VIT
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)
SCHOOL OF ADVANCED SCIENCES
DEPARTMENT OF MATHEMATICS
FALL SEMESTER – 2020~2021

MAT2001 – Statistics for Engineers
(Embedded Theory Component)

COURSE MATERIAL

Module 6
Hypothesis Testing – II

Syllabus:
Small Sample Tests – Student's t-Test – F-Test – Chi-Square Test – Goodness of Fit – Independence of Attributes – Design of Experiments – Analysis of Variance – One and Two Way Classifications - CRD-RBD-LSD.

Prepared By: **Prof. S. Roy (In-charge)**
Prof. V. Murugan
Prof. Padigepati Naveen

The course in-charges thankfully acknowledge the course materials preparation committee in-charge and members for their significant contribution in bringing out of this course material.

****
Dr. D. Easwaramoorthy
Dr. A. Manimaran
Course In-charges – MAT2001-SE,
Fall Semester 2020~2021,
Department of Mathematics,
SAS, VIT, Vellore.
****

Module-6: Small Sample Tests – Student's *t*-Test – *F*-Test:

Student's t-distribution:

The static '*t*' was introduced by W S.Gosset in 1908 who wrote under the name "Student". That is why it is called student's t test. Later on its distribution was rigorously established by Prof. R.A. Fisher in 1926. It can used when the population standard deviation is not known and the size of the sample is less than or equal to thirty.

A random variable X is said to follow t-distribution, if its probability density function is given by

$$f(t) = \frac{k}{\left(1+\frac{t^2}{v}\right)^{v+1/2}}, \quad -\infty < t < \infty$$

where v is known as the degrees of freedom and k is constant. The constant value k is chosen in such a way that $\int_{-\infty}^{\infty} f(t)dt = 1$. After simplification, we get $k = \frac{1}{\sqrt{v}\beta\left(\frac{1}{2},\frac{v}{2}\right)}$.

Assumption of t-distribution:

1) The population from which the sample is drawn is normal.

2) The sample is random and size $n \leq 30$.

3) The population S.D. σ is not known.

Properties of t-distribution:

1) The probability curve of t-distribution is symmetrical.

2) The tails of the curve are asymptotic to x-axis.

3) When n→ ∞, t- distribution tends to normal distribution.

4) The form of the t-dist. Varies with the degrees of freedom.

Application of t- distribution: The t- distribution is used

1) To test significance of the mean of sample.

2) To test the difference between two means or to compare two samples.

1. **To test significance of the mean of sample**:

If $x_1, x_2, \ldots, x_n$ is a random sample of '*n*' observations drawn from a normal population with mean μ and S.D. σ . To test the significance of a mean of a small sample under the null hypothesis H_0: $\mu=\mu_0$, the test statistic is given by

$$t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$$ and follows t-distribution with $v=n-1$ degree of freedom.

where $\bar{x} = \frac{\sum x_i}{n}$ be the mean of the sample then and $s^2 = \frac{\sum (x_i-\bar{x})^2}{n-1}$ be the variance of the sample

The alternative hypothesis in this case is either H_1: $\mu > \mu_0$ (right-tailed), or H_1: $\mu < \mu_0$ (left-tailed), or H_1: $\mu \neq \mu_0$ (two-tailed).
The rejection region for a level α is either $t \geq t_{\alpha,n-1}$(right-tailed), or $t \leq t_{\alpha,n-1}$(left-tailed), and or either $t \geq t_{\alpha/2,n-1}$ or $t \leq -t_{\alpha/2,n-1}$ (two-tailed).

2. **To test the difference between two means or to compare two samples.**

I. *If two small samples drawn from the same normal population:*

Let $\bar{x}_1$ and $\bar{x}_2$ be the sample means for two small samples drawn from a normal population. Under the null hypothesis H_0: $\mu_1=\mu_2$, the test statistic is given by

$$t = \frac{(\bar{x}_1-\bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$$ and follows t-distribution with $v=n_1+n_2-1$ degree of freedom.

where $\bar{x}_1 = \frac{\sum x_i}{n_1}$ and $\bar{x}_2 = \frac{\sum x_i}{n_2}$ be the means of the sample sizes n_1 and n_2. There variances are given by $s_1^2 = \frac{\sum (x_i-\bar{x}_1)^2}{n_1-1}$ and $s_2^2 = \frac{\sum (x_i-\bar{x}_2)^2}{n_2-1}$ respectively.

II. *If two small samples drawn from the normal populations having different means:*

Let $\bar{x}_1$ and $\bar{x}_2$ be the sample means for two small samples drawn from the normal population with different means μ_1 and μ_2, respectively. Under the null hypothesis H_0: $\mu_1=\mu_2$, the test statistic is given by

$$t = \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

where $S^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum(x_i - \bar{x}_1)^2 + \sum(x_j - \bar{x}_2)^2\right]$ with degrees of freedom $v=n_1+n_2-1$.

Also, $\bar{x}_1 = \frac{\sum x_i}{n_1}$ and $\bar{x}_2 = \frac{\sum x_i}{n_2}$ be the means of the sample sizes n_1 and n_2.

Problems:

1. Ten individuals are chosen at random from a population and their heights are found to be in inches 63,63,64,65,66,69,69,70,70,71 discuss the suggestion that the mean height in the universe is 65 inches given that for 9 degrees of freedom the value of *Student's t* and 5 percent level of significance is 2.262.

   Solution:

   For the calculation of sample mean and sample variance, we have taken the following into consideration

Serial no	x	$x-\bar{x}$	$(x-\bar{x})^2$
1	63	-4	16
2	63	-4	16
3	64	-3	9
4	65	-2	4
5	66	- 1	1
6	69	2	4
7	69	2	4
8	70	3	9
9	70	3	9
10	71	4	16
n = 10	$\sum x = 670$	-	$\sum (x-\bar{x})^2 = 88$

Sample mean, $\bar{x} = \frac{\sum x}{n} = \frac{670}{10} = 67$

Sample standard deviation

$$s = \sqrt{\frac{\sum (x-\bar{x})^2}{n}} = \sqrt{\frac{88}{9}} = 3.13 inches$$

Test static: $t = \frac{\bar{x}-\mu}{s/\sqrt{n}} \; or \; \frac{(x-\mu)\sqrt{n}}{s} \qquad = \frac{\bar{x}-M}{s}\sqrt{n} = \frac{(67-65)\sqrt{10}}{3.13} = 2.02$

H_0: the mean of the universe is 65 inches.

The number of degrees of freedom = v = 10 - 1 = 9.
Tabulated value for 9 d.f. at *5%* level of significance is 2.262.

Since calculated value of t is less than tabulated value for 9 d.f. (2.02 < 2.262). This error could have arisen due to fluctuations and we may conclude that the data are consistent with the assumption of mean height in the universe of 65 inches.

2. Two independent samples of 8 and 7 items respectively had the following values of the variable (weight in ounces):

Sample 1:	9	11	13	11	15	9	12	14
Sample 2:	10	12	10	14	9	8	10	

Is the difference between the means of the sample significant? Given $t_{0.05} = 2.16$.

Solution:

Assumed mean of x = 12, Assumed mean of y = 10

x	$(x-12)$	$(x-12)^2$	y	$(y-10)$	$(y-10)^2$
9	–3	9	10	0	0
11	–1	1	12	2	4
13	1	1	10	0	0
11	–1	1	14	4	16
15	3	9	9	–1	1
9	–3	9	8	–2	4
12	0	0	10	0	0
14	2	4	–	–	–
94	–2	34	73	3	25

$$\bar{x} = \frac{\Sigma x}{n} = \frac{94}{8} = 11.75$$

$$\sigma_x^2 = \frac{\Sigma (x-12)^2}{n} - \left(\Sigma \frac{(x-12)}{n}\right)^2 = \frac{34}{8} - \left(\frac{-2}{8}\right)^2 = 4.1875$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{73}{7} = 10.43$$

$$\sigma_y^2 = \frac{\Sigma (y-10)^2}{n} - \left[\frac{\Sigma (y-10)}{n}\right]^2 = \frac{25}{7} - \left(\frac{3}{7}\right)^2 = 3.438$$

$$s = \sqrt{\frac{(x-\bar{x})^2 + \Sigma (y-\bar{y})^2}{n_1 + n_2 - 2}} = \sqrt{\frac{34+25}{8+7-2}} = \sqrt{\frac{59}{13}} = \sqrt{4.54} = 2.13$$

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{11.75 - 10.43}{2.13\sqrt{\frac{1}{8} + \frac{1}{7}}} = \frac{1.32}{2.13\sqrt{0.268}} = \frac{1.32}{2.13 \times 0.518}$$

$$= \frac{1.32}{1.103} = 1.12$$

The 5% value of t for 13 degree of freedom is given to be 2.16. Since calculated value of t is 1.12 is less than 2.16, the difference between the means of samples is not significant.

3.

3. F-Test for Equality of Population Variances

A random variable X is said to follow F-distribution, if its probability density function is given by

$$f(F) = \frac{(v_1/v_2)^{v_1/2}}{\beta\left(\frac{v_1}{2}, \frac{v_2}{2}\right)} \cdot \frac{F^{\frac{v_1}{2}-1}}{\left(1 + \frac{v_1 F}{v_2}\right)^{(v_1+v_2)/2}}, \quad F > 0$$

where v_1 and v_2 are the degrees of freedom of samples.

Suppose we want to test

(i) Whether two independent samples $x_1, x_2, ..., x_{n1}$ and $y_1, y_2, ..., y_{n2}$ have been drawn from the normal population with the same variance σ^2.

(ii) Whether the two independent estimates of the population variance are homogeneous or not.

Under the null hypothesis (H_0) and when

i. Population variances are equal.
ii. Two independent estimates of population variance are homogeneous.

The test statistic F is given by $F = \frac{S_x^2}{S_y^2}, S_x^2 > S_y^2$ follows F-distribution with (n_1-1, n_2-1) degree's freedom. The value of F is greater than 1.

Problems:

4. Two independent samples of 8 and 7 items respectively had the following values of the variable:

Sample I	**9**	**11**	**13**	**11**	**15**	**9**	**12**	**14**
Sample II	**10**	**12**	**10**	**14**	**9**	**8**	**10**	

Does the estimate of Population variance differ significantly? Given that for 7 degrees of freedom the value of F at 5 % level of significance is 4.20 nearly

Solution:

Sample I		*Sample II*	
x	x^2	y	y^2
9	81	10	100
11	121	12	144
13	169	10	100
11	121	14	196
15	225	9	81
9	81	8	64
12	144	10	100
14	196	—	—
94	1138	73	785

$$\bar{x} = \frac{94}{8} = \mathbf{11.75}, \ \bar{y} = \frac{73}{7} = \mathbf{10.43}$$

$$\sum (x-\bar{x})^2$$

$$= 2\sum x^2 - 2\bar{x}\sum x + \sum \bar{x}^2$$

$$= 1138 - 2\times\frac{94}{8}\times 94 + 8\times\left(\frac{94}{8}\right)^2$$

$$= 1138 - \frac{94^2}{8}$$

$$= 33.5$$

Similarly, $\sum (y-\bar{y})^2 = 23.7$

$$S_1^2 = \frac{\sum (x-\bar{x})^2}{n_1 - 1} = \frac{33.5}{7}$$

$$S_2^2 = \frac{\sum (x-\bar{x})^2}{n_2 - 1} = \frac{23.7}{6}$$

$$\mathbf{F} = \frac{S_1^2}{S_2^2} = \frac{33.5\times 6}{7\times 23.7} = 1.21$$

This calculated value is less than the value of F at 5% level of significance.
Hence differences are not significant. Therefore the samples may well be drawn from the population with same variance.

21.82 CHI-SQUARE (χ^2) TEST

When a coin is tossed 200 times, the theoretical considerations lead us to expect 100 heads and 100 tails. But in practice, these results are rarely achieved. The quantity χ^2 (the Greek letter *chi* squared, pronounced chi-square) describes the magnitude of discrepancy between theory and observation. If $\chi = 0$, the observed and expected frequencies completely coincide. The greater the discrepancy between the observed and expected frequencies, the greater the value of χ^2. Thus χ^2 **affords a measure of the correspondence between theory and observation**.

If O_i ($i = 1, 2, \ldots, n$) is a set of observed (experimental) frequencies and E_i ($i = 1, 2, \ldots, n$) is the corresponding set of expected (theoretical or hypothetical) frequencies, then χ^2 **is defined as**

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

where $\Sigma O_i = \Sigma E_i = N$ (total frequency) and degrees of freedom (*d.f.*) = $(n - 1)$.

Note. (*i*) If $\chi^2 = 0$, the observed and theoretical frequencies agree exactly.

(*ii*) If $\chi^2 > 0$ they do not agree exactly.

21.82.1 Degrees of Freedom

While comparing the calculated value of χ^2 with the table value, we have to determine the degrees of freedom.

If we have to choose any four numbers whose sum is 50, we can exercise our independent choice for any three numbers only, the fourth being 50 minus the total of the three numbers selected. Thus, though we are to choose any four numbers, our choice is reduced to three because of an imposed condition. There is only one restraint on our freedom and our degrees of freedom are $4 - 1 = 3$. If two restrictions are imposed, our freedom to choose will be further curtailed and the degrees of freedom will be $4 - 2 = 2$.

In general, the number of degrees of freedom is the total number of observations less the number of independent constraints imposed on the observations. Degrees of freedom (*d.f.*) are usually denoted by ν (the letter *nu* of the Greek alphabet).

Thus, $\nu = n - k$, where k is the number of independent constraints in a set of data of n observations.

Note. (*i*) For a $p \times q$ contingency table (p columns and q rows), $\nu = (p - 1)(q - 1)$

(*ii*) In the case of a contingency table, the expected frequency of any class

$$= \frac{\text{Total of row in which it occurs} \times \text{Total of columns in which it occurs}}{\text{Total number of observations}}$$

The χ^2 test is one of the simplest and the most general tests known. It is applicable to a very large number of problems in practice, which can be summed up under the following heads:

(*i*) as a test of goodness of fit.
(*ii*) as a test of independence of attributes.
(*iii*) as a test of homogeneity of independent estimates of the population variance.
(*iv*) as a test of the hypothetical value of the population variance σ^2.
(*v*) as a list of the homogeneity of independent estimates of the population correlation coefficient.

21.82.2 Conditions for Applying the χ^2 Test

Following are the conditions that should be satisfied before the χ^2 test can be applied.

(*a*) N, the total number of frequencies, should be large. It is difficult to say what constitutes largeness, but as an arbitrary figure, we may say that **N should be at least 50**, however few the cells.

(*b*) No theoretical cell-frequency should be small. Here again, it is difficult to say what constitutes smallness, but 5 should be regarded as the very minimum and **10 is better**. If small theoretical frequencies occur (*i.e.*, < 10), the difficulty is overcome by grouping two or more classes together before calculating (O – E). **It is important to remember that the number of degrees of freedom is determined with the number of classes after regrouping.**

(*c*) The constraints on the cell frequencies, if any, should be linear.

Note. If any one of the theoretical frequencies is less than 5, we then apply a correction given by F. Yates, which is usually known as "Yates's correction for continuity," we add 0.5 to the cell frequency that is less than 5 and adjust the remaining cell frequency suitably so that the marginal total is not changed.

21.82.3 The χ^2 Distribution

For large sample sizes, the sampling distribution of χ^2 can be closely approximated by a continuous curve known as the chi-square distribution. The probability function of χ^2 distribution is given by

$$f(\chi^2) = c(\chi^2)^{(\nu/2-1)} e^{-x^2/2}$$

where $e = 2.71828$, ν = number of degrees of freedom; c = a constant depending only on ν.

Symbolically, the degrees of freedom are denoted by the symbol ν or by d.f. and are obtained by the rule $\nu = n - k$, where k refers to the number of independent constraints.

In general, when we fit a binomial distribution the number of degrees of freedom is one less than the number of classes; when we fit a Poisson distribution, the degrees of freedom are 2 less than the number of classes, because we use the total frequency and the arithmetic mean to get the parameter of the Poisson distribution. When we fit a normal curve, the number of degrees of freedom are 3 less than the number of classes, because in this fitting we use the total frequency, mean, and standard deviation.

If the data is given in a series of "n" numbers then degrees of freedom = $n - 1$.

In the case of Binomial distribution d.f. = $n - 1$.

In the case of Poisson distribution d.f. = $n - 2$.

In the case of Normal distribution d.f. = $n - 3$.

21.82.4 The χ^2 Test as a Test of Goodness of Fit

The χ^2 test enables us to ascertain how well the theoretical distributions such as Binomial, Poisson, or Normal, etc. fit empirical distributions, *i.e.*, distributions obtained from sample data.

If the **calculated value of χ^2 is less than the table value** at a specified level (generally 5%) of significance, the **fit is considered to be good**, *i.e.*, the divergence between actual and expected frequencies is attributed to fluctuations of simple sampling. If the calculated value of χ^2 is greater than the table value, the fit is considered to be poor.

ILLUSTRATIVE EXAMPLES

Example 1. *The following table gives the number of accidents that took place in an industry during various days of the week. Test whether accidents are uniformly distributed over the week.*

Day	*Mon*	*Tue*	*Wed*	*Thu*	*Fri*	*Sat*
No. of accidents	*14*	*18*	*12*	*11*	*15*	*14*

Sol. Null hypothesis H_0. The accidents are uniformly distributed over the week.

Under this H_0, the expected frequencies of the accidents on each of these days = $\frac{84}{6}$ = 14.

Observed frequency O_i	14	18	12	11	15	14
Expected frequency E_i	14	14	14	14	14	14
$(O_i - E_i)^2$	0	16	4	9	1	0

$$\chi^2 = \frac{\Sigma(O_i - E_i)^2}{E_i} = \frac{30}{14} = 2.1428.$$

Conclusion. Table value of χ^2 at 5% level for (6 – 1 = 5 d.f.) is 11.09.

Since the calculated value of χ^2 is less than the tabulated value, H_0 is accepted, *i.e.*, the accidents are uniformly distributed over the week.

Example 2. *A die is thrown 270 times and the results of these throws are given below:*

No. appeared on the die	*1*	*2*	*3*	*4*	*5*	*6*
Frequency	*40*	*32*	*29*	*59*	*57*	*59*

Test whether the die is biased or not.

Sol. Null hypothesis H_0. Die is unbiased.

Under this H_0, the expected frequencies for each digit is $\frac{276}{6}$ = 46.

To find the value of χ^2

O_i	40	32	29	59	57	59
E_i	46	46	46	46	46	46
$(O_i - E_i)^2$	36	196	289	169	121	169

$$\chi^2 = \frac{\Sigma(O_i - E_i)^2}{E_i} = \frac{980}{46} = 21.30.$$

Conclusion. The tabulated value of χ^2 at 5% level of significance for (6 – 1 = 5) d.f. is 11.09. Since the calculated value of $\chi^2 = 21.30 > 11.07$ the tabulated value, H_0 is rejected.
I.e., the die is not unbiased or the die is biased.

Example 3. *The following table shows the distribution of digits in numbers chosen at random from a telephone directory:*

Digits	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Frequency	*1026*	*1107*	*997*	*966*	*1075*	*933*	*1107*	*972*	*964*	*853*

Test whether the digits may be taken to occur equally frequently in the directory.

Sol. Null hypothesis H_0. The digits taken in the directory occur with equal frequency, *i.e.*, there is no significant difference between the observed and expected frequency.

Under H_0, the expected frequency is given by = $\frac{10,000}{10}$ = 1000

To find the value of χ^2

O_i	1026	1107	997	996	1075	1107	933	972	964	853
E_i	1000	1000	1000	1000	1000	1000	1107	1000	1000	1000
$(O_i - E_i)^2$	676	11449	9	1156	5625	11449	4489	784	1296	21609

$$\chi^2 = \frac{\Sigma(O_i - E_i)^2}{E_i} = \frac{58542}{1000} = 58.542.$$

Conclusion. The tabulated value of χ^2 at 5% level of significance for 9 d.f. is 16.919. Since the calculated value of χ^2 is greater than the tabulated value, H_0 is rejected.
I.e., there is a significant difference between the observed and theoretical frequency.
I.e., the digits taken in the directory do not occur with equal frequency.

Example 4. *Records taken of the number of male and female births in 800 families having four children are as follows:*

No. of male births	*0*	*1*	*2*	*3*	*4*
No. of female births	*4*	*3*	*2*	*1*	*0*
No. of families	*32*	*178*	*290*	*236*	*94*

Test whether the data are consistent with the hypothesis that the binomial law holds and the chance of male birth is equal to that of female birth, namely $p = q = 1/2$.

Sol. H_0: The data are consistent with the hypothesis of equal probability for male and female births, *i.e.*, $p = q = 1/2$.

We use binomial distribution to calculate theoretical frequency given by:

$$N(r) = N \times P(X = r)$$

where N is the total frequency. $N(r)$ is the number of families with r male children:

$$P(X = r) = {}^nC_r p^r q^{n-r}$$

where p and q are the probability of male and female births, n is the number of children.

N(0) = No. of families with 0 male children = $800 \times {}^4C_0 \left(\frac{1}{2}\right)^4 = 800 \times 1 \times \frac{1}{2^4} = 50$

$$N(1) = 800 \times {}^4C_1 \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^3 = 200; \quad N(2) = 800 \times {}^4C_2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^2 = 300$$

$$N(3) = 800 \times {}^4C_3 \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^3 = 200; \quad N(4) = 800 \times {}^4C_4 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^4 = 50$$

Observed frequency O_i	32	178	290	236	94
Expected frequency E_i	50	200	300	200	50
$(O_i - E_i)^2$	324	484	100	1296	1936
$\frac{(O_i - E_i)^2}{E_i}$	6.48	2.42	0.333	6.48	38.72

$$\chi^2 = \frac{\Sigma(O_i - E_i)^2}{E_i} = 54.433.$$

Conclusion. The table value of χ^2 at 5% level of significance for 5 – 1 = 4 d.f. is 9.49.

Since the calculated value of χ^2 is greater than the tabulated value, H_0 is rejected. *I.e.*, the data are not consistent with the hypothesis that the binomial law holds and that the chance of a male birth is not equal to that of a female birth.

Note. Since the fitting is binomial, the degrees of freedom $\nu = n - 1$, *i.e.*, $\nu = 5 - 1 = 4$.

Example 5. *Verify whether the Poisson distribution can be assumed from the data given below*:

No. of defects	*0*	*1*	*2*	*3*	*4*	*5*
Frequency	*6*	*13*	*13*	*8*	*4*	*3*

Sol. H_0 : The Poisson fit is a good fit to the data.

Mean of the given distribution = $\frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{94}{47} = 2$

To fit a Poisson distribution we require m. Parameter $m = \bar{x} = 2$.

By the Poisson distribution the frequency of r success is

$$N(r) = N \times e^{-m} \cdot \frac{m^r}{r!}, \text{ N is the total frequency.}$$

$$N(0) = 47 \times e^{-2} \cdot \frac{(2)^0}{0!} = 6.36 \approx 6; \qquad N(1) = 47 \times e^{-2} \cdot \frac{(2)^1}{1!} = 12.72 \approx 13$$

$$N(2) = 47 \times e^{-2} \cdot \frac{(2)^2}{2!} = 12.72 \approx 13; \qquad N(3) = 47 \times e^{-2} \cdot \frac{(2)^3}{3!} = 8.48 \approx 9$$

$$N(4) = 47 \times e^{-2} \cdot \frac{(2)^4}{4!} = 4.24 \approx 4; \qquad N(5) = 47 \times e^{-2} \cdot \frac{(2)^5}{5!} = 1.696 \approx 2.$$

X	0	1	2	3	4	5
O_i	6	13	13	8	4	3
E_i	6.36	12.72	12.72	8.48	4.24	1.696
$\frac{(O_i - E_i)^2}{E_i}$	0.2037	0.00616	0.00616	0.02716	0.0135	1.0026

$$\chi^2 = \frac{\Sigma(O_i - E_i)^2}{E_i} = 1.2864.$$

Conclusion. The calculated value of χ^2 is 1.2864. The tabulated value of χ^2 at 5% level of significance for $\gamma = 6 - 2 = 4$ d.f. is 9.49. Since the calculated value of χ^2 is less than that of the tabulated value, H_0 is accepted, *i.e.*, the Poisson distribution provides a good fit to the data.

Example 6. *The theory predicts the proportion of beans in the four groups,* G_1, G_2, G_3, G_4 *should be in the ratio 9 : 3 : 3 : 1. In an experiment with 1600 beans the numbers in the four groups were 882, 313, 287, and 118. Does the experimental result support the theory?*

Sol. H_0. The experimental result supports the theory, *i.e.*, there is no significant difference between the observed and theoretical frequency under H_0; the theoretical frequency can be calculated as follows:

$$E(G_1) = \frac{1600 \times 9}{16} = 900; \qquad E(G_2) = \frac{1600 \times 3}{16} = 300;$$

$$E(G_3) = \frac{1600 \times 3}{16} = 300; \qquad E(G_4) = \frac{1600 \times 1}{16} = 100$$

To calculate the value of χ^2

Observed frequency O_i	882	313	287	118
Expected frequency E_i	900	300	300	100
$\frac{(O_i - E_i)^2}{E_i}$	0.36	0.5633	0.5633	3.24

$$\chi^2 = \frac{\Sigma(O_i - E_i)^2}{E_i} = 4.7266.$$

Conclusion. The table value of χ^2 at 5% level of significance for 3 d.f. is 7.815. Since the calculated value of χ^2 is less than that of the tabulated value, hence H_0 is accepted.
I.e., the experimental results support the theory.

TEST YOUR KNOWLEDGE

1. The following table gives the frequency of occupance of the digits 0, 1, . . . , 9 in the last place in four logarithms of numbers 10–99. Examine whether there is any peculiarity.

Digits :	0	1	2	3	4	5	6	7	8	9
Frequency :	6	16	15	10	12	12	3	2	9	5

2. The sales in a supermarket during a week are given below. Test the hypothesis that the sales do not depend on the day of the week, using a significance level of 0.05.

Days :	Mon	Tues	Wed	Thurs	Fri	Sat
Sales (in \$10000) :	65	54	60	56	71	84

3. A survey of 320 families with 5 children each revealed the following information:

No. of boys :	5	4	3	2	1	0
No. of girls :	0	1	2	3	4	5
No. of families :	14	56	110	88	40	12

Is this result consistent with the hypothesis that male and female births are equally probable?

4. 4 coins were tossed at a time and this operation was repeated 160 times. It is found that 4 heads occur 6 times, 3 heads occur 43 times, 2 heads occur 69 times, and one head occur 34 times. Discuss whether the coin may be regarded as unbiased.

5. Fit a Poisson distribution to the following data and the best goodness of fit:

x :	0	1	2	3	4
f :	109	65	22	3	1

6. In the accounting department of a bank, 100 accounts are selected at random and estimated for errors. The following results were obtained:

No. of errors :	0	1	2	3	4	5	6
No. of accounts :	35	40	19	2	0	2	2

Does this information verify that the errors are distributed according to the Poisson probability law?

7. In a sample analysis of examination results of 500 students, it was found that 280 students have failed, 170 have gotten C's, 90 have gotten B's, and the rest, A's. Do these figures support the general belief that the above categories are in the ratio 4 : 3 : 2 : 1 respectively?

Answers

1. no **2.** accepted **3.** accepted **4.** unbiased
5. Poisson law fits the data **6.** maybe **7.** yes

21.82.5 The χ^2 Test as a Test of Independence

With the help of the χ^2 test, we can find whether or not two attributes are associated. We take the null hypothesis that there is no association between the attributes under study, *i.e.*, **we assume that the two attributes are independent**. **If the calculated value of χ^2 is less than the table value** at a specified level (generally 5%) of significance, the hypothesis holds true, *i.e.*, **the attributes are independent** and do not bear any association. On the other hand, if the calculated value of χ^2 is greater than the table value at a specified level of significance, we say that the results of the experiment do not support the hypothesis. In other words, the attributes are associated. Thus a very useful application of the χ^2 test is to investigate the relationship between trials or attributes, which can be classified into two or more categories.

The sample data are set out into a two-way table, called a **contingency table.**

Let us consider two attributes A and B divided into r classes $A_1, A_2, A_3, \ldots, A_r$ and B divided into s classes $B_1, B_2, B_3, \ldots, B_s$. If (A_i), (B_j) represents the number of people possessing the attributes A_i, B_j respectively, ($i = 1, 2, \ldots, r$, $j = 1, 2, \ldots, s$) and $(A_i\ B_j)$ represent the number of people possessing attributes A_i and B_j. Also we have $\sum_{i=1}^{r} A_i = \sum_{i=1}^{s} B_j$ where N is the total frequency. The contingency table for $r \times s$ is given below:

A / *B*	A_1	A_2	A_3	$\ldots A_r$	*Total*
B_1	(A_1B_1)	(A_2B_1)	(A_3B_1)	$\ldots (A_rB_1)$	B_1
B_2	(A_1B_2)	(A_2B_2)	(A_3B_2)	$\ldots (A_rB_2)$	B_2
B_3	(A_1B_3)	(A_2B_3)	(A_3B_3)	$\ldots (A_rB_3)$	B_3
...	...	...	...	...	...
...	...	...	...	...	...
B_s	(A_1B_s)	(A_2B_s)	(A_3B_s)	$\ldots (A_rB_s)$	(B_s)
Total	(A_1)	(A_2)	(A_3)	$\ldots (A_r)$	N

H_0 : Both the attributes are independent, *i.e.*, A and B are independent under the null hypothesis; we calculate the expected frequency as follows:

$$\mathrm{P}(\mathrm{A}_i) = \text{ Probability that a person possesses the attribute } \mathrm{A}_i = \frac{(\mathrm{A}_i)}{\mathrm{N}}\, i = 1, 2, \ldots, r$$

$$\mathrm{P}(\mathrm{B}_j) = \text{ Probability that a person possesses the attribute } \mathrm{B}_j = \frac{(\mathrm{B}_j)}{\mathrm{N}}$$

$$\mathrm{P}(\mathrm{A}_i\mathrm{B}_j) = \text{Probability that a person possesses both attributes } \mathrm{A}_i \text{ and } \mathrm{B}_j = \frac{(\mathrm{A}_i\mathrm{B}_j)}{\mathrm{N}}$$

If $(\mathrm{A}_i\mathrm{B}_j)_0$ is the expected number of people possessing both the attributes A_i and B_j

$$(\mathrm{A}_i\mathrm{B}_j)_0 = \mathrm{NP}(\mathrm{A}_i\mathrm{B}_j) = \mathrm{NP}(\mathrm{A}_i)(\mathrm{B}_j)$$

$$= \mathrm{N}\frac{(\mathrm{A}_i)}{\mathrm{N}}\frac{(\mathrm{B}_j)}{\mathrm{N}} = \frac{(\mathrm{A}_i)(\mathrm{B}_j)}{\mathrm{N}} \qquad (\because \text{ A and B are independent})$$

Hence $$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\left[\frac{\left[(\mathrm{A}_i\mathrm{B}_j) - (\mathrm{A}_i\mathrm{B}_j)_0\right]^2}{(\mathrm{A}_i\mathrm{B}_j)_0}\right]$$

which is distributed as a χ^2 variate with $(r-1)(s-1)$ degrees of freedom.

Note 1. For a 2 × 2 contingency table where the frequencies are $\begin{array}{c|c} a & b \\ c & d \end{array}$ χ^2 can be calculated from independent frequencies as $\chi^2 = \dfrac{(a+b+c+d)(ad-bc)^2}{(a+b)(c+d)(b+d)(a+c)}$.

Note 2. If the contingency table is not 2 × 2, then the formula for calculating χ^2 as given in Note 1, cannot be used. Hence, we have another formula for calculating the expected frequency $(A_iB_j)_0 = \dfrac{(A_i)(B_j)}{N}$

I.e., the expected frequency in each cell is = $\dfrac{\text{Product of column total and row total}}{\text{whole total}}$.

Note 3. If $\dfrac{a \mid b}{c \quad d}$ is the 2 × 2 contingency table with two attributes, $Q = \dfrac{ad-bc}{ad+bc}$ is called the coefficient of association.

If the attributes are independent then $\dfrac{a}{b} = \dfrac{c}{d}$.

Note 4. Yate's Correction. In a 2 × 2 table, if the frequencies of a cell is small, we make Yates's correction to make χ^2 continuous.

Decrease by $\frac{1}{2}$ those cell frequencies that are greater than expected frequencies, and increase by $\frac{1}{2}$ those that are less than expected. This will not affect the marginal columns. This correction is known as Yates's correction to continuity.

After Yates's correction
$$\chi^2 = \frac{N\left(bc - ad - \frac{1}{2}N\right)^2}{(a+c)(b+d)(c+d)(a+b)} \quad \text{when} \quad ad - bc < 0$$

$$\chi^2 = \frac{N\left(ad - bc - \frac{1}{2}N\right)^2}{(a+c)(b+d)(c+d)(a+b)} \quad \text{when} \quad ad - bc > 0.$$

ILLUSTRATIVE EXAMPLES

Example 1. *What are the expected frequencies of the 2 × 2 contingency tables given below*:

(*i*)

a	*b*
c	*d*

(*ii*)

2	10
6	6

Sol. Observed frequencies → Expected frequencies

(*i*)

a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	$a+b+c+d=N$

→

$\dfrac{(a+c)(a+b)}{a+b+c+d}$	$\dfrac{(b+d)(a+b)}{a+b+c+d}$
$\dfrac{(a+c)(c+d)}{a+b+c+d}$	$\dfrac{(b+d)(c+d)}{a+b+c+d}$

	Observed frequencies				Expected frequencies	
(*ii*)	2	10	12		$\frac{8\times12}{24}=4$	$\frac{16\times12}{24}=8$
	6	6	12	$\rightarrow$	$\frac{8\times12}{24}=4$	$\frac{16\times12}{24}=8$
	8	16	24			

Example 2. *From the following table regarding the color of eyes of fathers and sons test whether the color of the son's eye is associated with that of the father.*

		Eye color of son	
		Light	*Not light*
Eye color of father	*Light*	*471*	*51*
	Not light	*148*	*230*

Sol. Null hypothesis H_0. The color of the son's eye is not associated with that of the father, *i.e.*, they are independent.

Under H_0, we calculate the expected frequency in each cell as

$$=\frac{\text{Product of column total and row total}}{\text{whole total}}$$

Expected frequencies are:

Eye color of son / *Eye color of father*	*Light*	*Not light*	*Total*
Light	$\frac{619\times522}{900}=359.02$	$\frac{289\times522}{900}=167.62$	522
Not light	$\frac{619\times378}{900}=259.98$	$\frac{289\times378}{900}=121.38$	378
Total	619	289	900

$$\chi^2=\frac{(471-359.02)^2}{359.02}+\frac{(51-167.62)^2}{167.62}+\frac{(148-259.98)^2}{259.98}+\frac{(230-121.38)^2}{121.38}$$
$$=261.498.$$

Conclusion. Tabulated value of χ^2 at 5% level for 1 d.f. is 3.841.

Since the calculated value of χ^2 > the tabulated value of χ^2, H_0 is rejected. They are dependent, *i.e.*, the color of the son's eye is associated with that of the father.

Example 3. *The following table gives the number of good and bad parts produced by each of the three shifts in a factory*:

	Good parts	*Bad parts*	*Total*
Day shift	*960*	*40*	*1000*
Evening shift	*940*	*50*	*990*
Night shift	*950*	*45*	*995*
Total	*2850*	*135*	*2985*

Test whether or not the production of bad parts is independent of the shift on which they were produced.

Sol. Null hypothesis H_0. The production of bad parts is independent of the shift on which they were produced.
I.e., the two attributes, production and shifts, are independent.

Under H_0,
$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{3}\left[\frac{\left[(A_iB_j)_0 - (A_iB_j)\right]^2}{(A_iB_j)_0}\right]$$

Calculation of expected frequencies

Let A and B be two attributes, namely, production and shifts. A is divided into two classes A_1, A_2, and B is divided into three classes B_1, B_2, B_3.

$$(A_1B_1)_0 = \frac{(A_1)(B_2)}{N} = \frac{(2850)\times(1000)}{2985} = 954.77$$
$$(A_1B_2)_0 = \frac{(A_1)(B_2)}{N} = \frac{(2850)\times(990)}{2985} = 945.226$$
$$(A_1B_3)_0 = \frac{(A_1)(B_3)}{N} = \frac{(2850)\times(995)}{2985} = 950$$
$$(A_2B_1)_0 = \frac{(A_2)(B_1)}{N} = \frac{(135)\times(1000)}{2985} = 45.27$$
$$(A_2B_2)_0 = \frac{(A_2)(B_2)}{N} = \frac{(135)\times(990)}{2985} = 44.773$$
$$(A_2B_3)_0 = \frac{(A_2)(B_3)}{N} = \frac{(135)\times(995)}{2985} = 45.$$

To calculate the value of χ^2

Class	O_i	E_i	$(O_i - E_i)^2$	$(O_i - E_i)^2 / E_i$
(A_1B_1)	960	954.77	27.3529	0.02864
(A_1B_2)	940	945.226	27.3110	0.02889
(A_1B_3)	950	950	0	0
(A_2B_1)	40	45.27	27.7729	0.61349
(A_2B_2)	50	44.773	27.3215	0.61022
(A_2B_3)	45	45	0	0
				1.28126

Conclusion. The tabulated value of χ^2 at 5% level of significance for 2 degrees of freedom $(r-1)(s-1)$ is 5.991. Since the calculated value of χ^2 is less than the tabulated value, we accept H_0, *i.e.*, the production of bad parts is independent of the shift on which they were produced.

Example 4. *From the following data, find whether hair color and sex are associated.*

Sex \ *Color*	*Fair*	*Red*	*Medium*	*Dark*	*Black*	*Total*
Boys	*592*	*849*	*504*	*119*	*36*	*2100*
Girls	*544*	*677*	*451*	*97*	*14*	*1783*
Total	*1136*	*1526*	*955*	*216*	*50*	*3883*

Sol. Null hypothesis H_0. The two attributes of hair color and sex are not associated, *i.e.*, they are independent.

Let A and B be the attributes of hair color and sex, respectively. A is divided into 5 classes $(r = 5)$. B is divided into 2 classes $(s = 2)$.

$\therefore$ Degrees of freedom $= (r-1)(s-1) = (5-1)(2-1) = 4$

Under H_0, we calculate $\chi^2 = \sum_{i=1}^{5}\sum_{j=1}^{2} \frac{\left[(A_iB_j)_0 - (A_iB_j)\right]^2}{(A_iB_j)_0}$

Calculate the expected frequency $(A_iB_j)_0$ as follows:

$$(A_1B_1)_0 = \frac{(A_1)(B_1)}{N} = \frac{1136 \times 2100}{3883} = 614.37$$

$$(A_1B_2)_0 = \frac{(A_1)(B_2)}{N} = \frac{1136 \times 1783}{3883} = 521.629$$

$$(A_2B_1)_0 = \frac{(A_2)(B_1)}{N} = \frac{1526 \times 2100}{3883} = 852.289$$

$$(A_2B_2)_0 = \frac{(A_2)(B_2)}{N} = \frac{1526 \times 1783}{3883} = 700.71$$

$$(A_3B_1)_0 = \frac{(A_3)(B_1)}{N} = \frac{955 \times 2100}{3883} = 516.482$$

$$(A_3B_2)_0 = \frac{(A_3)(B_2)}{N} = \frac{955 \times 1783}{3883} = 483.517$$

$$(A_4B_1)_0 = \frac{(A_4)(B_1)}{N} = \frac{216 \times 2100}{3883} = 116.816$$

$$(A_4B_2)_0 = \frac{(A_4)(B_2)}{N} = \frac{216 \times 1783}{3883} = 99.183$$

$$(A_5B_1)_0 = \frac{(A_5)(B_1)}{N} = \frac{50 \times 2100}{3883} = 27.04$$

$$(A_5B_2)_0 = \frac{(A_5)(B_2)}{N} = \frac{50 \times 1783}{3883} = 22.959$$

Calculation of χ^2

Class	O_i	E_i	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
A_2B_1	592	614.37	500.416	0.8145
A_1B_2	544	521.629	500.462	0.959
A_2B_1	849	852.289	10.8175	0.0127
A_2B_2	677	700.71	562.1641	0.8023
A_3B_1	504	516.482	155.800	0.3016
A_3B_2	451	438.517	155.825	0.3553
A_4B_1	119	116.816	4.7698	0.0408
A_4B_2	97	99.183	4.7654	0.0480
A_5B_1	36	27.04	80.2816	2.9689
A_5B_2	14	22.959	80.2636	3.495
				9.79975

$$\chi^2 = 9.799.$$

Conclusion. Table of χ^2 at 5% level of significance for 4 d.f. is 9.488.

Since the calculated value of χ^2 < tabulated value H_0 is rejected, *i.e.*, the two attributes are not independent, *i.e.*, the hair color and sex are associated.

Example 5. *Can vaccination be regarded as a preventive measure of smallpox as evidenced by the following data of 1482 people exposed to small pox in a locality? 368 in all were attacked of these 1482 people, and 343 were vaccinated, and of these only 35 were attacked.*

Sol. For the given data we form the contingency table. Let the two attributes be vaccination and exposed to smallpox. Each attribute is divided into two classes.

Disease smallpox B \ Vaccination A	*Vaccinated*	*Not*	*Total*
Attacked	35	333	368
Not	308	806	1114
Total	343	1139	1482

Null hypothesis H_0. The two attributes are independent, *i.e.*, vaccination cannot be regarded as a preventive measure of smallpox.

Degrees of freedom $\nu = (r-1)(s-1) = (2-1)(2-1) = 1$

Under H_0,
$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\left[(A_iB_j)_0 - (A_iB_j)\right]^2}{(A_iB_j)_0}$$

Calculation of expected frequency

$$(A_1B_1)_0 = \frac{(A_1)(B_1)}{N} = \frac{343 \times 368}{1482} = 85.1713$$

$$(A_1B_2)_0 = \frac{(A_1)(B_2)}{N} = \frac{343 \times 1114}{1482} = 257.828$$

$$(A_2B_1)_0 = \frac{(A_2)(B_1)}{N} = \frac{1139 \times 368}{1482} = 282.828$$

$$(A_2B_2)_0 = \frac{(A_2)(B_2)}{N} = \frac{1139 \times 1114}{1482} = 856.171$$

Calculation of χ^2

Class	O_i	E_i	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
(A_1B_1)	35	85.1713	2517.159	29.554
(A_1B_2)	308	257.828	2517.229	8.1728
(A_2B_1)	333	282.828	2517.2295	7.5592
(A_2B_2)	806	856.171	2517.1292	2.9399
				48.2261

Calculated value of $\chi^2 = 48.2261$.

Conclusion. Tabulated value of χ^2 at 5% level of significance for 1 d.f. is 3.841. Since the calculated value of $\chi^2 >$ tabulated value H_0 is rejected.

I.e., the two attributes are not independent, *i.e.*, the vaccination can be regarded as a preventive measure of smallpox.

TEST YOUR KNOWLEDGE

1. In a locality 100 people were randomly selected and asked about their educational achievements. The results are given below:

Education				
Sex		*Middle*	*High school*	*College*
	Male	10	15	25
	Female	25	10	15

Based on this information, can you say the education depends on sex?

2. The following data is collected on two characteristics:

	Smokers	*Nonsmokers*
Literate	83	57
Illiterate	45	68

Based on this information can you say that there is no relation between habit of smoking and literacy?

3. 500 students at school were graded according to their intelligences and economic conditions of their homes. Examine whether there is any association between economic condition and intelligence, from the following data:

Economic conditions	Intelligence	
	Good	Bad
Rich	85	75
Poor	165	175

4. In an experiment on the immunization of goats from anthrax, the following results were obtained. Derive your inferences on the efficiency of the vaccine.

	Died from anthrax	*Survived*
Inoculated with vaccine	2	10
Not inoculated	6	6

Answers

1. Yes **2.** No **3.** No **4.** Not effective.

Design of experiments

For a scientific investigation we conduct an experiment by collection of data or measurement of an object according to certain sampling procedure. Suppose we conduct an agricultural experiment to verify the truth to claim that the fertilizers increase the yeild of wheat. Then the two variables, fertilizers and yeild of wheat are involved directly. These two variables are called as experimental variables. In addition, quality of seed, climate, nature of soil and all other associated variables are known as extraneous variables.

The main aim of design of experiments are to control the extraneous variables, to minimize the experimental error.

Experiment

An experiment is a device or a means of getting an answer to the problem under consideration.

Absolute experiments deals with determining the absolute value of some characteristics like obtaining the average intelligence quotient of a group of people.

Comparitive experiments are designed to compare the effect of two or more objects on some population characteristics.

eg. Comparison of different kinds of varities of crops.

Treatment.

Various objects of comparison in a comparitive experiment are termed as treatment.

eg. In a field experiment, different fertilizers, dif. varities of crop, diff. methods of cultivation.

Experimental unit.

The smallest division of the experimental material to which we apply the treatments and on which we make observations of the variable under study is termed as experimental unit.

eg. land

Yield

The measurement of the variable under study on different experimental units are termed as yields.

Experimental error.

It includes all types of extraneous variations due to

i. the inherent variability in the experimental material to which treatments are applied.

ii. the lack of uniformity in the methodology of conducting the experiment or in other words failure of standardised experimental techniques.

To control the effect of extraneous variables, we use grouping and blocking.

By grouping we mean combining sets of homogenous experimental units. The different groups need not have the same no. of "

By blocking we mean assigning same no. of experimental units in different blocks. Each block will have comparitively homogenous experimental units.

Completely Randomised Design,
Randomised Block " ,
Latin square " ,
2^2 Factorial design, Taguchi's robust parameter design

CRD.

CRD is the simplest of all the designs based on the principles of randomisation and replication. In this design, treatments are allocated at random to experimental units over the entire experimental material.

Let us suppose that we have v treatments. The i^{th} treatment being replicated r_i times ($i = 1$ to v). Then the whole experimental material is divided into $n = \Sigma r_i$ experimental units and the treatments are distributed completely at random over the units subject to the condition that the i^{th} treatment occurs r_i times. Randomization assures that extraneous factors do not continuously influence one factor.

Advantages of CRD.

There is a complete flexibility in the model as the no. of replications is not fixed.

Analysis can be performed even if some observations are missing.

CRD results in the maximum use of the experimental units since all the experimental material can be used.

Disadvantage.

The experimental error is large as compared to the other designs since the homogenity of the unit is ignored.

Statistical analysis of CRD.

ANOVA is a technique used to test the means of more than two samples. It divides the total variance in the group into parts, which are associated to different factors. This variation is split into two components as

variation within subgroups.

variation between the subgroups.

Model eqn.

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad 1 \le i \le v, \quad 1 \le j \le n \text{ where}$$

y_{ij} - yield or response from the j^{th} unit receiving the i^{th} treatment

α_i - effect due to the i^{th} treatment.

ε_{ij} - error effect (independent normally distributed random variables with zero means and common variance σ^2)

μ - grand mean

H_0: The v pop. means are equal. or $\alpha_1 = \alpha_2 = \dots = \alpha_v = 0$.

H_1: Atleast two of the pop means are unequal. $\alpha_i \neq 0$ for some i.

ANOVA Table

Source of variation	Degrees of freedom	Sum of squares	Mean Square	F ratio
Treatments	$v-1$	SST	$MST = \frac{SST}{v-1}$	$F = \frac{MST}{MSE}$
Error	$n-v$	SSE	$MSE = \frac{SSE}{n-v}$	
Total	$n-1$	$SST+SSE$		

where v - no. of treatments

$$n = \sum_{i=1}^{v} r_i \text{ - total no. of experimental units}$$

$$SST = \sum_{i=1}^{v} \frac{T_i^2}{r_i} - C.F \quad , \; C.F \text{ - Correction Factor} = \frac{G^2}{n}, \; G \text{ - Grand total}$$

$$SSE = TSS - SST,$$

$$TSS = \sum_j \sum_i y_{ij}^2 - C.F$$

For convenience we denote

SST = Q_1 , SSE = Q_2

and TSS = $Q_1 + Q_2$

Table value $F_\alpha(v-1, n-v)$

1. A set of data involving four tropical feeds A, B, C, D tried on 20 chicks are given below. All the 20 chicks are treated alike in all aspects except the feeding treatment. Each feeding treatment is given to 5 chicks. Analyse the data.

	Weight gain					Total
A	55	49	42	21	52	219
B	61	112	30	89	63	355
C	42	97	81	95	92	407
D	169	137	169	85	154	714
						1695

H_0 : Effects due to four feeds are equal.

i.e. $H_0 : \alpha_A = \alpha_B = \alpha_C = \alpha_D$.

$H_1 : \alpha_A \neq \alpha_B \neq \alpha_C \neq \alpha_D$.

$G = 1695, \; n = 20, \; v = 4, \; r_i = 5, \; 1 \le i \le 4$

$$C.F = \frac{G^2}{n} = 143651.25$$

$TSS = 181445 - 143651.25 = 37793.75$

$SST = 169886.2 - \text{''} = 26234.95$

$SSE = 11558.8$

ANOVA Table

Source of variation	Degrees of freedom	Sum of squares	Mean square	F
Treatment	$v-1 = 3$	26234.95	8744.983	12.105
Error	$n-v=16$	11558.8	722.425	
Total	19	37793.95		

$F_{0.05}^{(3,16)} = 3.24$

$F > F_\alpha$.

$\therefore$ Reject H_0.

2. Set the ANOVA table for the following per hectare yield for three varieties of wheat each grown in 4 plots.

Plots	Varities of wheat		
.	A_1	A_2	A_3
1	6	5	5
2	7	5	4
3	3	3	3
4	8	7	4
Total	24	20	16	60

Test whether there is a significant dif. among yield in the three varieties of wheat.

$H_0: \alpha_{A_1} = \alpha_{A_2} = \alpha_{A_3}$

$G = 60, \ v = 3, \ n = 12.$

$C.F = 300$

$TSS = 332 - 300 = 32$

$SST = 8$

$SSE = 24$

Source of variation	Degrees of freedom	S.S	M.S	F
Treatment	2	8	4	1.5003
Error	9	24	2.666	
Total	11	32		

$F_{0.05}(2,9) = 4.26$

$F < F_\alpha$.

$\therefore$ Accept H_0.

3. A completely randomised design 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

Analyse the results for treatment effects.

Solution:

Rearranging the data according to the treatments, we have the following table:

Treatment	Yield from plots (x_{ij})				T_i	T_i^2.	n_i	$\frac{T_i^2}{n_i}$
A	5	7	3	1	16	256	4	64
B	4	4	7	–	15	225	3	75
C	3	5	1	–	9	81	3	27
				Total	$T = 40$	–	$N = 10$	166

$$\sum\sum x_{ij}^2 = (25 + 49 + 9 + 1) + (16 + 16 + 49) + (9 + 25 + 1)$$
$$= 84 + 81 + 35 = 200$$

$$Q = \sum\sum x_{ij}^2 - \frac{T^2}{N} = 200 - \frac{40^2}{10} = 200 - 160 = 40$$

$$Q_1 = \sum \frac{T_i^2}{n_i} - \frac{T^2}{N} = 166 - 160 = 6$$

$$Q_2 = Q - Q_1 = 40 - 6 = 34$$

ANOVA table

S.V.	S.S.	d.f.	M.S.	F_0
Between classes (treatments)	$Q_1 = 6$	$h - 1 = 2$	3.0	$\frac{4.86}{3.0}$
Within classes	$Q_2 = 34$	$N - h = 7$	4.86	$= 1.62$
Total	$Q = 40$	$N - 1 = 9$	–	–

From the F-table, $F_{5\%}\ (\nu_1 = 7, \nu_2 = 2) = 19.35$
We note that $F_0 < F_{5\%}$
Let H_0 : The treatments do not differ significantly.
$\therefore$ The null hypothesis is accepted.
i.e., the treatments are not significantly different.

4. A manufacturing company has purchased 3 new machines of different makes and wise to determine whether one of them is faster than the others in producing certain output. Five hourly production figures are observed at random from each machine and the results are given below.

observations.	A_1	A_2	A_3
1	25	31	24
2	30	39	30
3	36	38	28
4	38	42	25
5	31	35	28

Use analysis of variance and determine whether the machines are different in their mean speed.

Ans: Here $N = 15$, $k = 3$, $n_1 = n_2 = n_3 = 5$

H_0: The machines are donot differ significantly.

H_1: The machines are differ significantly.

Table:

machines	observations	T_i	T_i^2	n_i	T_i^2/n_i
A_1	25 30 36 38 31	160	25600	5	5120
A_2	31 39 38 42 35	185	34225	5	6845
A_3	24 30 28 25 28	135	18225	5	3645
Total		$T=480$	-	$N=15$	15610

$\Rightarrow \sum\sum x_{ij}^2 = 15810$

$$Q = \sum\sum x_{ij}^2 - \frac{T^2}{N} = 15810 - \frac{480^2}{15}$$

$$= 450$$

$$Q_1 = \sum_i \frac{T_i^2}{n_i} - \frac{T^2}{N} = \frac{160^2}{5} + \frac{185^2}{5} + \frac{135^2}{5} - \frac{480^2}{15}$$

$$= 5120 + 6845 + 3645 - 15360$$

$$= 15610 - 15360$$

$$= 250$$

$Q_1 = 250$

$Q_2 = Q - Q_1 = 450 - 250 = 200$

$k = 3$

∴ Table: (ANOVA)

S.V	S.S	d.f	M.S.S	F
Between machine	$Q_1 = 250$	$h-1=2$	$\frac{250}{2} = 125$	$F = \frac{125}{16.67}$
Within machine	$Q_2 = 200$	$N-h = 15-3 = 12$	$\frac{200}{12} = 16.67$	$= 7.50$
Total	450	14		

$\Rightarrow F_{cal} = 7.50$

from table: $F_{5\%}(\nu_1 = 2, \nu_2 = 12) = 3.89$

Since $F_{cal} > F_{tab}$

$\Rightarrow$ we reject H_0.

That is There is a significance difference in the three machine.

Exercise:

The following table shows the lives in hours of four brands of electric lamps:

Brand

A : 1610, 1610, 1650, 1680, 1700, 1720, 1800

B : 1580, 1640, 1640, 1700, 1750

C : 1460, 1550, 1600, 1620, 1640, 1660, 1740, 1820

D : 1510, 1520, 1530, 1570, 1600, 1680

Perform an analysis of variance and test the homogeneity of the mean lives of the four brands of lamps.

Randomised Block Design (2 way classification)

If the whole experimental area is not homogenous, then a simple method of controlling the variability of experimental material consists in grouping the whole area into relatively homogenous subgroups. The treatments can be applied in a random manner to relatively homogenous units within each subgroup / block and replicated over all the blocks. This design is known as RBD.

Layout.

Let us consider 5 treatments A, B, C, D, E each replicated four times. We divide the whole experimental area into 4 relatively homogenous blocks and each block into 5 units. Treatments are allocated at random to the plots of blocks.

Blocks

I	A	E	B	D	C
II	E	D	C	B	A
III	C	B	A	E	D
IV	A	D	E	C	B

Advantages.

1. It has a simple layout
2. This design controls the variability in the experimental units and gives the treatments equivalence to show their effects.

Disadv. It is not suitable for large no. of treatments.

Model eqn.

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \quad , \quad 1 \le i \le a, \; 1 \le j \le b$$

a - no. of treatments

b - " " blocks

μ - grand mean

α_i - effect of i^{th} treatment

β_j - " " j^{th} block

ε_{ij} - independent normally distributed random variable having mean 0 and variance σ^2.

ANOVA Table

Source of variation	Degrees of freedom	Sum of squares	M.S.	F
Treatments	$a-1$	SST	$MST = \frac{SST}{a-1}$	$F_T = \frac{MST}{MSE}$
Blocks	$b-1$	SSB	$MSB = \frac{SSB}{b-1}$	$F_B = \frac{MSB}{MSE}$
Error (residual)	$(a-1)(b-1)$	SSE	$MSE = \frac{SSE}{(a-1)(b-1)}$	
Total	$ab-1$	TSS		

$$C.F = \frac{G^2}{ab}$$

$$TSS = \sum_j \sum_i y_{ij}^2 - C.F$$

$$SST = \frac{\sum_{i=1}^{a} T_i^2}{b} - C.F$$

$$SSB = \frac{\sum_{j=1}^{b} B_j^2}{a} - C.F$$

$$SSE = TSS - (SST + SSB)$$

If $F_T < F_\alpha((a-1),(a-1)(b-1))$, accept H_α

If $F_B < F_\alpha((b-1),(a-1)(b-1))$, accept H_β.

For convenience we denote

SST = Q_1, SSB = Q_2, SSE = Q_3 and TSS = $Q_1 + Q_2 + Q_3$

1. Consider the results given in the fol. table involving 6 treatments in 4 randomised blocks. The treatments indicated by numbers within paranthesis. Analyse whether there is any significant difference between the treatments and blocks are homogenous.

Treatments and yield

Blocks						
1	(1) 24.7	(2) 27.7	(3) 20.6	(4) 16.2	(5) 16.2	(6) 24.9
2	(3) 22.9	(2) 28.8	(1) 27.3	(4) 15	(6) 22.5	(5) 17
3	(6) 26.3	(4) 19.6	(1) 38.5	(3) 36.8	(2) 39.5	(5) 15.4
4	(5) 17.7	(2) 31	(1) 28.5	(4) 14.1	(3) 34.9	(6) 22.6

$H_T : \alpha_1 = \alpha_2 = \cdots = \alpha_6$

$H_B : \beta_1 = \beta_2 = \cdots = \beta_4$

	(1)	(2)	(3)	(4)	(5)	(6)	Total
1	24.7	27.7	20.6	16.2	16.2	24.9	130.3
2	27.3	28.8	22.9	15	17	22.5	133.5
3	38.5	39.5	36.8	19.6	15.4	26.3	176.1
4	28.5	31	34.9	14.1	17.7	22.6	148.8
Total	119	127	115.2	64.9	66.3	96.3	588.7

$C.F = 14,440.32$

$TSS = 15789.89 - C.F = 1349.57$

$SST = 15360.6075 - C.F = 920.28$

$SSB = 14658.831 - '' = 218.51$

$SSE = 210.78$

Treatment	5	920.28	184.056	13.098
Block	3	218.51	72.836	5.1833
Error	15	210.78	14.052	

$F_{\alpha}(5,15) = 2.9$ $\quad F_T > F_{\alpha}(5,15)$

$F_{\alpha}(3,15) = 3.29$ $\quad F_B > F_{\alpha}(3,15)$

Reject H_T, H_B

2. An experiment was designed to study the performance of 4 different detergents for cleaning fuel injectors. The following cleanness readings were obtained with specially designed equipment for 12 tanks of gas distributed over 3 different models of engines.

Detergent	Engine 1	2	3
A	45	43	51
B	47	46	52
C	48	50	55
D	42	37	49

Looking on the detergents as treatments and the engines as blocks, obtain the two way analysis of variance table and test at 1% level of significance whether there are differences in the detergents or in the engine.

3. Three varieties of a crop are tested in a randomised block design with four replications, the layout being as given below: The yields are given in kilograms. Analyse for significance

C48	A51	B52	A49
A47	B49	C52	C51
B49	C53	A49	B50

Solution:

Rewriting the data such that the rows represent the blocks and the columns represent the varieties of the crop (as assumed in the discussion of analysis of variance for two factors of classification), we have the following table:

Crops

Blocks	*A*	*B*	*C*
1	47	49	48
2	51	49	53
3	49	52	52
4	49	50	51

We shift the origin to 50 and work out with the new values of x_{ij}.

Crops

Blocks	*A*	*B*	*C*	T_i	T^2_i/k	$\sum_j x_{ij}^2$
1	-3	-1	-2	-6	$36/3 = 12$	14
2	1	-1	3	3	$9/3 = 3$	11
3	-1	2	2	3	$9/3 = 3$	9
4	-1	0	1	0	$0/3 = 0$	2
T_j	-4	0	4	$T = 0$	$\sum \frac{T_i^2}{k} = 18$	36
T^2_j/h	$\frac{16}{4} = 4$	$\frac{0}{4} = 0$	$\frac{16}{4} = 4$	$\sum \frac{T_j^2}{h} = 8$		
$\sum_i x_{ij}^2$	12	6	18	36		

$$Q = \sum\sum x_{ij}^2 - \frac{T^2}{N} = 36 - \frac{0^2}{12} = 36$$

$$Q_1 = \frac{1}{k}\sum T_i^2 - \frac{T^2}{N} = 18 - 0 = 18$$

$$Q_2 = \frac{1}{h}\sum T_j^2 - \frac{T^2}{N} = 8 - 0 = 8$$

$$Q_3 = Q - Q_1 - Q_2 = 36 - 18 - 8 = 10$$

ANOVA table

S.V.	S.S.	d.f.	M.S.	F_0
Between rows (blocks)	$Q_1 = 18$	$h-1=3$	6	$\frac{6}{1.67} = 3.6$
Between columns (crops)	$Q_2 = 8$	$k-1=2$	4	$\frac{4}{1.67} = 2.4$
Residual	$Q_3 = 10$	$(h-1)(k-1) = 6$	1.67	–
Total	$Q = 36$	$hk-1 = 11$	–	–

From F-tables, $F_{5\%}(v_1 = 3, v_2 = 6) = 4.76$ and $F_{5\%}(v_1 = 2, v_2 = 6) = 5.14$

Considering the difference between rows, we see that $F_0 (= 3.6) < F_{5\%} (= 4.76)$

Hence the difference between the rows is not significant. (H_0 is accepted) viz., the blocks do not differ significantly with respect to the yield.

Considering the difference between columns, we see that $F_0 (= 2.4) < F_{5\%} (= 5.14)$

Hence the difference between the columns is not significant. (H_0 is accepted) viz., the varieties of crop do not differ significantly with respect to the yield.

4. A tea company appoints four salesmen A, B, C and D and observes their sales in three seasons – summer, winter and monsoon. The figures (in lakhs) are given in the following table.

Season	Salesman A	B	C	D	Total
Summer	36	36	21	25	128
Winter	28	29	31	32	120
Monsoon	26	28	29	29	112
Total	90	93	81	96	$\boxed{360}$

(a) Do the salesmen significantly differ in performance?

(b) Is there a significant difference between the salesmans

Ans: $N = 12$

We shift the origin to 30 by subtracting to each.

Seasons \ Salesman	A	B	C	D	T_i	T_i^2/k	$\sum_j x_{ij}^2$
Summer	6	6	-9	5	8	$\frac{64}{4} = 16$	178
Winter	-2	-1	1	2	0	$0/4 = 0$	10
Monsoon	-4	-2	-1	-1	-8	$\frac{64}{4} = 16$	22
T_j	0	3	-9	6	$T = 0$	$\frac{1}{k}\sum T_i^2 = 32$	210
T_j^2/h	$\frac{0}{3} = 0$	3	27	12	$\frac{\sum T_j^2}{h} = \boxed{42}$		
$\sum_i x_{ij}^2$	56	41	83	30			$\sum\sum x_{ij}^2 = 210$

$$\therefore Q = \sum\sum x_{ij}^2 - \frac{T^2}{N} = 210 - \frac{0^2}{12} = 210$$

$$Q_1 = \frac{1}{k}\sum T_i^2 - \frac{T^2}{N} = 32 - \frac{0^2}{12} = 32$$

$$Q_2 = \frac{1}{h}\sum T_j^2 - \frac{T^2}{N} = 42 - 0 = 42$$

$$Q_3 = Q - Q_1 - Q_2 = 210 - 32 - 42$$
$$= 136$$

S.V.	S.S.	d.f	m.s.	F
Between rows (seasons)	$Q_1 = 32$	$h-1 = 2$	16	$F = 1.4125$
Between cols (salesman)	$Q_2 = 42$	$k-1 = 3$	14	$F = 1.6142$
Residual	$Q_3 = 136$	$(h-1)(k-1) = 6$	$\frac{136}{6} = 22.\bar{6}$	
Total	210	$hk-1 = 11$		

Consider the seasons.

$F_{cal} = 1.4125$ $\Rightarrow F_{cal} < F_{tab}$

$F_{tab} = F_{5\%}(6, 2) = 19.33$

$\Rightarrow H_0$ accepted.

There is no difference in performance

Consider the salemen:

$F_{cal} = 1.6142$ $\Rightarrow F_{cal} < F_{tab}$

$F_{tab} = F_{5\%}(6, 3) = 8.94$ $\Rightarrow H_0$ is accepted

There is no significant difference with in Salesmen.

Exercise:

The following data represent the number of units of production per day turned out by 5 different workers using 4 different types of machines:

Machine Type

		A	*B*	*C*	*D*
	1	44	38	47	36
	2	46	40	52	43
Workers:	3	34	36	44	32
	4	43	38	46	33
	5	38	42	49	39

(a) Test whether the five men differ with respect to mean productivity.

(b) Test whether the mean productivity is the same for the four different machine types.

Latin Square design

It is used to eliminate the effects of 2 extraneous sources of variability. An nxn Latin square is a square array of n distinct letters, with each letter appearing once and only once in each row and each column.

Layout.

Suppose there are four treatments A, B, C, D each applied once in each row and each column.

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

4 x 4 Latin Square

Advantages.

1. Latin square design controls more of the variation than RBD.
2. Analysis is simple.
3. Even with missing data the analysis remains relatively simple.

Dis.adv.. It cannotbe applied for all experiments

Model equation.

$$y_{ijk} = \mu + \alpha_i + \beta_j + \tau_k + \varepsilon_{ijk} \quad , \quad i,j,k = 1,2,\dots,m$$

y_{ijk} - yield obtained from the i^{th} row and j^{th} column by applying k^{th} treatment.

μ - grand mean

α_i - effect of the i^{th} row

β_j - " " " j^{th} column

τ_k - " " " k^{th} treatment

ε_{ijk} - indep. normally distributed random variables with zero means and common variance σ^2.

$H_R : \alpha_i = 0 \ \forall i \qquad H_{R_1} :$ Atleast one $\alpha_i \neq 0$

$H_C : \beta_j = 0 \ \forall j$

$H_T : \gamma_k = 0 \ \forall k$.

ANOVA Table

Source of variation	Degrees of freedom	S.S	MS	F
Row	$m-1$	SSR	$MSR = \frac{SSR}{m-1}$	$F_R = MSR/MSE$
Column	$m-1$	SSC	$MSC = \frac{SSC}{m-1}$	$F_C = MSC/MSE$
Treatment	$m-1$	SST	$MST = \frac{SST}{m-1}$	$F_T : MST/MSE$
Error	$(m-1)(m-2)$	SSE	$MSE = \frac{SSE}{(m-1)(m-2)}$	
Total	m^2-1	TSS		

$C.F = \frac{G^2}{m^2}$, m = No. of rows = No. of columns = No. of treatments

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

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

$$SST = \frac{\sum T_k^2}{m} - C.F$$

$$TSS = \sum\sum y_{ij}^2 - C.F$$

$$SSE = TSS - (SSR + SSC + SST)$$

If $F_R < F_\alpha(m-1, (m-1)(m-2))$, accept H_R

" F_C " " " H_C

If F_T " " " H_T.

1. Set up the ANOVA for the following results of a Latin square design

A 12	C 19	B 10	D 8	49
C 18	B 12	D 6	A 7	43
B 22	D 10	A 5	C 21	58
D 12	A 7	C 27	B 17	63
64	48	48	53	213

SV	Df	SS	MSS	F
R	3	60.1875	20.0625	1.5165
C	3	42.6875	14.2292	1.0756
T	3	465.1875	155.0625	11.7212
E	6	79.375	13.2292	

$F_{0.05}(3,6) = 4.76$

Accept H_R, H_C, Reject H_T

Five doctors each test five treatments for a certain disease and observe the no. of days each patient takes to recover. Discuss the dif between i) the doctors ii) the treatments for the fol. data

Da.	Treat 1	2	3	4	5
1	10	14	23	18	20
2	11	15	24	17	21
3	9	12	20	16	19
4	8	13	17	17	20
5	12	15	19	15	22

Treat.	4	406.64	101.66	47.06
Doc.	4	25.84	6.46	2.99
Error	16	34.56	2.16	
		467.04		

$F_\alpha(4, 16) = 3.01$

2. The sample data in the following Latin Square are the scores obtained by 9 college students of various ethnic backgrounds and various professional interests in an American history test. A, B, C are the three instructors by whom the 9 college students were taught the course in American history. Use 5% level of significance to analyze the design and test the following hypotheses. whether differences in
i) the ethnic background have no effect on the scores and
ii) professional interests have no effect on the scores.

Ethnic background

	Mexican	German	Polish
Law	A : 75	B : 86	C : 69
Medicine	B : 95	C : 79	A : 86
Engineering	C : 70	A : 83	B : 93

SV	Df	SS	MSS	F	
Row	2	150.23	75.11	27.11	2.407
Column	2	14.23	7.11	2.567	1.478
Treatment	2	523.56	261.78	94.505	2.222
Error	2	5.54	2.77		
Total	8	693.56			

$F_{0.05}^{(2,2)} = 19$

Accept H_C

Reject H_R, H_T