population means are not all equal.

Level of significance: α

Analysis of variance table:

Source of variation	Degrees of freedom	Sum of squares	Mean square	F
Treatments	K - 1	SS(Tr)	MS(Tr) = SS(Tr)/(k – 1)	MS(Tr)/MSE
Error	N - k	SSE	MSE = SSE/(N – k)	
Total	N - 1	SST		

Decision: Reject the null hypothesis if F exceeds $F_\alpha(k - 1, N - k)$.

Problem 1: Suppose that a sheet of tin plate, sufficiently long and wide, is selected and that the 48 disks are cut as circles. The 12 disks cut from strip 1 are sent to the laboratory 1, The 12 disks cut from strip 2 are sent to the laboratory 2, and so forth. Each laboratory measures the tin-coating weights of 12 disks and that the results are as follows:

Laboratory 1	Laboratory 2	Laboratory 3	Laboratory 4
0.25	0.18	0.19	0.23
0.27	0.28	0.25	0.30
0.22	0.21	0.27	0.28
0.30	0.23	0.24	0.28

0.27	0.25	0.18	0.24
0.28	0.20	0.26	0.34
0.32	0.27	0.28	0.20
0.24	0.19	0.24	0.18
0.31	0.24	0.25	0.24
0.26	0.22	0.20	0.28
0.21	0.29	0.21	0.22
0.28	0.16	0.19	0.21

Perform an analysis of variance to test at the 0.05 level of significance whether the differences among the sample means at the four laboratories are significant.

Solution: Given k = 4 samples and the size of each sample is $n_i = 12$, i = 1,2,3.4

$T_1 = 3.21$ (the sum of all the values under laboratory 1)

$T_2 = 2.72$ (the sum of all the values under laboratory 2)

$T_3 = 2.76$ (the sum of all the values under laboratory 3)

$T_4 = 3.00$ (the sum of all the values under laboratory 4)

The grand total, $T_\bullet = T_1 + T_2 + T_3 + T_4 = 11.69$

$N = n_1 + n_2 + n_3 + n_4 = 12 + 12 + 12 + 12 = 48$

$$C = \frac{T_\bullet^2}{N} = \frac{11.69^2}{48} = 2.8470$$

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - C$$

$$= (0.25)^2 + (0.27)^2 + (0.22)^2 + \ldots + (0.22)^2 + (0.21)^2 - 2.8470 = 0.0809$$

$$SS(Tr) = \sum_{i=1}^{k} \frac{T_i^2}{n_i} - C$$

$$= \frac{(3.21)^2}{12} + \frac{(2.72)^2}{12} + \frac{(2.76)^2}{12} + \frac{(3.00)^2}{12} - 2.8470 = 0.0130$$

SSE = SST – SS(Tr) = 0.0809 – 0.0130 = 0.0679

$F_{0.05}(4-1,48-4) = F_{0.05}(3,44) = 2.82$

Analysis of variance table:

Source of variation	Degrees of freedom	Sum of squares	Mean square	F
Treatments	k – 1 **3**	SS(Tr) **0.0130**	MS(Tr) = SS(Tr)/(k – 1) **0.0043**	MS(Tr)/MSE **2.87**
Error	N - k **44**	SSE **0.0679**	MSE = SSE/(N – k) **0.0015**	
Total	N – 1 **47**	SST **0.0809**		

Decision: Since F = 2.87 exceeds the value of $F_{0.05}(3,44) = 2.82$, the null hypothesis can be rejected at the 0.05 level of significance. So, we conclude that the laboratories are not obtaining consistent results.

Problem 2: As a part of the investigation of the collapse of the roof of a building, a testing laboratory is given all the available bolts that connected the steel structure at three 3 different positions on the roof. The forces required to shear each of these bolts are as follows:

Position 1	90	82	79	98	83	91	
Position 2	105	89	93	104	89	95	86
Position 3	83	89	80	94			

Perform an analysis of variance to test at the 0.05 level of significance whether the differences among the sample means at the 3 positions are significant.

Solution: Given k = 3 samples and the sizes are $n_1 = 6$, $n_2 = 7$ and $n_3 = 4$

$T_1 = 523$ (the sum of all the values under Position 1)

$T_2 = 661$ (the sum of all the values under Position 2)

$T_3 = 346$ (the sum of all the values under Position 3)

The grand total, $T_{\bullet} = T_1 + T_2 + T_3 = 1530$

$N = n_1 + n_2 + n_3 = 6 + 7 + 4 = 17$

$$C = \frac{T_{\bullet}^2}{N} = \frac{1530^2}{17} = 137700$$

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - C$$

$= (90)^2 + (82)^2 + (79)^2 + \ldots + (80)^2 + (94)^2 - 137700$

$= 138638 - 137700 = 938$

$$SS(Tr) = \sum_{i=1}^{k} \frac{T_i^2}{n_i} - C$$

$$= \frac{(523)^2}{6} + \frac{(661)^2}{7} + \frac{(346)^2}{4} - 137700 = 234$$

SSE = SST – SS(Tr) = 938 – 234 = 704

$F_{0.05}(3\text{-}1,17\text{-}3) = F_{0.05}(2,14) = 3.74$

Analysis of variance table:

Source of variation	Degrees of freedom	Sum of squares	Mean square	F
Treatments	k – 1 **2**	SS(Tr) **234**	MS(Tr) = SS(Tr)/(k – 1) **117**	MS(Tr)/MSE **2.33**
Error	N - k **14**	SSE **704**	MSE = SSE/(N – k) **50.3**	
Total	N – 1 **16**	SST **938**		

Decision: Since F = 2.33 does not exceed the value of $F_{0.05}(2,14) = 3.74$, the null hypothesis cannot be rejected. So, accept the null hypothesis. That is there is no difference in the mean shear strengths of the bolts at the three different positions on the roof.

Problem 3: The following are the number of mistakes made in 5 successive days for 4 technicians working for a photographic laboratory:

Technician-I	Technician-II	Technician-III	Technician-IV

6	14	10	9
14	9	12	12
10	12	7	8
8	10	15	10
11	14	11	11

Test at the level of significance α = 0.01 whether the differences among 4 sample means can be attributed to chance.

Solution: Given k = 4 samples and the size of each sample is $n_i = 5$, i = 1,2,3.4

$T_1 = 49$ (the sum of all the values under laboratory 1)

$T_2 = 59$ (the sum of all the values under laboratory 2)

$T_3 = 55$ (the sum of all the values under laboratory 3)

$T_4 = 50$ (the sum of all the values under laboratory 4)

The grand total, $T_\bullet = T_1 + T_2 + T_3 + T_4 = 213$

$N = n_1 + n_2 + n_3 + n_4 = 5 + 5 + 5 + 5 = 20$

$$C = \frac{T_\bullet^2}{N} = \frac{213^2}{20} = 2268.45$$

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - C$$

$$= (6)^2 + (14)^2 + (10)^2 + \ldots + (10)^2 + (11)^2 - 2268.45 = 114.55$$

$$SS(Tr) = \sum_{i=1}^{k} \frac{T_i^2}{n_i} - C$$

$$= \frac{(49)^2}{5} + \frac{(59)^2}{5} + \frac{(55)^2}{5} + \frac{(50)^2}{5} - 2.268.45 = 12.95$$

SSE = SST – SS(Tr) = 101.6

$F_{0.01}(4-1,20-4) = F_{0.01}(3,16) = 5.29$

Analysis of variance table:

Source of variation	Degrees of freedom	Sum of squares	Mean square	F
Treatments	k – 1 **3**	SS(Tr) **12.95**	MS(Tr) = SS(Tr)/(k – 1) **4.3167**	MS(Tr)/MSE **0.6798**
Error	N - k **16**	SSE **101.6**	MSE = SSE/(N – k) **6.35**	
Total	N – 1 **19**	SST **114.55**		

Decision: Since F = 0.6798 does not exceed the value of $F_{0.01}(3,16) = 3.2389$, the null hypothesis can't be rejected at the 0.01 level of significance. So, we conclude that the 4 sample means are all equal.

Problem 4: The following are the weight losses of certain machine parts (in milligrams) due to friction when three different lubricants were used under controlled conditions:

Lub A	12.2	11.8	13.1	11.0	3.9	4.1	10.9	8.4
Lub B	10.9	5.7	13.5	9.4	11.4	15.7	10.8	14.0
Lub C	12.7	19.9	13.6	11.7	18.3	14.3	22.8	20.4

Test at the 0.01 level of significance whether the differences among the means can be attributed to chance. Also estimate the parameters of the model used in the analysis of experiment.

Solution: Given k = 3 samples and the sizes are $n_1 = 8$, $n_2 = 8$ and $n_3 = 8$

$T_1 = 74.8$ (the sum of all the values under Lub A)

$T_2 = 91.4$ (the sum of all the values under Lub B)

$T_3 = 133.7$ (the sum of all the values under Lub C)

The grand total, $T\bullet = T_1 + T_2 + T_3 = 299.9$

$N = n_1 + n_2 + n_3 = 8 + 8 + 8 = 24$

$$C = \frac{T\bullet^2}{N} = \frac{(299.9)^2}{24} = 3747.5$$

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - C$$

$= (12.2)^2 + (11.8)^2 + (13.1)^2 + \ldots + (22.8)^2 + (20.4)^2 - 3747.5$

$= 507.3896$

$$SS(Tr) = \sum_{i=1}^{k} \frac{T_i^2}{n_i} - C$$

$= 230.5858$

SSE = SST – SS(Tr) = 276.8038

$F_{0.01}(3\text{-}1, 24\text{-}3) = F_{0.01}(2,21) = 5.78$

Analysis of variance table:

Source of variation	Degrees of freedom	Sum of squares	Mean square	F
Treatments	k – 1 **2**	SS(Tr) **230.5858**	MS(Tr) = SS(Tr)/(k – 1) **115.2929**	MS(Tr)/MSE **8.7468**
Error	N - k **21**	SSE **276.8038**	MSE = SSE/(N – k) **13.1811**	
Total	N – 1 **23**	SST **507.3896**		

Decision: Since F = 8.7468 exceed the value of $F_{0.01}(2,21) = 5.78$, the null hypothesis must be rejected. So, accept the alternative hypothesis. That is the weight losses are not the same for the three lubricants.

Simple linear regression

Principle of Least squares: Principle of least squares says that for a curve of best fit to the given data points the sum of the squares of the errors is a minimum.

The simple linear regression model is a model with a single regressor x that has a relationship with a response y that is a straight line. This model is given by

$$\hat{y} = a + bx \quad \text{---------- (1)}$$

where a is the intercept, b is the slope.

The parameters a and b are called regression coefficients.

<u>Least Squares Estimation of the parameters:</u>

We will use the method of least squares to estimate the parameters a and b in (1). That is we will estimate a and b so that the sum of squares of the differences between the observations y_i and the straight line $\hat{y} = a + bx$ is a minimum.

From (1), we have

$$y_i = a + bx_i \quad \text{---------- (2)}, \quad i = 1,2,\ldots\ldots n$$

Equation (1) may be viewed as a population regression model whereas (4) is a sample regression model, written in terms of the n pairs of data (x_i,y_i), i = 1,2,……. n.

According to least squares criterion,

$$f(a, b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \quad \text{---------- (3) is a minimum.}$$

So we obtain the following equations:

$$\left(\frac{\partial f}{\partial a}\right) = 0 \Rightarrow -2 \sum_{i=1}^{n} (y_i - a - bx_i) = 0$$

$$\left(\frac{\partial f}{\partial b}\right)=0 \Rightarrow -2\sum_{i=1}^{n}(y_i - a - bx_i)x_i = 0$$

Simplifying the above equations, we get

$$\left.\begin{aligned} na + b\sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \\ a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i y_i \end{aligned}\right\} \rightarrow (4)$$

Solving (5), we get $a = \hat{a}$ and $b = \hat{b}$. Note that

$\hat{a}$ and $\hat{b}$ are called the least squares estimators of a and b. Here the equations (5) are called the least squares normal equations.

Since $(\bar{x}, \bar{y})$ lies on the least squares line, we have $\bar{y} = \hat{a} + \hat{b}\bar{x}$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} \qquad \rightarrow (5)$$

And $$\hat{b} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \text{--------} (6)$$

Equation (5) is obtained from the first equation of (4), after dividing with n, where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

$\therefore$ The fitted simple linear regression model is given by

$$\hat{y} = \hat{a} + \hat{b}x \rightarrow (7)$$

Equation (7) gives the point estimate of the mean of y for a particular x.

Let us denote $S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

And $$S_{xy} = \sum_{i=1}^{n} x_i\, y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$$

Now equation (6) can be written as $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \rightarrow (7)$

The difference between the observed value y_i and the corresponding fitted value $\hat{y}_i$ is called the residual.

That is, $e_i = y_i - \hat{y}_i$

$= y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x\right)$, i = 1,2,……n $\rightarrow$(8)

<u>Problem 1:</u>The following are measurements of the air velocity and evaporation coefficient of burning fuel droplets in an impulse engine:

Air velocity (cm/sec) :	20	60	100	140	180	220	260	300	340	380
Evo. Coeff(mm^2/sec):	0.18	0.37	0.35	0.78	0.56	0.75	1.18	1.36	1.17	1.65

Fit a simple linear regression line to the above data.

Solution: Let $\hat{y} = \hat{a} + \hat{b}x$ is the simple linear regression line.

Here n = 10

$\sum_{i=1}^{10} x_i = 2,000$, $\sum_{i=1}^{10} x_i^2 = 5,32,000$, $\sum_{i=1}^{10} y_i = 8.35$, $\sum_{i=1}^{10} x_i y_i = 2,175.40$, $\sum_{i=1}^{8} y_i^2 = 9.1097$

$$S_{xx} = \sum_{i=1}^{10} x_i^2 - \frac{\left(\sum_{i=1}^{10} x_i\right)^2}{n}$$

$$= 532000 - \frac{(2000)^2}{10} = 1,32,000$$

$$S_{xy} = \sum_{i=1}^{10} x_i y_i - \frac{\sum_{i=1}^{10} x_i \sum_{i=1}^{10} y_i}{n}$$

$$= 2,175.40 - \frac{(2,000)(8.35)}{10} = 505.40$$

$$S_{yy} = \sum_{i=1}^{10} y_i^2 - \frac{\left(\sum_{i=1}^{10} y_i\right)^2}{n}$$

$$= 9.1097 - \frac{(8.35)^2}{10} = 2.13745$$

Now, $\hat{b} = \frac{S_{xy}}{S_{xx}} = \frac{505.40}{1,32,000} = 0.00383$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

$$= \frac{8.35}{10} - 0.00383\frac{2000}{10}$$

$$= 0.069$$

$\therefore y = 0.069 + 0.00383x$ is the least square regression line.

Or you can solve the normal equations

10 a + 2000b = 8.35

2000a + 532000b = 2175.40 to find the values of a and b.

Problem: The following data pertain to the number of computer jobs per day and the central processing unit time required,

No.of jobs (x)	1	2	3	4	5
CPU Time(y)	2	5	4	9	10

(i) Fit a least squares line $\hat{y} = \hat{a} + \hat{b}x$

(ii) Predict the mean CPU time when $x = 3.5$

Solution: (i) $\sum_i x_i = 15, \sum_i y_i = 30, \sum_i x_i^2 = 55, \sum_i y_i^2 = 226, \sum_i x_i y_i = 110, \bar{x} = 3 \; and \; \bar{y} = 6$

$S_{xx} = 10, \; S_{yy} = 46, S_{xy} = 20$

$$\hat{b} = \frac{S_{xy}}{S_{xx}} = \frac{20}{10} = 2$$

$$\begin{aligned} \hat{a} &= \bar{y} - \hat{b}\bar{x} \\ &= 6 - (2)(3) \\ &= 0 \end{aligned}$$

$\therefore$ The least squares line is $\hat{y} = 2x$

(ii) When x = 3.5, $\hat{y}$ = 2 * 3.5 = 7

Problem: A chemical company, wishing to study the effect of extraction time on the efficiency of an extraction operation, obtained the data shown in the following table:

Extraction time, x (minutes)	Extraction efficiency, y (%)
27	57
45	64
41	80

19	46
35	62
39	72
19	52
49	77
15	57
31	68

Fit a straight line to the given data by the method of least squares and use it to predict the extraction efficiency one can expect when the extraction time is 35 minutes.

<u>Solution</u>:

Let $\hat{y} = \hat{a} + \hat{b}x$ be the least squares straight line.

$$\sum_i x_i = 320, \sum_i y_i = 635, \sum_i x_i^2 = 11490, \sum_i y_i^2 = 41395, \sum_i x_i y_i = 21275, \bar{x} = 32 \text{ and } \bar{y} = 63.5$$

$$S_{xx} = 1250,\ S_{yy} = 1072.5, S_{xy} = 955$$

$$\hat{b} = \frac{S_{xy}}{S_{xx}} = \frac{955}{1250} = 0.764$$

$$\begin{aligned} \hat{a} &= \bar{y} - \hat{b}\bar{x} \\ &= 63.5 - (0.764)(32) \\ &= 39.052 \end{aligned}$$

$\therefore$ The least squares line is $\hat{y} = 39.052 + 0.764x$

When x = 35, $\hat{y}$ = 39.052+0.764*35 = 65.792

<u>Regression :</u>

The process of estimating the best possible values of one variable given the values of another variable through a least squares curve is called regression.

Regression is of two types. One is linear regression and the other one is curvilinear regression.

<u>Curvilinear regression</u>: The process of estimating the best possible values of one variable from the known values of the other variable through a least squares curve is called curvilinear regression.

Suppose we have to fit a curve $y = a + bx + cx^2$ to the given data points (x_i,y_i); $i = 1,2,..,n$

Normal equations to fit the above curve are

$$na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3 = \sum_{i=1}^{n} x_i y_i$$

$$a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4 = \sum_{i=1}^{n} x_i^2 y_i$$

Solving the above equations we will get the values of a, b and c. Let us denote respectively them by $\hat{a}, \hat{b}\ and\ \hat{c}$. So, The least squares best fit to the given data points is $\hat{y} = \hat{a} + \hat{b}x + \hat{c}x^2$

<u>Problem</u>: The following are data on the drying time of a certain varnish and the amount of an additive that is intended to reduce the drying time:

Amount of varnish additive (grams) x	Drying Time (Hours) y
0	12.0
1	10.5
2	10.0
3	8.0
4	7.0
5	8.0
6	7.5
7	8.5
8	9.0

Fit a second degree polynomial $y = a + bx + cx^2$ by the method of least squares. Also predict the drying time of the varnish when 6.5 grams of additive is used.

<u>Solution</u>: Normal equations to fit y = a + bx + cx^2 are

$$na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3 = \sum_{i=1}^{n} x_i y_i$$

$$a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4 = \sum_{i=1}^{n} x_i^2 y_i$$

Here n = 9

Table:

x	**y**	**x^2**	**x^3**	**x^4**	**xy**	**x^2y**
0	12	0	0	0	0	0
1	10.5	1	1	1	10.5	10.5
2	10	4	8	16	20	40
3	8	9	27	81	24	72
4	7	16	64	256	28	112
5	8	25	125	625	40	200
6	7.5	36	216	1296	45	270
7	8.5	49	343	2401	59.5	416.5
8	9	64	512	4096	72	576
36	**80.5**	**204**	**1296**	**8772**	**299**	**1697**

Now we have the equations 9a + 36b + 204c = 80.5

36a + 204b + 1296c = 299

204a + 1296b + 8772c = 1697

Solving the above equations by the known methods we get

A = 12.2, b= -1.85 and c = 0.183

Hence the least squares second degree polynomial is $\hat{y}$ = 12.2 – 1.85x + 0.183x^2.

When x = 6.5, $\hat{y}$ = 12.2 – 1.85*6.5 + 0.183*6.5^2 = 7.9.

Problem: Fit a second degree polynomial $y = a + bx + cx^2$ by the method of least squares to the following data

x	y
1.5	1.3
2	1.6
2.5	2
3	2.7
3.5	3.4
4	4.1

Solution: Normal equations to fit $y = a + bx + cx^2$ are

$$na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3 = \sum_{i=1}^{n} x_i y_i$$

$$a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4 = \sum_{i=1}^{n} x_i^2 y_i$$

Here n = 7

Table:

x	y	x^2	x^3	x^4	xy	x^2y
1	1.1	1	1	1	1.1	1.1
1.5	1.3	2.25	3.375	5.0625	1.95	2.925
2	1.6	4	8	16	3.2	6.4
2.5	2	6.25	15.625	39.063	5	12.5
3	2.7	9	27	81	8.1	24.3
3.5	3.4	12.25	42.875	150.06	11.9	41.65
4	4.1	16	64	256	16.4	65.6
17.5	**16.2**	**50.75**	**161.88**	**548.19**	**47.65**	**154.48**

Now we have the equations $7a + 17.5b + 50.75c = 16.2$

$$17.5a + 50.75b + 161.88c = 47.65$$

$$50.75a + 161.88b + 548.19c = 154.48$$

Solving the above equations by the known methods we get

a = 1.0477, b= - 0.204 and c = 0.2451

Hence the least squares second degree polynomial is $\hat{y}$ = 1.0477 – 0.204x + $0.2451x^2$.

The Regression analysis for studying more than two variables at a time is known as Multiple Regression.

<u>Fitting of other curves</u>

(1) To fit $y = ab^x$ ---------- (1)

Taking natural logarithms on both sides we get

ln y = ln a + x ln b

Y = A + Bx---------- (2)

Where Y = ln y, A = ln a and B = ln b

Normal equations to fit (2) are

$$\left.\begin{aligned} nA + B\sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} Y_i \\ A\sum_{i=1}^{n} x_i + B\sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i Y_i \end{aligned}\right\} \rightarrow (3)$$

Solve (3) for A and B and hence find the values of A and B. Where $a = e^A$

and $b = e^B$.

(2) To fit $y = ax^b$ ---------- (1)

Taking natural logarithms on both sides we get

ln y = ln a + b ln x

i.e., Y = A + bX ---------- (2)

Where Y = ln y, A = ln a and X = ln x

Normal equations to fit (2) are

$$\left.\begin{aligned} nA + b\sum_{i=1}^{n} X_i &= \sum_{i=1}^{n} Y_i \\ A\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} X_i^2 &= \sum_{i=1}^{n} X_i Y_i \end{aligned}\right\} \rightarrow (3)$$

Solve (3) for A and b and hence find the values of a and b. Where a = e^A.

(3) To fit y = ae^{bx} ---------- (1)

Taking natural logarithms on both sides we get

ln y = ln a + bx

i.e., Y = A + bx ---------- (2)

Normal equations to fit (2) are

$$\left.\begin{aligned} nA + b\sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} Y_i \\ A\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i Y_i \end{aligned}\right\} \rightarrow (3)$$

Solve (3) for A and b and hence find the values of a and b. Where a = e^A.

Problem: An experiment gave the following values:

x :	61	26	7	26
y :	350	400	500	600

Use the principle of least squares, fit a curve y = ax^b.

Solution: Normal equations to fit y = ax^b are

$$\left.\begin{aligned} nA + b\sum_{i=1}^{n} X_i &= \sum_{i=1}^{n} Y_i \\ A\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} X_i^2 &= \sum_{i=1}^{n} X_i Y_i \end{aligned}\right\} \rightarrow (1)$$

<u>Table</u>

x	y	X = ln x	Y = ln y	X^2	XY
61	350	4.11	5.86	16.90	24.08
26	400	3.26	5.99	10.62	19.52
7	500	1.95	6.21	3.79	12.09
26	600	3.26	6.40	10.62	20.84
120	**1850**	**12.57**	**24.46**	**41.92**	**76.54**

Then (1) becomes 4A + 12.57b = 24.46 ---------- (2)

12.57A + 41.92b = 76.54 ---------- (3)

Solving (2) and (3) we get A = 6.538 and b = - 0.1346

Now a = $e^{6.538}$ = 690.90 and b = -0.1346. So, the curve is y = $690.90x^{-0.1346}$.

Home work:

(1) Fit a curve y = ab^x to the data given below:

x :	2	3	4	5	6
y :	144	172	207	248	298

(2) Fit a curve y = ae^{bx} to the data given below:

x :	2	4	6	8
y :	25	38	56	84

Multiple Linear regression

Let y denote the dependent variable that is linearly related to k independent variables x_1, x_2, .., x_k through the parameters $\beta_1, \beta_2, ..., \beta_k$. The multiple linear regression model with k parameters is given by

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \quad -----(1)$$

Data for multiple linear regression

Observations	Response y	Regressors			
		x_1	x_2		x_k
1	y_1	x_{11}	x_{12}		x_{1k}
2	y_2	x_{21}	x_{22}		x_{2k}
n	y_n	x_{n1}	x_{n2}		x_{nk}

The sample regression model corresponding to (1) is given by

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i, i = 1, 2, ..., n ----(2)$$

The least squares function is given by

$$S(\beta_0, \beta_1, ..., \beta_k) = \sum_{i=1}^{n} \varepsilon_i^2$$

$$= \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij})^2 ----(3)$$

The function S to have minimum $\frac{\partial S}{\partial \beta_0} = 0, \frac{\partial S}{\partial \beta_1} = 0, ..., \frac{\partial S}{\partial \beta_k} = 0 ----(4)$

Solving (4) for $\beta_0, \beta_1, ..., \beta_k$ we get $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ respectively.

The least squares normal equations are given by

$$\left.\begin{aligned}
&n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} = \sum_{i=1}^{n} y_i \\
&\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1}x_{i2} + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_{i1}x_{ik} = \sum_{i=1}^{n} x_{i1}y_i \\
&. \\
&. \\
&. \\
&\hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} x_{ik}x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{ik}x_{i2} + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik}^2 = \sum_{i=1}^{n} x_{ik}y_i
\end{aligned}\right\} \quad \ldots\ldots\ldots\ldots(5)$$

Problem: Fit the linear regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ for the following data:

y	41	49	69	65	40	50	58	57	31	36	44	57	19	31	33	43
x_1	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
x_2	5	5	5	5	10	10	10	10	15	15	15	15	20	20	20	20

Find y when $x_1 = 2.5$ and $x_2 = 12$

Solution:

<u>Table</u>

y	**x_1**	**x_2**	**x_1^2**	**x_1x_2**	**x_2^2**	**x_1y**	**x_2y**
41	1	5	1	5	25	41	205
49	2	5	4	10	25	98	245
69	3	5	9	15	25	207	345
65	4	5	16	20	25	260	325
40	1	10	1	10	100	40	400
50	2	10	4	20	100	100	500
58	3	10	9	30	100	174	580
57	4	10	16	40	100	228	570
31	1	15	1	15	225	31	465
36	2	15	4	30	225	72	540
44	3	15	9	45	225	132	660
57	4	15	16	60	225	228	855
19	1	20	1	20	400	19	380
31	2	20	4	40	400	62	620
33	3	20	9	60	400	99	660
43	4	20	16	80	400	172	860
723	40	200	120	500	3000	1963	8210

From the above table n = 16, $\sum x_1$ = 40, $\sum x_2$ = 200, $\sum x_1^2$ = 120, $\sum x_1x_2$ = 500

$\sum x_2^2$ = 3000, $\sum y$ = 723, $\sum x_1y$ = 1963 and $\sum x_2y$ = 8210

Normal equations to fit the model are

$$n\hat{\beta}_0 + \hat{\beta}_1\sum x_1 + \hat{\beta}_2\sum x_2 = \sum y$$
$$\hat{\beta}_0\sum x_1 + \hat{\beta}_1\sum x_1^2 + \hat{\beta}_2\sum x_1x_2 = \sum x_1y$$
$$\hat{\beta}_0\sum x_2 + \hat{\beta}_1\sum x_1x_2 + \hat{\beta}_2\sum x_2^2 = \sum x_2y$$

Substituting the above values in the normal equations we get

$$16\hat{\beta}_0 + 40\hat{\beta}_1 + 200\hat{\beta}_2 = 723$$
$$40\hat{\beta}_0 + 120\hat{\beta}_1 + 500\hat{\beta}_2 = 1963$$
$$200\hat{\beta}_0 + 500\hat{\beta}_1 + 3000\hat{\beta}_2 = 8210$$

Solving the above system of equations by known method we get

$$\hat{\beta}_0 = \frac{743}{16}, \hat{\beta}_1 = \frac{311}{40}, \hat{\beta}_2 = -\frac{331}{200}$$

Therefore $\hat{y} = 46.43 + 7.775x_1 - 1.655x_2$. So when x_1 = 2.5 and x_2 = 12 the value of y is given by $\hat{y} = 46$.

<u>Bivariate Data:</u>

In a Bivariate data, two variables are observed. One variable is independent and the other is dependent. These variables are usually denoted by X and Y. So, here we analyze the changes occurred between the two variables. So, Bivariate analysis is the analysis of exactly two variables. Multivariate analysis is the analysis of more than two variables.

Ex: Income – Expenditure

In this example, Income is the independent variable and Expenditure is the dependent variable. The Expenditure is determined by Income. Having more Income increases the Expenditure, but increasing Expenditure will not increase the Income. Such type of variables are called mutually dependent variables.

Correlation :

The Correlation is a statistical technique which studies the relationship between the two variables. In other words, Correlation is a statistical technique that is used to measure and describe the strength and direction of the relationship between two variables. It is defined as when the changes in the values of one variable are associated with the changes in the values of the other variable is called correlation.

Types of Correlation:

(a) **Positive Correlation:** If the values of two variables deviate in the same direction i.e., if the increase(decrease) in the values of one variable results in a corresponding increase(decrease) in the values of the other variable, such type of Correlation is said to be positive correlation.
Ex: Price and Supply of a commodity, Rainfall and Yield of crop, Heights and weights etc.

(b) **Negative Correlation:** Correlation is said to be Negative if the values deviate in the opposite direction i.e., if the increase(decrease) in the values of one variable results in a corresponding decrease(increase) in the values of the other variable.
Ex: Price and Demand of a commodity, Volume and Pressure, Sale of woolen garments and the Day temperature etc.

(c) **Linear and Non-linear Correlation:** The Correlation between two variables is said to be linear if corresponding to a unit change in one variable, there is a constant change in the other variable.

Ex: X :	1	2	3	4	5
Y :	5	7	9	11	13

The relationship between two variables in said to be Non-linear if corresponding to a unit change in one variable, the other variable does not change at a constant rate but at fluctuating rate.

(d) If the change in one variable does not affect the other variable, and the two variables are said to be **"Uncorrelated".**
Ex: Rainfall and Intelligence.

Scatter diagram: It is a diagram which is obtained by plotting the paired observations of mutually dependent variables. With the help of the Scatter diagram we are able to identify the type of correlation that exists between two mutually dependent variables.

Correlation Coefficient :

The quantitative measure for Correlation is called **"Correlation Coefficient"**. Otherwise, it is a measure of intensity or degree of linear relationship between two variables. It was proposed by Karl Pearson.

Correlation Coefficient between two variables X and Y, usually denoted by r_{XY} and is defined as

$$r_{XY} = \frac{Cov(X,Y)}{\sigma_X \, \sigma_Y}$$

$$= \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}}$$

Problem: Calculate the correlation coefficient between the Age and B.P. from the following data.

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

Solution:

Table:

x	**y**	**xy**	**x^2**	**y^2**
56	147	8232	3136	21609
42	125	5250	1764	15625
72	160	11520	5184	25600
36	118	4248	1296	13924
63	149	9387	3969	22201
47	128	6016	2209	16384
55	150	8250	3025	22500
49	145	7105	2401	21025
38	115	4370	1444	13225
42	140	5880	1764	19600
68	152	10336	4624	23104
60	155	9300	3600	24025
628	**1684**	**89894**	**34416**	**238822**

$$r_{XY} = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}}$$

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

r_{XY} = 0.8961

<u>Problem :</u> The following are the numbers of minutes it took 10 mechanics to assemble a piece of machinery in the morning, x, and late afternoon, y:

x	11.1	10.3	12.0	15.1	13.7	18.5	17.3	14.2	14.8	15.3
y	10.9	14.2	13.8	21.5	13.2	21.1	16.4	19.3	17.4	19.0

Calculate the correlation coefficient between x and y.

Solution:

Table:

x	**y**	**xy**	**x^2**	**y^2**
11.1	10.9	120.99	123.21	118.81
10.3	14.2	146.26	106.09	201.64
12	13.8	165.6	144	190.44
15.1	21.5	324.65	228.01	462.25
13.7	13.2	180.84	187.69	174.24
18.5	21.1	390.35	342.25	445.21
17.3	16.4	283.72	299.29	268.96
14.2	19.3	274.06	201.64	372.49
14.8	17.4	257.52	219.04	302.76
15.3	19	290.7	234.09	361
142.3	166.8	2434.69	2085.31	2897.8

$$r_{XY} = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}}$$

$$= \frac{10 \times 2434.69 - 142.3 \times 166.8}{\sqrt{10 \times 2085.31 - (142.3)^2}\sqrt{10 \times 2897.8 - (166.8)^2}}$$

$$= \frac{611.26}{\sqrt{603.81}\sqrt{1155.76}}$$

$= 0.7317$

Problem: From the following table calculate the coefficient of correlation by Karl Pearson's method.

X :	23	27	28	29	30	31	33	35	36	39
Y :	18	22	23	24	25	26	28	29	30	32