Hypothesis Testing - II

Overview

Overview

Overview

Overview

1. Introduction
2. Small Sample tests
   - Test of single mean
   - Test of difference of means
   - F-test (Test of equality of variance)
3. Chi-square Test
   - Test for Independence
   - Test for goodness of fit
4. Design of Experiments
   - One-way ANOVA
   - Two-way ANOVA

Small sample tests

1. If the population is normally distributed and σ is known (OR) if σ is unknown and $n \geq 30$ then we can apply Z test (standard normal distribution).

2. If the population is normally distributed, σ is unknown, and $n < 30$, then we apply t-test (Student's t distribution).

Small sample tests

1. If the population is normally distributed and σ is known (OR) if σ is unknown and $n \geq 30$ then we can apply Z test (standard normal distribution).

2. If the population is normally distributed, σ is unknown, and $n < 30$, then we apply t-test (Student's t distribution).

Student's t-distribution

The p.d.f of the t-distribution is

$$f(t) = \frac{\Gamma(\frac{(r+1)}{2})}{\sqrt{\pi r}\Gamma(\frac{r}{2})} \frac{1}{(1+\frac{t^2}{r})^{\frac{(r+1)}{2}}}$$

with r degrees of freedom (the number of independent values or quantities which can be assigned to a statistical distribution).

Test of single mean

Null Hyp $H_0 : \mu = \mu_0$

Test of single mean

Null Hyp $H_0 : \mu = \mu_0$

Test statistic :

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

follows t-distribution with $n - 1$ degrees of freedom.

Test of single mean

Here,

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

is an unbiased estimator of population standard deviation σ^2.

Test of single mean

Here,

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

is an unbiased estimator of population standard deviation σ^2.
The relation between S and s (sample standard deviation) is

$$S = s(\sqrt{\frac{n}{n-1}})$$

Test of single mean

Here,

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

is an unbiased estimator of population standard deviation σ^2.
The relation between S and s (sample standard deviation) is

$$S = s(\sqrt{\frac{n}{n-1}})$$

Standard error $= \frac{S}{\sqrt{n}}$
$1 - \alpha$ confidence limits for the mean are

$$(\bar{x} - t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}})$$

T-table

Table entry for p and C is the critical value t^* with probability p lying to its right and probability C lying between $-t^*$ and t^*.

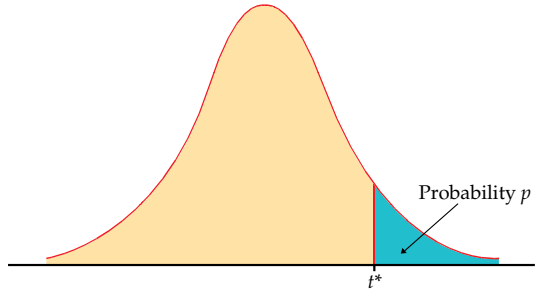

TABLE D

t distribution critical values

	Upper-tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646

Problems

1. A random sample of 16 households is taken from a large block of flats, and shows that household expenditure on food is 42 dollars per week, with a standard deviation of 10 dollars. Assuming that household expenditure on food is normally distributed, find the 95% confidence interval for the population mean.

Problems

1. A random sample of 16 households is taken from a large block of flats, and shows that household expenditure on food is 42 dollars per week, with a standard deviation of 10 dollars. Assuming that household expenditure on food is normally distributed, find the 95% confidence interval for the population mean.

Solution:

Given that $n = 16, \bar{x} = 42, s = 10, \alpha = 0.05$

Problems

1. A random sample of 16 households is taken from a large block of flats, and shows that household expenditure on food is 42 dollars per week, with a standard deviation of 10 dollars. Assuming that household expenditure on food is normally distributed, find the 95% confidence interval for the population mean.

Solution:

Given that $n = 16, \bar{x} = 42, s = 10, \alpha = 0.05$

Since $n < 30$, we use t-distribution.

$$S = s(\sqrt{\frac{n}{n-1}}) = 10(\sqrt{\frac{16}{15}}) = 10.33$$

$\frac{\alpha}{2} = 0.025$, degrees of freedom $= n - 1 = 15$.

Solution contd.

From the table for t-distribution, $t_{0.025,15} = 2.1314$.

Thus the 95% confidene interval for population mean μ is

$$(\bar{x} - t_{0.025,n-1}\frac{S}{\sqrt{n}}, \bar{x} + t_{0.025,n-1}\frac{S}{\sqrt{n}})$$

$$(42 - 2.1314(\frac{10.33}{4}), 42 + 2.1314(\frac{10.33}{4}))$$

$$(36.5, 47.5)$$

Test of single mean

2. The heights of 10 males of a given locality are found to be 70,67,62,68,61,68,70,64,64,66 inches. Is it reasonable to believe that the average height is greater than 64 inches?

Test of single mean

2. The heights of 10 males of a given locality are found to be 70,67,62,68,61,68,70,64,64,66 inches. Is it reasonable to believe that the average height is greater than 64 inches?

Solution

Given $n = 10$

Test of single mean

2. The heights of 10 males of a given locality are found to be 70,67,62,68,61,68,70,64,64,66 inches. Is it reasonable to believe that the average height is greater than 64 inches?

Solution

Given $n = 10$

$\bar{x} = 66, S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = 10$

Test of single mean

2. The heights of 10 males of a given locality are found to be 70,67,62,68,61,68,70,64,64,66 inches. Is it reasonable to believe that the average height is greater than 64 inches?

Solution

Given $n = 10$

$\bar{x} = 66, S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = 10$

1. $H_0 : \mu = 64$ against $H_1 : \mu > 64$ (Right tailed test)
2. Level of significance $\alpha = 0.05$

Test of single mean

solution contd.

3. Test statistic:
   Since population standard deviation is not known and $n < 30$, we use t-test.

$$t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}} = \frac{66 - 64}{\frac{\sqrt{10}}{\sqrt{10}}} = 2$$

   follows t-distribution with 9 degrees of freedom.

4. Rejection region:
   $\alpha = 0.05$, Critical value is $t_{\alpha,9} = 1.833$
   The critical region is $t \geq 1.833$.
   Since Cal $t > t_{\alpha,9}$, we reject H_0.

5. Conclusion:
   There is sufficient evidence to believe that the average height is greater than 64 inches.

Test of difference of means

Null Hyp $H_0 : \mu_1 - \mu_2 = d$

Test of difference of means

Null Hyp $H_0 : \mu_1 - \mu_2 = d$

Test statistic :

$$t = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

follows t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

Test of difference of means

Here,

$$S^2 = \frac{\sum(x_{1i} - \bar{x_1})^2 + \sum(x_{2i} - \bar{x_2})^2}{n_1 + n_2 - 2}$$

OR

$$S^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$$

Problems:

1. Samples of two types of electric light bulbs were tested for length of life and the following data were obtained.

Type I: $n_1 = 8, \bar{x_1} = 1234$ hours, $s_1 = 36$ hours

Type II: $n_2 = 7, \bar{x_2} = 1036$ hours, $s_2 = 40$ hours

Is the differnce in mean sufficient to warrant that Type I is superior than Type II regarding the length of life?

Solution

Given $n_1 = 8, n_2 = 7, \bar{x_1} = 1234, \bar{x_2} = 1036$.

1. $H_0 : \mu_1 = \mu_2$ aginst $H_1 : \mu_1 > \mu_2$
2. Level of significance $\alpha = 0.05$
3. Test statistic:
   Since population standard deviation is unknown and $n < 30$, we apply t-distribution with $n_1 + n_2 - 2 = 13$ degrees of freedom.

$$t = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = 9.3925$$

   follows t-distribution with 13 degrees of freedom.

Solution contd

4. Critical region:
   $t_{\alpha,13} = 1.77$ and critical region is $t > 1.77$
   Since Cal $t > 1.77$, we reject H_0.
5. Conclusion:
   There is a statistical evidence that Type I is superior to Type II.

F-distribution

F-distribution is used to test the equality of the variances of two populations from which two samples have been drawn.

F-distribution

F-distribution is used to test the equality of the variances of two populations from which two samples have been drawn.

$H_0 : \sigma_1^2 = \sigma_2^2$

Test Statistic:

$$F = \frac{S_1^2}{S_2^2}$$

where $S_1^2 = \frac{\sum (x_{1i} - \bar{x_1})^2}{n_1 - 1}$ and $S_2^2 = \frac{\sum (x_{2i} - \bar{x_2})^2}{n_2 - 1}$.
(Note: The larger among S_1^2 and S_2^2 will be the numerator)

Here F follows F-distribution with $(n_1 - 1, n_2 - 1)$ degrees of freedom.
The critical value is $F_{(n_1 - 1, n_2 - 1)}$

Test of Variances

Two samples of 6 and 7 items have the following values for a variable.
Sample 1: 39,41,42,42,44,40
Sample 2: 40,42,39,45,38,39,40
Do the sample variances differ significantly?

Test of Variances

Two samples of 6 and 7 items have the following values for a variable.
Sample 1: 39,41,42,42,44,40
Sample 2: 40,42,39,45,38,39,40
Do the sample variances differ significantly?

Solution

$n_1 = 6, n_2 = 7, \bar{x_1} = 41.33, \bar{x_2} = 40.43$

Test of Variances

Two samples of 6 and 7 items have the following values for a variable.
Sample 1: 39,41,42,42,44,40
Sample 2: 40,42,39,45,38,39,40
Do the sample variances differ significantly?

Solution

$n_1 = 6, n_2 = 7, \bar{x_1} = 41.33, \bar{x_2} = 40.43$

1. $H_0 : \sigma_1^2 = \sigma_2^2$ (There is no significant difference between the variances)
   $H_1 : \sigma_1^2 \neq \sigma_2^2$
2. Level of significance $\alpha = 0.05$
3. Test statistic:

$$F = \frac{S_1^2}{S_2^2}$$

Test of variances - problem

solution contd.

where $S_1^2 = \frac{\sum (x_{1i} - \bar{x_1})^2}{n_1 - 1} = 3.06668$ and $S_2^2 = \frac{\sum (x_{2i} - \bar{x_2})^2}{n_2 - 1} = 5.61905$.

Test of variances - problem

solution contd.

where $S_1^2 = \frac{\sum (x_{1i} - \bar{x_1})^2}{n_1 - 1} = 3.06668$ and $S_2^2 = \frac{\sum (x_{2i} - \bar{x_2})^2}{n_2 - 1} = 5.61905$.
Hence, $F = 1.8323$.

Test of variances - problem

solution contd.

where $S_1^2 = \frac{\sum (x_{1i} - \bar{x_1})^2}{n_1 - 1} = 3.06668$ and $S_2^2 = \frac{\sum (x_{2i} - \bar{x_2})^2}{n_2 - 1} = 5.61905$.
Hence, $F = 1.8323$.

From the table of F for (5,6) degrees of freedom, $F_{(6,5)} = 4.95$.

The critical region is $F > 4.95$

Test of variances - problem

solution contd.

where $S_1^2 = \frac{\sum (x_{1i} - \bar{x_1})^2}{n_1 - 1} = 3.06668$ and $S_2^2 = \frac{\sum (x_{2i} - \bar{x_2})^2}{n_2 - 1} = 5.61905$.

Hence, $F = 1.8323$.

From the table of F for (5,6) degrees of freedom, $F_{(6,5)} = 4.95$.

The critical region is $F > 4.95$

Since cal $F < 4.95$, we accept H_0.

Test of variances - problem

solution contd.

where $S_1^2 = \frac{\sum (x_{1i} - \bar{x_1})^2}{n_1 - 1} = 3.06668$ and $S_2^2 = \frac{\sum (x_{2i} - \bar{x_2})^2}{n_2 - 1} = 5.61905$.
Hence, $F = 1.8323$.
From the table of F for (5,6) degrees of freedom, $F_{(6,5)} = 4.95$.
The critical region is $F > 4.95$
Since cal $F < 4.95$, we accept H_0.

5. Conclusion:
   There is no significant difference between the population variances.

Try these problems

1. The price of a popular tennis racket at a national chain store is 179 dollars. Ria bought five of the same racket at an online auction site for the following prices: 155, 179,175,175,161. Assuming that the auction prices of rackets are normally distributed, determine whether there is sufficient evidence in the sample, at the 5% level of significance, to conclude that the average price of the racket is less than 179 dollars if purchased at an online auction.
(Hint: $n = 5$. Calculate $\bar{x}$ and S from the data. $H_0 : \mu = 179$ against $H_1 : \mu < 179$. Apply t-test.)

Problems

2. Find the rejection region for each hypothesis test based on the information given. The population is normally distributed.
(a) $H_0 : \mu = 27$ Against $H_1 : \mu < 27$, $\alpha = 0.05, n = 12, \sigma = 2.2$ (Hint: Since σ is given and population is normally distributed, we apply Z-test.)

(b) $H_0 : \mu = 52$ Against $H_1 : \mu \neq 52$,$\alpha = 0.05, n = 6, \sigma$ is unknown. (Hint: Since σ is unknown and $n < 30$, we apply t-test)

(c) $H_0 : \mu = -105$ Against $H_1 : \mu > -105, \alpha = 0.10, n = 24, \sigma$ is unknown. (Hint: Since σ is unknown and $n < 30$, we apply t-test)

Try these

3. An economist wishes to determine whether people are driving less than in the past. In one region of the country, the number of miles driven per household per year in the past was 18.59 thousand miles. A sample of 15 households produced a sample mean of 16.23 thousand miles for the last year, with sample standard deviation 4.06 thousand miles. Assuming a normal distribution of household driving distances per year, perform the relevant test at the 5% level of significance.

4. Two random samples gave the following results:
Sample 1: $n_1 = 10, \bar{x_1} = 15, \sum(x_{1i} - \bar{x_1})^2 = 90$ (sum of squared deviations from the mean)
Sample 2: $n_2 = 12, \bar{x_2} = 14, \sum(x_{2i} - \bar{x_2})^2 = 108$
Test whether the samples come from the same normal population at 5% significance level. (Hint: We need to test for (i) population mean and (ii) population variance. $H_0 : \sigma_1^2 = \sigma_2^2$ and $\mu_1 = \mu_2$. First apply F-test and then apply t-test)

Chi-square distribution

The sum of k independent squared standard normal variables is a Chi-square random variable with k degrees of freedom. That is,
$\chi^2 = Z_1^2 + Z_2^2 + \ldots + Z_k^2$

Chi-square distribution

The sum of k independent squared standard normal variables is a Chi-square random variable with k degrees of freedom. That is,
$\chi^2 = Z_1^2 + Z_2^2 + \ldots + Z_k^2$

- The curve is non symmetrical and skewed to the right
- The curve differs for each degrees of freedom

Chi-square distribution

The sum of k independent squared standard normal variables is a Chi-square random variable with k degrees of freedom. That is,
$\chi^2 = Z_1^2 + Z_2^2 + \ldots + Z_k^2$

- The curve is non symmetrical and skewed to the right
- The curve differs for each degrees of freedom

Applications

- Goodness of fit
- Test for independence

Test for independence

Procedure

1. Formulate the null and alternate hypothesis:

   H_0: Two variables are independent

   H_1: Two variables are not independent

Test for independence

Procedure

1. Formulate the null and alternate hypothesis:
   H_0: Two variables are independent
   H_1: Two variables are not independent
2. Calculate the Expected frequencies
   E = (row total)(column total) / sample size
   Note that each expected value must be greater than or equal to 5 for the chi square test to be valid

Test for independence

Procedure

1. Formulate the null and alternate hypothesis:
   H_0: Two variables are independent
   H_1: Two variables are not independent
2. Calculate the Expected frequencies
   E = (row total)(column total) / sample size
   Note that each expected value must be greater than or equal to 5 for the chi square test to be valid
3. Calculate the test statistic:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Test for independence

4 Find the degrees of freedom

$$df = (r - 1)(c - 1)$$

where r is the number of rows and c is the number of columns

Test for independence

4 Find the degrees of freedom

$$df = (r-1)(c-1)$$

where r is the number of rows and c is the number of columns

5 Calculate the critical value (cv) at the given LOS

Test for independence

4. Find the degrees of freedom

$$df = (r-1)(c-1)$$

where r is the number of rows and c is the number of columns

5. Calculate the critical value (cv) at the given LOS
6. Conclusion: If $\chi^2 < cv$, then accept H_0 (Variables are independent)
else reject H_0 (Variables are not independent)

Problems

1. The side effects of a new drug are being tested against a placebo. A simple random sample of 565 patients yields the results below. At a significance level $\alpha = 0.10$, is there enough evidence to conclude that the treatment is independent of the side effect of nausea?

Result	Drug	Placebo	Total
Nausea	36	13	49
No Nausea	254	262	516
Total	290	275	565

Problems

1. The side effects of a new drug are being tested against a placebo. A simple random sample of 565 patients yields the results below. At a significance level $\alpha = 0.10$, is there enough evidence to conclude that the treatment is independent of the side effect of nausea?

Result	Drug	Placebo	Total
Nausea	36	13	49
No Nausea	254	262	516
Total	290	275	565

Solution

H_0: The treatment and the response are independent.

H_1: The treatment and the response are dependent.

Problems

1. The side effects of a new drug are being tested against a placebo. A simple random sample of 565 patients yields the results below. At a significance level $\alpha = 0.10$, is there enough evidence to conclude that the treatment is independent of the side effect of nausea?

Result	Drug	Placebo	Total
Nausea	36	13	49
No Nausea	254	262	516
Total	290	275	565

Solution

H_0: The treatment and the response are independent.

H_1: The treatment and the response are dependent.

$\alpha = 0.10$

solution contd.

Expected frequencies

Result	Drug	Placebo	Total
Nausea	25.15	23.85	49
No Nausea	264.85	251.15	516
Total	290	275	565

solution contd.

Expected frequencies

Result	Drug	Placebo	Total
Nausea	25.15	23.85	49
No Nausea	264.85	251.15	516
Total	290	275	565

degrees of freedom $df = (2-1)(2-1) = 1$

solution contd.

Expected frequencies

Result	Drug	Placebo	Total
Nausea	25.15	23.85	49
No Nausea	264.85	251.15	516
Total	290	275	565

degrees of freedom $df = (2-1)(2-1) = 1$

Test statistic: $\chi^2 = 10.53$

solution contd.

Expected frequencies

Result	Drug	Placebo	Total
Nausea	25.15	23.85	49
No Nausea	264.85	251.15	516
Total	290	275	565

degrees of freedom $df = (2-1)(2-1) = 1$

Test statistic: $\chi^2 = 10.53$

Critical value is 2.71 ($df = 1$, $\alpha = 0.10$)

solution contd.

Expected frequencies

Result	Drug	Placebo	Total
Nausea	25.15	23.85	49
No Nausea	264.85	251.15	516
Total	290	275	565

degrees of freedom $df = (2-1)(2-1) = 1$
Test statistic: $\chi^2 = 10.53$
Critical value is 2.71 ($df = 1$, $\alpha = 0.10$)
Since $\chi^2 > 2.71$, there is enough evidence to reject H_0. Hence, there is a relation between the treatment and response.

Problems

2. Suppose the undergraduate degrees are BA, BE, BBA, and several others. There are three possible majors for the MBA students which are accounting, finance, and marketing. Can the statistician conclude that the undergraduate degree affects the choice of major from the given table?

UG/ MBA	Accounting	Finance	Marketing	Total
BA	31	13	16	60
BE	8	16	7	31
BBA	12	10	17	39
Other	10	5	7	22
Total	61	44	47	152

Solution

1. H_0: The undergraduate degree and MBA major are independent
   H_1: The undergraduate degree and MBA major are dependent
2. Expected frequencies:

UG/MBA	Accounting	Finance	Marketing	Total
BA	24.08	17.37	18.55	60
BE	12.44	8.97	9.59	31
BBA	15.65	11.29	12.06	39
Other	8.83	6.37	6.8	22
Total	61	44	47	152

Solution contd.

3 Test statistic:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$\chi^2 = \frac{(31-24.08)^2}{24.08} + \frac{(13-17.37)^2}{17.37} + \ldots + \frac{(7-6.8)^2}{6.8} = 14.7$$

Solution contd.

3 Test statistic:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$\chi^2 = \frac{(31-24.08)^2}{24.08} + \frac{(13-17.37)^2}{17.37} + \ldots + \frac{(7-6.8)^2}{6.8} = 14.7$$

4 degrees of freedom $df = (r-1)(c-1) = (4-1)(3-1) = 6$

Solution contd.

3 Test statistic:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$\chi^2 = \frac{(31-24.08)^2}{24.08} + \frac{(13-17.37)^2}{17.37} + \ldots + \frac{(7-6.8)^2}{6.8} = 14.7$$

4 degrees of freedom $df = (r-1)(c-1) = (4-1)(3-1) = 6$

5 $\alpha = 0.05$, Critical value $\chi^2_{(0.05,6)} = 12.59$

Solution contd.

3 Test statistic:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$\chi^2 = \frac{(31-24.08)^2}{24.08} + \frac{(13-17.37)^2}{17.37} + \ldots + \frac{(7-6.8)^2}{6.8} = 14.7$$

4 degrees of freedom $df = (r-1)(c-1) = (4-1)(3-1) = 6$

5 $\alpha = 0.05$, Critical value $\chi^2_{(0.05,6)} = 12.59$
The critical region is $\chi^2 > 12.59$.
Since calculated $\chi^2 > 12.59$, we reject H_0.

6 Conclusion: There is sufficient evidence to the claim that the undergraduate degree and the MBA major are related.

Problems

3. The operations manager of a company that manufactures tyres wants to determine whether there are any differences in the quality of workmanship among the three daily shifts. She randomly selects 496 tyres and carefully inspects them. Each tyre is either classified as perfect, satisfactory, or defective, and the shift that produced it is also recorded. The two categorical variables of interest are : shift and condition of the tyre produced. Do these data provide sufficient evidence at 5% significance level to infer that there are differences in quality among the three shifts?

	Perfect	Satisfactory	Defective	Total
Shift 1	106	114	11	231
Shift 2	67	70	16	153
Shift 3	37	65	10	112
Total	210	249	37	496

Problems

4. Various countries are compared using two variables- composition of economy and growth band as shown in the table.

	High growth	Medium growth	Low growth
Predominant agriculture	20	25	5
Predominant manufacturing	40	5	6
Predominant services	5	55	20

Test whether the predominant function in an economy has an impact on the growth of the economy.

Goodness of fit

Chi-square test enables us to check whether the given data fits the theoretical distributions such as Binomial, Poisson, Normal, etc.

Goodness of fit

Chi-square test enables us to check whether the given data fits the theoretical distributions such as Binomial, Poisson, Normal, etc.

- Formulate null and alternate hypothesis
- Calculate the expected frequencies
- Level of significance
- Test statistic:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

- degrees of freedom $df = k - 1$ where k represents the number of categories
- Calculate the critical value (cv) at the given LOS

Goodness of fit

Chi-square test enables us to check whether the given data fits the theoretical distributions such as Binomial, Poisson, Normal, etc.

- Formulate null and alternate hypothesis
- Calculate the expected frequencies
- Level of significance
- Test statistic:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

- degrees of freedom $df = k - 1$ where k represents the number of categories
- Calculate the critical value (cv) at the given LOS
- Conclusion: If $\chi^2 < cv$, then accept H_0
  else reject H_0

Problems

1. The number of defects per unit in a sample of 330 units of a manufactured product was found as follows:

No. of defects:	0	1	2	3	4
No. of units:	214	92	20	3	1

Fit a Poisson distribution to the data and test for goodness of fit.

Problems

1. The number of defects per unit in a sample of 330 units of a manufactured product was found as follows:

No. of defects:	0	1	2	3	4
No. of units:	214	92	20	3	1

Fit a Poisson distribution to the data and test for goodness of fit.

Solution

H_0: The data fits the Poisson distribution

H_1: The data doesnot fit the Poisson distribution

Problems

1. The number of defects per unit in a sample of 330 units of a manufactured product was found as follows:

No. of defects:	0	1	2	3	4
No. of units:	214	92	20	3	1

Fit a Poisson distribution to the data and test for goodness of fit.

Solution

H_0: The data fits the Poisson distribution

H_1: The data doesnot fit the Poisson distribution

Mean $= \lambda = \frac{\sum fx}{\sum f} = \frac{145}{330} = 0.439$

Expected frequencies:

$P(X = 0) = \frac{e^{-\lambda}\lambda^0}{0!} = 0.645$

Similarly calculate $P(X = 1), P(X = 2), P(X = 3), P(X = 4)$

Solution contd.

X	0	1	2	3	4
O	214	92	20	3	1
E	212.75	93.4	20.5	3	0.35

Test statistic:

$$\chi^2 = \frac{(O-E)^2}{E} = 0.0292$$

Solution contd.

X	0	1	2	3	4
O	214	92	20	3	1
E	212.75	93.4	20.5	3	0.35

Test statistic:

$$\chi^2 = \frac{(O-E)^2}{E} = 0.0292$$

Degrees of freedom $df = k - 1 = (5 - 3) - 1 = 1$

Solution contd.

X	0	1	2	3	4
O	214	92	20	3	1
E	212.75	93.4	20.5	3	0.35

Test statistic:

$$\chi^2 = \frac{(O-E)^2}{E} = 0.0292$$

Degrees of freedom $df = k - 1 = (5 - 3) - 1 = 1$

Critical value is $\chi^2_{0.05,1} = 3.84$

Solution contd.

X	0	1	2	3	4
O	214	92	20	3	1
E	212.75	93.4	20.5	3	0.35

Test statistic:

$$\chi^2 = \frac{(O-E)^2}{E} = 0.0292$$

Degrees of freedom $df = k - 1 = (5 - 3) - 1 = 1$
Critical value is $\chi^2_{0.05,1} = 3.84$
Critical region is $\chi^2 > 3.84$.

Solution contd.

X	0	1	2	3	4
O	214	92	20	3	1
E	212.75	93.4	20.5	3	0.35

Test statistic:

$$\chi^2 = \frac{(O-E)^2}{E} = 0.0292$$

Degrees of freedom $df = k - 1 = (5 - 3) - 1 = 1$
Critical value is $\chi^2_{0.05,1} = 3.84$
Critical region is $\chi^2 > 3.84$.
Since calculated $\chi^2 < 3.84$, we accept H_0. Hence the given data fits well with Poisson distribution.

Problems

1. An experiment is conducted in which a die is rolled 240 times. The outcomes are in the table below. At a significance level $\alpha = 0.05$, is there enough evidence to support the hypothesis that the die is unbiased?

Outcome	1	2	3	4	5	6
Frequency	34	44	30	46	51	35

Problems

1. An experiment is conducted in which a die is rolled 240 times. The outcomes are in the table below. At a significance level $\alpha = 0.05$, is there enough evidence to support the hypothesis that the die is unbiased?

Outcome	1	2	3	4	5	6
Frequency	34	44	30	46	51	35

Solution

H_0: The die is unbiased

H_1: The die is biased

Solution contd.

Expected frequencies: $E = \frac{240}{6} = 40$ where Total frequency is 240.

Solution contd.

Expected frequencies: $E = \frac{240}{6} = 40$ where Total frequency is 240.

Outcome	1	2	3	4	5	6
Frequency	40	40	40	40	40	40

Solution contd.

Expected frequencies: $E = \frac{240}{6} = 40$ where Total frequency is 240.

Outcome	1	2	3	4	5	6
Frequency	40	40	40	40	40	40

Outcome	O	E	$(O-E)$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
1	34	40	-6	36	0.9
2	44	40	4	16	0.4
3	30	40	-10	100	2.5
4	46	40	6	36	0.9
5	51	40	11	121	3.025
6	35	40	-5	25	0.625

Solution contd.

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 8.35$$

Solution contd.

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 8.35$$

Degrees of freedom $df = k - 1 = 6 - 1 = 5$

Solution contd.

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 8.35$$

Degrees of freedom $df = k - 1 = 6 - 1 = 5$

Critical value at $\alpha = 0.05$ with 5 degrees of freedom is 11.07

Solution contd.

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 8.35$$

Degrees of freedom $df = k - 1 = 6 - 1 = 5$

Critical value at $\alpha = 0.05$ with 5 degrees of freedom is 11.07

Since Calculated $\chi^2 < 11.07$, we accept H_0.

Conclusion: There is no sufficient evidence that the die is biased OR The die is unbiased.

Problems

2. A sample analysis of examination results of 500 students was made. It was found that 220 students had failed, 170 had secured a third class, 90 were placed in the second class and 20 got a first class. Are these figures commensurate with the general examination result which is the ratio $4 : 3 : 2 : 1$ for various categories respectively?

Problems

2. A sample analysis of examination results of 500 students was made. It was found that 220 students had failed, 170 had secured a third class, 90 were placed in the second class and 20 got a first class. Are these figures commensurate with the general examination result which is the ratio $4 : 3 : 2 : 1$ for various categories respectively? (Expected frequencies are 200,150,100,50. $\chi^2 = 23.667$, critical value is 7.81, reject H_0.)

Problems

2. A sample analysis of examination results of 500 students was made. It was found that 220 students had failed, 170 had secured a third class, 90 were placed in the second class and 20 got a first class. Are these figures commensurate with the general examination result which is the ratio $4:3:2:1$ for various categories respectively? (Expected frequencies are 200,150,100,50. $\chi^2 = 23.667$, critical value is 7.81, reject H_0.)

3. The demand for a particular spare part in a factory was found to vary from day to day. In a sample study the following information was obtained:

Days	Mon	Tue	Wed	Thurs	Fri	Sat
Demand	1124	1125	1110	1120	1126	1115

Test the hypothesis that the number of parts demanded does not depend on the day of the week.

Problems

2. A sample analysis of examination results of 500 students was made. It was found that 220 students had failed, 170 had secured a third class, 90 were placed in the second class and 20 got a first class. Are these figures commensurate with the general examination result which is the ratio $4:3:2:1$ for various categories respectively? (Expected frequencies are 200,150,100,50. $\chi^2 = 23.667$, critical value is 7.81, reject H_0.)

3. The demand for a particular spare part in a factory was found to vary from day to day. In a sample study the following information was obtained:

Days	Mon	Tue	Wed	Thurs	Fri	Sat
Demand	1124	1125	1110	1120	1126	1115

Test the hypothesis that the number of parts demanded does not depend on the day of the week.

(Expected frequencies are 1120. $\chi^2 = 0.179$, critical value is 11.07, accept H_0.)

ANOVA - Analysis of Variance

ANOVA - Analysis of Variance

Why ANOVA?

- When comparing means across two samples - we use Z-test or t-test

ANOVA - Analysis of Variance

Why ANOVA?

- When comparing means across two samples - we use Z-test or t-test
- If more than two samples are test for their means, we use ANOVA

Design of Experiments

ANOVA - Analysis of Variance

Why ANOVA?

- When comparing means across two samples - we use Z-test or t-test
- If more than two samples are test for their means, we use ANOVA

We study,

- Completely Randomised Design or One-way ANOVA
- Randomized Block Design or Two-way ANOVA
- Latin Square Design or Three-way ANOVA

CSD or One-way ANOVA

The observations are independent.

$H_0 : \mu_1 = \mu_2 = \mu_3$

H_1 : At least there is one difference among the means.

$$F = \frac{\text{Between group variance}}{\text{Within group variance}}$$

with degrees of freedom $(k-1, N-k)$ where k denotes the number of groups and N denotes the sample size.

CSD or One-way ANOVA

Problem 1: Compare the means of these groups

I	II	III
1	2	2
2	4	3
5	2	4

CSD or One-way ANOVA

Problem 1: Compare the means of these groups

I	II	III
1	2	2
2	4	3
5	2	4

Solution:

1: $H_0 : \mu_1 = \mu_2 = \mu_3$

H_1 : At least there is one difference among the means.

$\alpha = 0.05$

CSD or One-way ANOVA

Problem 1: Compare the means of these groups

I	II	III
1	2	2
2	4	3
5	2	4

Solution:

1: $H_0 : \mu_1 = \mu_2 = \mu_3$
H_1 : At least there is one difference among the means.
$\alpha = 0.05$

2: Degrees of freedom
$DF_{Bet} = k - 1 = 3 - 1 = 2,$

CSD or One-way ANOVA

Problem 1: Compare the means of these groups

I	II	III
1	2	2
2	4	3
5	2	4

Solution:

1: $H_0 : \mu_1 = \mu_2 = \mu_3$
H_1 : At least there is one difference among the means.
$\alpha = 0.05$

2: Degrees of freedom
$DF_{Bet} = k - 1 = 3 - 1 = 2, \qquad DF_{Within} = N - k = 9 - 3 = 6$

CSD or One-way ANOVA

Problem 1: Compare the means of these groups

I	II	III
1	2	2
2	4	3
5	2	4

Solution:

1: $H_0 : \mu_1 = \mu_2 = \mu_3$
H_1 : At least there is one difference among the means.
$\alpha = 0.05$

2: Degrees of freedom
$DF_{Bet} = k - 1 = 3 - 1 = 2, \qquad DF_{Within} = N - k = 9 - 3 = 6$
$F_{(2,6)} = 5.14$

3.

$$G = \sum\sum x_{ij} = 25$$

3.

$$G = \sum\sum x_{ij} = 25$$

Correction Factor $C.F. = \frac{G^2}{N} = 69.444$

Sum of squares total

$$SST = \sum\sum x_{ij}^2 - \frac{G^2}{N} = 83 - 69.444 = 13.556$$

3.

$$G = \sum\sum x_{ij} = 25$$

Correction Factor $C.F. = \frac{G^2}{N} = 69.444$

Sum of squares total

$$SST = \sum\sum x_{ij}^2 - \frac{G^2}{N} = 83 - 69.444 = 13.556$$

Sum of squares between

$$SSB = \sum \frac{T_i^2}{r_i} - \frac{G^2}{N} = \frac{8^2}{3} + \frac{8^2}{3} + \frac{9^2}{3} - 69.444$$
$$SSB = 69.667 - 69.444 = 0.223$$

3.

$$G = \sum\sum x_{ij} = 25$$

Correction Factor $C.F. = \frac{G^2}{N} = 69.444$

Sum of squares total

$$SST = \sum\sum x_{ij}^2 - \frac{G^2}{N} = 83 - 69.444 = 13.556$$

Sum of squares between

$$SSB = \sum \frac{T_i^2}{r_i} - \frac{G^2}{N} = \frac{8^2}{3} + \frac{8^2}{3} + \frac{9^2}{3} - 69.444$$
$$SSB = 69.667 - 69.444 = 0.223$$

Sum of squares within

$$SSW = SST - SSB = 13.556 - 0.223 = 13.333$$

4. Mean sum of squares

$$MSB = \frac{SSB}{DF_{Between}} = \frac{.223}{2} = 0.115$$

4. Mean sum of squares

$$MSB = \frac{SSB}{DF_{Between}} = \frac{.223}{2} = 0.115$$

$$MSW = \frac{SSW}{DF_{Within}} = \frac{13.333}{6} = 2.222$$

4. Mean sum of squares

$$MSB = \frac{SSB}{DF_{Between}} = \frac{.223}{2} = 0.115$$

$$MSW = \frac{SSW}{DF_{Within}} = \frac{13.333}{6} = 2.222$$

5. ANOVA table:

Source	SS	DF	MS	F
Between groups	0.223	2	0.115	$\frac{MSB}{MSW} = 0.0517$
Within groups	13.333	6	2.222	

4. Mean sum of squares

$$MSB = \frac{SSB}{DF_{Between}} = \frac{.223}{2} = 0.115$$

$$MSW = \frac{SSW}{DF_{Within}} = \frac{13.333}{6} = 2.222$$

5. ANOVA table:

Source	SS	DF	MS	F
Between groups	0.223	2	0.115	$\frac{MSB}{MSW} = 0.0517$
Within groups	13.333	6	2.222	

Since Cal $F < F_{Critical} = 5.14$, we accept H_0.
There is no significant difference between the means of the group.

Problems

2. A random sample is selected from each of three makes of ropes and their breaking strength (in pounds) are measured with the following results:

Group A	70	72	75	80	83		
Group B	100	110	108	112	113	120	107
Group C	60	65	57	84	87	73	

Test whether the breaking strength of the ropes differs significantly.

Problems

2. A random sample is selected from each of three makes of ropes and their breaking strength (in pounds) are measured with the following results:

Group A	70	72	75	80	83		
Group B	100	110	108	112	113	120	107
Group C	60	65	57	84	87	73	

Test whether the breaking strength of the ropes differs significantly. $SST = 6964.44, SSB = 5838.44, SSW = 1126, F = 38.89, F_{(2,15)} = 3.68$, Reject H_0

Problems

3. An experiment with 10 plots and 3 treatments gave the following results:

Plot no.	1	2	3	4	5	6	7	8	9	10
Treatment	A	B	C	A	C	C	A	B	A	B
Yield	5	4	3	7	5	1	3	4	1	7

Test whether the treatments differs significantly.

Problems

3. An experiment with 10 plots and 3 treatments gave the following results:

Plot no.	1	2	3	4	5	6	7	8	9	10
Treatment	A	B	C	A	C	C	A	B	A	B
Yield	5	4	3	7	5	1	3	4	1	7

Test whether the treatments differs significantly.

Solution

A	5	7	1	3
B	4	4	7	
C	3	1	5	

Problems

3. An experiment with 10 plots and 3 treatments gave the following results:

Plot no.	1	2	3	4	5	6	7	8	9	10
Treatment	A	B	C	A	C	C	A	B	A	B
Yield	5	4	3	7	5	1	3	4	1	7

Test whether the treatments differs significantly.

Solution

A	5	7	1	3
B	4	4	7	
C	3	1	5	

$SST = 40, SSB = 6, SSW = 34, F = 1.619, F_{(2,7)} = 4.74$, Accept H_0

RBD or Two-Way ANOVA (without replications

- In a One-way ANOVA, we select the random sample for each group or column
- A Two-way ANOVA allows us to "account for variation" at the ROW level due to some other factor or grouping
- By adding blocks or factors to the ROWS, we can reduce the overall ERROR or WITHIN variance
- Now we have 4 types of Sum of Squares or Sources of Variation: (i) TOTAL (ii) COLUMNS or GROUPS (iii) ROWS or BLOCKS (iv) ERROR or WITHIN
- Note that SST=SSC+SSR+SSE

RBD or Two-way ANOVA without replication-Example

	Sydney	Brisbane	Melbourne
Shopper 1	75	75	90
Shopper 2	70	70	70
Shopper 3	50	55	75
Shopper 4	65	60	85
Shopper 5	80	65	80
Shopper 6	65	65	65
	$\dot{x}_{C1} = 67.5$	$\dot{x}_{C2} = 65$	$\dot{x}_{C3} = 77.5$

CITY VARIAT

Overall Mean

The mean of al scores taken together.

$$\bar{\bar{x}} = 70$$

Step 1:

Null hypothesis: There is no significant difference in the means of Columns (Groups) as well as Rows (Blocks). That is,

$H_{01} : \mu_1 = \mu_2 = \ldots = \mu_c$ (Columns)

$H_{02} : \mu_1 = \mu_2 = \ldots = \mu_r$ (Rows)

Alternate Hypothesis: There is at least one mean in the Columns which differs from others. Also there is at least one mean in the Rows which differs from the others.

Step 1:

Null hypothesis: There is no significant difference in the means of Columns (Groups) as well as Rows (Blocks). That is,

$H_{01} : \mu_1 = \mu_2 = \ldots = \mu_c$ (Columns)

$H_{02} : \mu_1 = \mu_2 = \ldots = \mu_r$ (Rows)

Alternate Hypothesis: There is at least one mean in the Columns which differs from others. Also there is at least one mean in the Rows which differs from the others.

Step 2:

Degrees of Freedom: $DF_{Columns} = c - 1$, $DF_{Rows} = r - 1$, $DF_{Error} = (c - 1)(r - 1)$.

Step 1:

Null hypothesis: There is no significant difference in the means of Columns (Groups) as well as Rows (Blocks). That is,

$H_{01} : \mu_1 = \mu_2 = \ldots = \mu_c$ (Columns)

$H_{02} : \mu_1 = \mu_2 = \ldots = \mu_r$ (Rows)

Alternate Hypothesis: There is at least one mean in the Columns which differs from others. Also there is at least one mean in the Rows which differs from the others.

Step 2:

Degrees of Freedom: $DF_{Columns} = c - 1$, $DF_{Rows} = r - 1$, $DF_{Error} = (c-1)(r-1)$.

Compute the Critical values $F_{(c-1,(c-1)(r-1))}$ and $F_{(r-1,(c-1)(r-1))}$.

Step 3: $G = \sum\sum x_{ij}$;

Step 3: $G = \sum\sum x_{ij}$; Correction Factor $C.F. = \frac{G^2}{N}$

Sum of squares total

$SST = \sum\sum x_{ij}^2 - C.F.$

Step 3: $G = \sum\sum x_{ij}$; Correction Factor $C.F. = \frac{G^2}{N}$

Sum of squares total

$$SST = \sum\sum x_{ij}^2 - C.F.$$

Sum of squares

$$SSC = \sum \frac{C_j^2}{c_j} - C.F.$$

where C_j represent the column sum of jth column and c_j represent the number of observations in the jth column.

Step 3: $G = \sum\sum x_{ij}$; Correction Factor $C.F. = \frac{G^2}{N}$

Sum of squares total

$$SST = \sum\sum x_{ij}^2 - C.F.$$

Sum of squares

$$SSC = \sum \frac{C_j^2}{c_j} - C.F.$$

where C_j represent the column sum of jth column and c_j represent the number of observations in the jth column.

$$SSR = \sum \frac{R_i^2}{r_i} - C.F.$$

where R_i represent the row sum of ith row and r_i represent the number of observations in the ith row.

Step 3: $G = \sum\sum x_{ij}$; Correction Factor $C.F. = \frac{G^2}{N}$

Sum of squares total

$$SST = \sum\sum x_{ij}^2 - C.F.$$

Sum of squares

$$SSC = \sum \frac{C_j^2}{c_j} - C.F.$$

where C_j represent the column sum of jth column and c_j represent the number of observations in the jth column.

$$SSR = \sum \frac{R_i^2}{r_i} - C.F.$$

where R_i represent the row sum of ith row and r_i represent the number of observations in the ith row.

$$SSE = SST - SSR - SSC$$

Step 4:

Source	SS	DF	MS	F
Columns (Groups)	SSC	$c-1$	$\frac{SSC}{c-1}$	$F_1 = \frac{MSC}{MSE}$
Rows (Blocks) (or) rows	SSR	$r-1$	$\frac{SSR}{r-1}$	$F_2 = \frac{MSR}{MSE}$
Within or Error	SSE	$(c-1)(r-1)$	$\frac{SSE}{(c-1)(r-1)}$	

RBD or Two-way ANOVA

Problem 1: The following data represent the number of units produced per day by different workers using 4 different types of machines.

Worker / Machines	I	II	III	IV
1	44	38	47	36
2	46	40	52	43
3	34	36	44	32
4	43	38	46	33
5	38	42	49	39

1. Test whether the five men differ with respect to mean productivity
2. Test whether the mean productivity is same for the four different machine types.

Solution

1. Null hypothesis H_0: (a) The mean productivity is same for the four different machines. (b) Five workers do not differ with respect to mean productivity.
   Alternate hypothesis H_1: (a) The mean productivity differs for at least one machine. (b) The mean productivity differs for at least a worker.

Solution

1. Null hypothesis H_0: (a) The mean productivity is same for the four different machines. (b) Five workers do not differ with respect to mean productivity.
   Alternate hypothesis H_1: (a) The mean productivity differs for at least one machine. (b) The mean productivity differs for at least a worker.

2. Degrees of Freedom: $DF_{Columns} = 4 - 1 = 3$, $DF_{Rows} = 5 - 1 = 4$, $DF_{Error} = 3 \times 4 = 12$. The critical values are $F_{(3,12)} = 3.49$ and $F_{(4,12)} = 3.26$

Solution

3 Calculate the deviation for all observations with respect to some origin, say 40.

Worker / Machines	I	II	III	IV	Total
1	4	-2	7	-4	5
2	6	0	12	3	21
3	-6	-4	4	-8	-14
4	3	-2	6	-7	0
5	-2	2	9	-1	8
Total	5	-6	38	-17	20

Solution

3 Calculate the deviation for all observations with respect to some origin, say 40.

Worker / Machines	I	II	III	IV	Total
1	4	-2	7	-4	5
2	6	0	12	3	21
3	-6	-4	4	-8	-14
4	3	-2	6	-7	0
5	-2	2	9	-1	8
Total	5	-6	38	-17	20

$G = 20; \qquad N = 20; \qquad C.F. = \frac{G^2}{N} = 20$

Solution

$$SST = \sum\sum x_{ij} - C.F. = 594 - 20 = 574$$

Solution

$$SST = \sum\sum x_{ij} - C.F. = 594 - 20 = 574$$
$$SSC = \sum \frac{C_j^2}{c_j} - C.F. = 181.5 - 20 = 161.5$$

Solution

$$SST = \sum\sum x_{ij} - C.F. = 594 - 20 = 574$$
$$SSC = \sum \frac{C_j^2}{c_j} - C.F. = 181.5 - 20 = 161.5$$
$$SSR = \sum \frac{R_i^2}{r_i} - C.F. = 358.8 - 20 = 338.8$$

Solution

$$SST = \sum\sum x_{ij} - C.F. = 594 - 20 = 574$$
$$SSC = \sum \frac{C_j^2}{c_j} - C.F. = 181.5 - 20 = 161.5$$
$$SSR = \sum \frac{R_i^2}{r_i} - C.F. = 358.8 - 20 = 338.8$$
$$SSE = SST - SSC - SSR = 73.7$$

Solution

Source	SS	DF	MS	F
Columns (Groups)	338.8	3	112.93	$F_1 = \frac{MSC}{MSE} = 18.393$
Rows (Blocks) (or) rows	161.5	4	40.375	$F_2 = \frac{MSR}{MSE} = 6.576$
Within or Error	73.7	12	6.14	

Solution

Source	SS	DF	MS	F
Columns (Groups)	338.8	3	112.93	$F_1 = \frac{MSC}{MSE} = 18.393$
Rows (Blocks) (or) rows	161.5	4	40.375	$F_2 = \frac{MSR}{MSE} = 6.576$
Within or Error	73.7	12	6.14	

Since $F_1 > 3.49$, we reject the corresponding H_0. That is, the mean productivity differs for at least one worker.
Since $F_2 > 3.26$, we reject the corresponding H_0. That is, the fiver workers differ with respect to the mean productivity.

Try these

2. Four kinds of fertilizer f_1, f_2, f_3 and f_4 are used to study the yield of beans. The soil is divided into 3 blocks, each containing 4 homogeneous plots. The yields in kilograms per plot and the corresponding treatments are as follows:

Block 1	$f_1 = 42.7,$	$f_3 = 48.5,$	$f_4 = 32.8,$	$f_2 = 39.3$
Block 2	$f_3 = 50.9,$	$f_1 = 50,$	$f_2 = 38,$	$f_4 = 40.2$
Block 3	$f_4 = 51.1,$	$f_2 = 46.3,$	$f_1 = 51.9,$	$f_2 = 53.5$

Conduct an analysis of variance at the 0.05 level of significance using the randomized block model.

Try these

3. Three varieties of potatoes are being compared for yield. The experiment is conducted by assigning each variety at random to one of 3 equal size plots at each of 4 different locations. The following yields for varieties A, B, and C, in 100 kilograms per plot, were recorded:

Location 1	B:13	A:18	C:12
Location 2	C:21	A:20	B:23
Location 3	C:9	B:12	A:14
Location 4	A:11	C:10	B:17

Perform a two-way analysis of variance to test at 5% level of significance.

Latin Square Design or Three-Way ANOVA

In addition to rows and columns, we need to consider an extra factor known as Treatments. Every treatment occurs only once in each row and in each column. Such a layout is known as Latin square design. For eg. if we are interested in studying the effects of n types of fertilizers on a yield of a certain variety of wheat, we conduct the experiment on a square field with n^2 plots of equal area and associate treatments with different fertilizers; row and column effects with variations in fertility of soil.

Procedure

Step 1:

Null hypothesis: There is no significant difference in the means of Columns (Groups) , Rows (Blocks), and Treatments **Alternate Hypothesis:** There is at least one mean in the Columns which differs from others. Also there is at least one mean in the Rows which differs from the others. Similarly for Treatments

Step 2:

Degrees of Freedom: $DF_{Columns} = n - 1$, $DF_{Rows} = n - 1$, $DF_{Treatments} = n - 1$, $DF_{Error} = (n - 1)(n - 2)$.
Compute the Critical value $F_{(n-1,(n-1)(n-2))}$

Step 3: $G = \sum\sum x_{ij}$; Correction Factor $C.F. = \frac{G^2}{N}$

Sum of squares total

$$SST = \sum\sum x_{ij}^2 - C.F.$$

Sum of squares

$$SSC = \sum \frac{C_j^2}{n} - C.F.$$

where C_j represent the column sum of jth column.

$$SSR = \sum \frac{R_i^2}{n} - C.F.$$

where R_i represent the row sum of ith row.

$$SSTr = \sum \frac{T_i^2}{n} - C.F.$$

where T_i represent the Treatment sum of ith treatment.

$$SSE = SST - SSR - SSC - SSTr$$

Step 4:

Source	SS	DF	MS	F
Columns (Groups)	SSC	$n-1$	$\frac{SSC}{n-1}$	$F_1 = \frac{MSC}{MSE}$
Rows (Blocks) (or) rows	SSR	$n-1$	$\frac{SSR}{n-1}$	$F_2 = \frac{MSR}{MSE}$
Treatments	SSTr	$n-1$	$\frac{SSTr}{n-1}$	$F_3 = \frac{MSTr}{MSE}$
Within or Error	SSE	$(n-1)(n-2)$	$\frac{SSE}{(n-1)(n-2)}$	

Problems

1. Analyze the variance in the Latin square of yields (in Kgs) of paddy where P, Q, R, S denote the different methods of cultivation

S 122	P 121	R 123	Q 122
Q 124	R 123	P 122	S 125
P 120	Q 119	S 120	R 121
R 122	S 123	Q 121	P 122

Step 1:

Null hypothesis: There is no significant difference in the means of Columns (Groups) , Rows (Blocks), and Treatments (methods of cultivation)

Alternate Hypothesis: There is at least one mean in the Columns which differs from others. Also there is at least one mean in the Rows which differs from the others. Similarly for Treatments

Step 2:

$n = 4$

Degrees of Freedom: $DF_{Columns} = 3$, $DF_{Rows} = 3$, $DF_{Treatments} = 3$, $DF_{Error} = 6$.

Critical value $F_{(3,6)} = 4.76$

Step 3:

Calculate the deviation about the origin as 120

S 2	P 1	R 3	Q 2
Q 4	R 3	P 2	S 5
P 0	Q −1	S 0	R 1
R 2	S 3	Q 1	P 2

Treatment sum: $P = 5, Q = 6, R = 9, S = 10$

$G = 30; N = 16$

$\text{C.F.} = \frac{G^2}{N} = 56.25$

$SST = \sum\sum x_{ij}^2 - C.F. = 35.75$

$SSC = \sum \frac{C_j^2}{n} - C.F. = 2.75$

$SSR = \sum \frac{R_i^2}{n} - C.F. = 24.75$

$SSTr = \sum \frac{T_i^2}{n} - C.F. = \frac{5^2}{4} + \frac{6^2}{4} + \frac{9^2}{4} + \frac{10^2}{4} - 56.25 = 4.25$

$SSE = SST - SSR - SSC - SSTr = 4$

Source	SS	DF	MS	F
Columns (Groups)	2.75	3	0.917	$F_1 = \frac{0.917}{0.667} = 1.375$
Rows (Blocks) (or) rows	24.75	3	8.25	$F_2 = \frac{8.25}{0.667} = 12.36$
Methods	4.25	3	1.417	$F_3 = \frac{1.417}{0.667} = 2.124$
Within or Error	SSE	6	0.667	

Comparing F_1, F_2, F_3 with Critical F, we accept H_0 (columns), accept H_0 (Methods or treatments), Reject H_0 (Rows).

Digital Assignment - 3

1. Perform a Latin square analysis for the design

C 25	B 23	A 20	D 20
A 19	D 19	C 21	B 18
B 19	A 14	D 17	C 20
D 17	C 20	B 21	A 15

2. Three varieties of potatoes are being compared for yield. The experiment is conducted by assigning each variety at random to one of 3 equal size plots at each of 4 different locations. The following yields for varieties A, B, and C, in 100 kilograms per plot, were recorded:

Location 1	B:13	A:18	C:12
Location 2	C:21	A:20	B:23
Location 3	C:9	B:12	A:14
Location 4	A:11	C:10	B:17

Perform a three-way analysis of variance to test at 5% level of