Goodness of fit test (χ^2-test)

Test for testing the significance of discrepancy between theory and experiment was given by Prof Karl Pearson in 1900 and is known as "Chi square test of goodness of fit".

Decision rule: Accept H_0 if $\chi^2 \leq \chi^2_\alpha(n-1)$ and reject H_0 if $\chi^2 > \chi^2_\alpha(n-1)$, where χ^2 is the calculated value of chi-square and $\chi^2_\alpha(n-1)$ is the tabulated value of chi-square for $(n-1)$ d.f and level of significance α.

(1) 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	Thur	Fri	Sat
No of parts demanded:	1124	1125	1110	1120	1126	1115

Test the hypothesis, that the number of parts demanded does not depend on the day of the week. (Given: the value of chi-square significance at 5, 6, 7, d.f are respectively 11.07, 12.59, 14.07 at the 5% level of significance)

Sol: Here we set up the null hypothesis

H_0: The number of parts demanded does not depend on the day of the week.

Under the null hypothesis, the expected frequencies of the spare part demanded on each of the six days would be

$$\frac{1}{6}(1124 + 1125 + 1110 + 1120 + 1126 + 1115) = \frac{6720}{6} = 1120$$

Calculation for χ^2

Days	observed (f_i)	expected (f_i)	$(f_i - e_i)^2$	$\frac{(f_i - e_i)^2}{e_i}$
Mon	1124	1120	16	0.014
Tues	1125	1120	25	0.022
Wed	1110	1120	100	0.089
Thurs	1120	1120	0	0
Fri	1126	1120	36	0.032
Sat	1115	1120	25	0.022
Total	6720	6720		0.179

$$\chi^2 = \sum \frac{(f_i - e_i)^2}{e_i} = 0.179$$

deg of freedom $= 6 - 1 = 5$

$\chi^2_{0.05}$ for 5 d.f $= 11.07$

Calculated χ^2 < Tabulated χ^2

Thus null hypothesis may be accepted at 5% level of significance. Hence we conclude that the number of parts demanded are same over the 6-day period.

Q(2) The following figure show the distribution of digits in numbers chosen at random from a telephone directory.

Digit :	0	1	2	3	4	5	6	7	8	9	Total
Frequency :	1026	1107	997	966	1075	933	1107	972	964	853	10000

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

Sol: Null hypothesis: The digits occur equally frequently on the directory.

Under the null hypothesis, the expected frequency for each of the digits $0, 1, 2, \ldots 9$ is $\frac{10000}{10} = 1000$

The value of χ^2 is calculated as follows.

Digits	Observed freq (f_i)	Exp freq (e_i)	$(f_i - e_i)$	$\frac{(f_i - e_i)^2}{e_i}$
0	1026	1000	676	0.676
1	1107	1000	11449	11.449
2	997	1000	9	0.009
3	966	1000	1156	1.156
4	1075	1000	5625	5.625
5	933	1000	4489	4.489
6	1107	1000	11149	11.449
7	972	1000	784	0.784
8	964	1000	1296	1.296
9	853	1000	21609	21.609
				58.542

$$\chi^2 = \sum \frac{(f_i - e_i)^2}{e_i} = 58.542$$

$\chi^2_{0.05}$ for 9 d.f $= 16.919$

Calculated $\chi^2 >$ Tab χ^2

So null hypothesis is rejected

Thus we conclude that the digits do not occur uniformly.

Example 15·13. *A sample analysis of examination results of 200 MBA's was made. It was found that 46 students had failed, 68 secured a third division, 62 secured a second division and the rest were placed in first division. Are these figures commensurate with the general examination result which is in the ratio of 4 : 3 : 2 : 1 for various categories respectively?*

Solution. Set up the *null hypothesis* that the observed figures do not differ significantly from the hypothetical frequencies which are in the ratio of 4 : 3 : 2 : 1. In other words *the given data are commensurate with the general examination result* which is in the ratio of 4 : 3 : 2 : 1 for the various categories.

Under the null hypothesis, the expected frequencies can be computed as shown in the adjoining table :

Category	*Frequency*	
	Observed (f_i)	*Expected* (e_i)
Failed	46	$\frac{4}{10} \times 200 = 80$
III Division	68	$\frac{3}{10} \times 200 = 60$
II Division	62	$\frac{2}{10} \times 200 = 40$
I Division	24	$\frac{1}{10} \times 200 = 20$
Total	200	200

TABLE 15·4 : CALCULATIONS FOR χ^2

Category	Frequency		$(f_i - e_i)^2$	$\frac{(f_i - e_i)^2}{e_i}$
	Observed (f_i)	Expected (e_i)		
Failed	46	80	1156	14·450
III Division	68	60	64	1·067
II Division	62	40	484	12·100
I Division	24	20	16	0·800
Total	200	200		28·417

$\chi^2 = \sum \frac{(f_i - e_i)^2}{e_i} = 28·417$

$d.f. = 4 - 1 = 3$, tabulated $\chi^2_{0·05}$ for 3 $d.f.$ = 7·815

Since the calculated value of χ^2 is greater than the tabulated value, it is significant and the null hypothesis is rejected at 5% level of significance. Hence we may conclude that data are not commensurate with the general examination result.

Example 15·14. *A survey of 800 families with four children each revealed the following distribution :*

No. of boys	:	0	1	2	3	4
No. of girls	:	4	3	2	1	0
No. of families	:	32	178	290	236	64

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

Solution. Let us set up the *null hypothesis* that *the data are consistent with the hypothesis of equal probability for male and female births.* Then under the null hypothesis :

$$p = \text{Probability of male birth} = \frac{1}{2} = q$$

$$p(r) = \text{Probability of } 'r' \text{ male births in a family of } 4 = {}^4C_r \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^{4-r} = {}^4C_r \left(\frac{1}{2}\right)^4$$

The frequency of r male births is given by :

$$f(r) = N.\ p(r) = 800 \times {}^4C_r \left(\frac{1}{2}\right)^4 = 50 \times {}^4C_r ;\ r = 0, 1, 2, 3, 4. \quad \ldots (*)$$

Substituting $r = 0, 1, 2, 3, 4$ successively in (*), we get the expected frequencies as follows :

$f(0) = 50 \times 1 = 50,$ $f(1) = 50 \times {}^4C_1 = 200,$ $f(2) = 50 \times {}^4C_2 = 300,$

$f(3) = 50 \times {}^4C_3 = 200,$ $f(4) = 50 \times {}^4C_4 = 50.$

Test of Independence of attributes.

Ex:- Two samples polls of votes for two candidates A and B for two candidates A and B for a public office are taken, One from among the residents of rural area. The results are given in the table below Examine whether the nature of the area is related to voting preference in this election.

Area	Votes for		Total
	A	B	
Rural	620	380	1000
Urban	550	450	1000
Total	1170	830	2000

[Discussion: Here we see that there are two attributes a voter is either from rural or urban and a voter is voting for A or B. We need to check whether there is any difference in voting pattern for rural and urban population]

Sol: Null hypothesis: The nature of area is independent of voting preference in the election

We get the expected frequencies as follow

$$E(620) = \frac{1170 \times 1000}{2000} = 585, \quad E(380) = \frac{830 \times 1000}{2000} = 415$$

$$E(550) = \frac{1170 \times 1000}{2000} = 585, \quad E(450) = \frac{830 \times 1000}{2000} = 415$$

$$\chi^2 = \sum \frac{(f_i - e_i)^2}{e_i}$$

$$= \frac{(620-585)^2}{585} + \frac{(380-415)^2}{415} + \frac{(550-585)^2}{585} + \frac{(450-415)^2}{415} = 10.0891$$

Degree ab freedom $= (2-1)(2-1) = 1$ [as the table is 2×2]

$\chi^2_{0.05}$ - for 1 d.f = 3.841 (given from table)

Cal χ^2 > Tab χ^2

So Null hypothesis is rejected at 5% level ab scgnebicance.

Thus we conclude that nature ab area is related to voting pattern.