13·7·2. Chi-square Test of Goodness of Fit. A very powerful test for testing the significance of the discrepancy between theory and experiment was given by Prof. Karl Pearson in 1900 and is known as "*Chi-square test of goodness of fit.*" It enables us to find if the deviation of the experiment from theory is just by chance or is it really due to the inadequacy of the theory to fit the observed data.

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 Karl Pearson's chi-square, given by

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{(O_i - E_i)^2}{E_i} \right], \qquad \left(\sum_{i=1}^{n} O_i = \sum_{i=1}^{n} E_i \right) \qquad \ldots(13{\cdot}15)$$

follows chi-square distribution with $(n - 1)$ d.f.

2. Conditions for the Validity of χ^2-test. χ^2-test is an approximate test for large values of n. For the validity of chi-square test of '*goodness of fit*' between theory and experiment, the following conditions must be satisfied :

(*i*) The sample observations should be independent.

(*ii*) Constraints on the cell frequencies, if any, should be linear, *e.g.*, $\sum n_i = \sum \lambda_i$ or $\sum O_i = \sum E_i$.

(*iii*) N, the total frequency should be reasonably large, say, greater than 50.

(*iv*) No theoretical cell frequency should be less than 5. (The chi square distribution is essentially a continuous distribution but it cannot maintain its character of continuity if cell frequency is less than 5). If any theoretical cell frequency is less than 5, then for the application of χ^2-test, it is pooled with the preceding or succeeding frequency so that the pooled frequency is more than 5 and finally adjust for the d.f. lost in pooling.

3. It may be noted that the χ^2-test depends only on the set of observed and expected frequencies and on degrees of freedom (*d.f.*). It does not make any assumptions regarding the parent population from which the observations are taken. Since χ^2 defined in (13·8) does not involve any population parameters, it is termed as a statistic and the test is known as *Non-Parametric Test* or *Distribution-Free Test*.

Example 13·11. *The following figures show the distribution of digits in numbers chosen at random from a telephone directory :*

Digits :	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*Total*
Frequency :	*1026*	*1107*	*997*	*966*	*1075*	*933*	*1107*	*972*	*964*	*853*	*10,000*

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

Solution. Here we set up the *null hypothesis* that the digits occur equally frequently in the directory.

Under the null hypothesis, the expected frequency for each of the digits 0, 1,2, ..., 9 is 10000/10 = 1000. The value of χ^2 is computed as follows :

CALCULATIONS FOR χ^2

Digits	*Observed Frequency* (O)	*Expected Frequency* (E)	$(O-E)^2$	$(O-E)^2/E$
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	11449	11·449
7	972	1000	784	0·784
8	964	1000	1296	1·296
9	853	1000	21609	21·609
Total	10,000	10,000		58·542

$$\therefore \qquad \chi^2 = \sum\left[\frac{(O-E)^2}{E}\right] = 58{\cdot}542$$

The number of degrees of freedom = 10 – 1 = 9, (since we are given 10 frequencies subjected to only one linear constraint $\sum O = \sum E = 10{,}000$).

The tabulated $\chi^2_{0{\cdot}05}$ for 9 d.f. = 16·919

Since the calculated χ^2 is much greater than the tabulated value, it is highly significant and we reject the null hypothesis. Thus we conclude that the digits are not uniformly distributed in the directory.

Example 13·12. *The following table gives the number of aircraft accidents that occurs during the various days of the week. Find whether the accidents are uniformly distributed over the week.*

Days	...	*Sun.*	*Mon.*	*Tues.*	*Wed.*	*Thus.*	*Fri.*	*Sat.*
No. of accidents	...	14	16	8	12	11	9	14

(Given : the values of chi-square significant at 5, 6, 7, d.f. are respectively 11·07, 12·59, 14·07 at the 5% level of significance.

Solution. Here we set up the *null hypothesis that the accidents are uniformly distributed over the week.*

Under the null hypothesis, the expected frequencies of the accidents on each of the days would be :

Days ...	Sun.	Mon.	Tues.	Wed.	Thus.	Fri.	Sat.	Total
No. of accidents ...	12	12	12	12	12	12	12	84

$$\chi^2 = \frac{(14-12)^2}{12} + \frac{(16-12)^2}{12} + \frac{(8-12)^2}{12} + \frac{(12-12)^2}{12} + \frac{(11-12)^2}{12} + \frac{(9-12)^2}{12} + \frac{(14-12)^2}{12}$$

$$= \frac{1}{12}(4+16+16+0+1+9+4) = \frac{50}{12}$$

$$= 4{\cdot}17$$

The number of degrees of freedom

= Number of observations – Number of independent constraints.

$= 7 - 1 = 6$

The tabulated $\chi^2_{0{\cdot}05}$ for 6 d.f. = 12·59

Since the calculated χ^2 is much less than the tabulated value, it is highly insignificant and we accept the null hypothesis. Hence we conclude that the accidents are uniformly distributed over the week.

Chi-Square (χ^2) Distribution

Area to the Right of Critical Value

Degrees of Freedom	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Example 13·14. *A survey of 320 families with 5 children each revealed the following distribution :*

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 ?

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 :

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

$$p(r) = \text{Probability of 'r' male births in a family of 5}$$

$$= \binom{5}{r} p^r q^{5-r} = \binom{5}{r} \left(\frac{1}{2}\right)^5$$

The frequency of r male births is given by :

$$f(r) = N.\ p(r) = 320 \times \binom{5}{r} \times \left(\frac{1}{2}\right)^5$$

$$= 10 \times \binom{5}{r} \qquad \ldots(*)$$

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

$$f(0) = 10 \times 1 = 10, \qquad f(1) = 10 \times {}^5C_1 = 50$$

$$f(2) = 10 \times {}^5C_2 = 100, \qquad f(3) = 10 \times {}^5C_3 = 100$$

$$f(4) = 10 \times {}^5C_4 = 50, \qquad f(5) = 10 \times {}^5C_5 = 10$$

CALCULATIONS FOR χ^2

Observed Frequencies (O)	*Expected Frequencies* (E)	$(O-E)^2$	$(O-E)^2/E$
14	10	16	1·6000
56	50	36	0·7200
110	100	100	1·0000
88	100	144	1·4400
40	50	100	2·0000
12	10	4	0·4000
Total 320	320		7·1600

$$\therefore \qquad \chi^2 = \sum\left[\frac{(O-E)^2}{E}\right] = 7{\cdot}16$$

Tabulated $\chi^2_{0.05}$ for $6-1=5$ d.f. is 11·07.

Calculated value of χ^2 is less than the tabulated value, it is not significant at 5% level of significance and hence the null hypothesis of equal probability for male and female births may be accepted.

Independence of Attributes.

Example 13·6. *Two sample polls of votes for two candidates A and B for a public office are taken, one from among the residents of rural areas. The results are given in the table. Examine whether the nature of the area is related to voting preference in this election.*

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

Solution. Under the *null hypothesis* that the nature of the area is independent of the voting preference in the election, we get the observed frequencies as follows :

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

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

Aliter. In a 2×2 contingency table, since

$$\text{d.f.} = (2-1)(2-1) = 1,$$

only one of the cell frequencies can be filled up independently and the remaining will follow immediately, since the observed and theoretical marginal totals are fixed. Thus having obtained any one of the theoretical frequencies, (say), $E(620) = 585$, the remaining theoretical frequencies can be easily obtained as follows :

$$E(380) = 1000 - 585 = 415, \; E(550) = 1170 - 585 = 585.$$

and
$$E(450) = 1000 - 585 = 415$$

$$\therefore \qquad \chi^2 = \sum\left[\frac{(O-E)^2}{E}\right] = \frac{(620-585)^2}{585} + \frac{(380-415)^2}{415} + \frac{(550-585)^2}{585} + \frac{(450-415)^2}{415}$$

$$= (35)^2\left[\frac{1}{585} + \frac{1}{415} + \frac{1}{585} + \frac{1}{415}\right]$$

$$= (1225)[2 \times 0{\cdot}002409 + 2 \times 0{\cdot}001709] \doteq 10{\cdot}0891$$

Tabulated $\chi^2_{0\cdot05}$ for $(2-1)(2-1) = 1$ d.f. is $3{\cdot}841$. Since calculated χ^2 is much greater than the tabulated value, it is highly significant and null hypothesis is rejected at 5% level of significance. Thus we conclude that nature of area is related to voting preference in the election.

- Q) Out of 8000 graduates in a town 800 are females, out of 1600 graduate employee 120 are females. Use chi-square to determine if any distinction is made in appointment on the basis of gender. Value of chi square at 5 % level for one degree of freedom is 3.84.

Null Hypothesis: There is no distinction in appointment on the basis of gender.

Alternate hypothesis: the distinction is made on the basis of gender.

TABLE NO. OBSERVED FREQUENCIES — EXPECTED FREQUENCIES

	Employed	*Not employed*	*Total*	*Employed*	*Not employed*	*Total*
Male	1480	5720	7200	$\frac{7200 \times 1600}{8000} = 1440$	$7200 - 1440 = 5760$	7200
Female	120	680	800	$1600 - 1440 = 160$	$6400 - 5760 = 640$	800
Total	1600	6400	8000	1600	6400	8000

TABLE 15·8 : CALCULATIONS FOR χ^2

Class	*Frequency* Observed (f_i)	Expected (e_i)	$(f_i - e_i)$	$\frac{(f_i - e_i)^2}{e_i}$
Male employed	1480	1440	40	$\frac{1600}{1440} = 1{\cdot}11$
Male unemployed	5720	5760	– 40	$\frac{1600}{5760} = 0{\cdot}28$
Female employed	120	160	– 40	$\frac{1600}{160} = 10{\cdot}00$
Female unemployed	680	640	40	$\frac{1600}{640} = 2{\cdot}50$

$$\chi^2 = \sum \frac{(f_i - e_i)^2}{e_i} = 13{\cdot}89$$

$$d.f. = (2-1)(2-2) = 1$$

Tabulated $\chi^2_{0{\cdot}05}$ for 1 $d.f.$ = 3·841.