Chapter 6:

Chi square tests

October 30, 2021

A Chi-square test is a hypothesis testing method. Chi square test statistic is used for comparing between the observed and the expected values in order to answer questions or to test some claims concerning the experiment in two types of Chi square tests for

1. Goodness of fit test with a probability distribution.
2. Independence and Homogeneity between two variables.

For Goodness of fit:
We consider a test to determine if a population has a specified theoretical distribution. The test is based on how good a fit we have between the frequency of occurrence of observations in an observed sample and the expected frequencies obtained from the hypothesized distribution.
we test whether the null hypothesis

H_0 : The given data is well fitted a certain Probability Distribution,

H_1 : The data is not fitted.

1. The data range is divided into k intervals (or cells or classes) and the frequency count O_i for each interval (cell, or class) is tabulated.
2. The expected frequency E_i based on the theoretical pdf f with cumulative distribution F is calculated for each interval (cell, or class)

$$E_i = \quad np_i,$$

where n is is the number of data points, and $\sum_i^k O_i = \sum_i^k E_i = n$.

$p_i = F(a_i) - F(a_{i-1})$, where a_i and a_{i-1} are the endpoint of the interval,

or

$p_i = f(a_i)$ if the interval is a single point

3. Under the fact that $E_i \geq 5, i = 1, \ldots, k$, The test statistic is $\chi_0^2 = \sum_{i=1}^k \frac{\left(O_i - E_i\right)^2}{E_i}$.

The test statistic (χ_0^2) is approximated to the chi-square distribution with $\nu = k - m - 1$ degrees of freedom, where m is the number of unknown parameters of the theoretical probability distribution estimated from sample data.

- If $E_i < 5$, E_i is combined to a neighbor level.
- The null hypothesis is rejected if the test statistic $\chi_0^2 > \chi_{\alpha,k-m-1}^2$.
  $\chi_{\alpha,\nu}^2$ is the value of Chi-Square distribution for significance level α with $df = \nu$ in the Chi-Square distribution table.

Example 1:

We consider the tossing of a die. We hypothesize that the die is honest, which is equivalent to testing the hypothesis that the distribution of outcomes is the discrete uniform distribution

$$f(x) = \frac{1}{6}, x = 1, 2, \ldots, 6.$$

Suppose that the die is tossed 120 times and each outcome is recorded. Theoretically, if the die is balanced, we would expect each face to occur 20 times. The results are given by

Face	1	2	3	4	5	6
Observed	20	22	17	18	19	24
Expected	20	20	20	20	20	20

By comparing the observed frequencies with the corresponding expected frequencies, the hypotheses,

H_0 : the die is fair $(f(x) = \frac{1}{6}, x = 1, \ldots, 6)$,

H_1 die is not ($f(x)$ is not discrete uniform distribution).

Solution:
We find, from the table, that the χ_0^2- test statistic value to be

$$\chi_0^2 = \frac{(20-20)^2}{20} + \frac{(22-20)^2}{20} + \frac{(17-20)^2}{20} + \frac{(18-20)^2}{20} + \frac{(19-20)^2}{20} + \frac{(24-20)^2}{20} = 1.7$$

Using chi-squared table, we $\chi^2_{0.05,5} = 11.070$. Since 1.7 is less than the critical value ($\chi^2_{0.05,5}$), we fail to reject H_0. We conclude the die is fair.

Example 2:
In a study for the quality control for a certain machine to be used for paving the public roads, a researcher to select a sample of 4 machines within a period 200 days for the production line and to record number of machines needed maintains in each day. He obtained the following records:

No. of machines needed maintains (x_i)	0	1	2	3	4
No. of days to be repeated (O_i)	101	79	19	1	0

Does this data suggest that No. of machines that need maintains follows Bin(4,p) at level 0.01? (Hint; The parameter p should be estimated.)

Solution:

The estimation of the parameter p is:

$$\text{Given } \mu = np \implies \hat{\mu} = 4\hat{p}$$

$$\hat{\mu} = \frac{\sum_0^4 (O_i \times x_i)}{200} = \frac{(0\times 101 + 1\times 79 + 2\times 19 + 3\times 1 + 4\times 0)}{200} = 0.6$$

$$\text{So } \hat{p} = \frac{0.6}{4} = 0.15$$

The hypothesis:

H_0 : No. of machines that need maintains follows Bin (4, 0.15),

H_1 : No. of machines that need maintains does not follow Bin (4, 0.15).

Under H_0, the expected numbers are:

$$E_i = np_i = 200 \times \begin{pmatrix} 4 \\ i \end{pmatrix} (0.15)^i (0.85)^{4-i}$$

$$
\begin{aligned}
E_0 &= 200p_0 = 200 \times 0.5220 = 104.4 \\
E_1 &= 200p_1 = 200 \times 0.3685 = 73.7 \\
E_2 &= 200p_2 = 200 \times 0.0975 = 19.5 \\
E_3 &= 200p_3 = 200 \times 0.00115 = 2.3 \\
E_4 &= 200p_4 = 200 \times 0.0005 = 0.1
\end{aligned}
$$

We may summarize our calculation as in the following table:

No. of machines needed maintains (x_i)	0	1	2	3	4
No. of days to be repeated (O_i)	101	79	19	1	0
corresponding expected values (E_i)	104.4	73.7	19.5	2.3	0.1

We note that the expected values $E_3 = 2.3$ and $E_4 = 0.1$ correspond to $i = 3, 4$ are less than 5, so we cancel these two categories and (in the time) to add both the observed and expected values to that category before, and therefore we have now the following table to be used for our calculated

No. of machines needed maintains (x_i)	0	1	≥ 2
No. of days to be repeated (O_i)	101	79	20
The corresponding expected values (E_i)	104.4	73.7	21.9

The number of observed values became $k = 3$, and the test statistic is:

$$\chi_0^2 = \sum_{j=1}^{3} \frac{(O_j - E_j)^2}{E_j} = \frac{(101-104.4)^2}{104.4} + \frac{(79-73.7)^2}{73.7} + \frac{(20-21.9)^2}{21.9} = 0.657$$

The df of χ_0^2 is: $\nu = k–m–1 = 3 - 1 - 1 = 1$, and m=1 as we estimated one parameter. The rejection region (R.R) is: $(\chi^2_{1,0.01}, \infty) = (6.63, \infty)$, so H_0 is accepted at the level 0.01, that is; the No. of machines that need maintains follows Bin(4, 0.15).
Remark: if p is not estimated but is given in advance, then m=0 and $\nu = 2$.

Example 3:

A third example is to test the hypothesis that the frequency distribution of battery lives given in this Table

class	Class Boundaries	Frequency
1	1.45-1.95	2
2	1.95-2.45	1
3	2.45-2.95	4
4	2.95-3.45	15
5	3.45-3.95	10
6	3.95-4.45	5
7	4.45-4.95	3

The frequency may be approximated by a normal distribution with mean $\mu = 3.5$ and standard deviation $\sigma = 0.7$. Test the hypothesis

$$H_0: \quad \text{the data is normal } N(3.5, 0.7),$$
$$H_1: \quad \text{the data is not normal.}$$

at level of significance $\alpha = 0.05$

Solution:

Under H_0, The z-values corresponding to the boundaries of the first class are $z_1 = (1.45 - 3.5)/0.7 = -2.93$ and $z_2 = (1.95 - 3.5)/0.7 = -2.21$.

From standard normal table, we find the area between $z_1 = -2.93$ and $z_2 = -2.21$ to be area equal to

$$\begin{aligned} p_1 &= P(-2.93 < Z < -2.21) \\ &= P(Z < -2.21) - P(Z < -2.93) \\ &= F(-2.21) - F(-2.93) \\ &= 0.0119 \end{aligned}$$

.

Hence, the expected frequency for the first class is

$$E_1 = np_1 = 40(0.0119) = 0.5$$

Similarly, we get

$$E_2 = np_2 = 40[P(-2.21 < Z < -1.5)] = 2.1$$

$$
\begin{aligned}
E_3 &= 40[P(-1.5 < Z < -0.79)] = 5.9 \\
E_4 &= 40[P(-0.79 < Z < -0.07)] = 10.3 \\
E_5 &= 40[P(-0.07 < Z < 0.64)] = 10.7 \\
E_6 &= 40[P(0.64 < Z < 1.36)] = 7.0 \\
E_7 &= 40[P(1.36 < Z < 2.07)] = 3.5
\end{aligned}
$$

We note that the expected values $E_1 = 0.5, E_2 = 2.1$ and $E_7 = 3.5$ correspond to class 1, 2 and 7 are less than 5, so we cancel these three classes and (in the time) to add both class 1 and 2 the observed and expected values to that class 3 and add class 7 to class 6, and therefore we have now the following table to be used for our calculated:

class	1	2	3	4
class boundary	1.45-2.95	2.95-3.45	3.45-3.95	3.95- 4.95
observed (O)	7	15	10	8
Expected (E)	8.5	10.3	10.7	10.5

The number of observed values is reduced to $k = 4$, and the test statistic is

$$\chi_0^2 = \frac{(7-8.5)^2}{8.5} + \frac{(15-10.3)^2}{10.3} + \frac{(10-10.7)^2}{10.7} + \frac{(8-10.5)^2}{10.5} = 3.05$$

Since the computed χ_0^2-value is less than $\chi^2_{0.05,\nu=k-m-1} = \chi^2_{0.05,4-0-1} = \chi^2_{0.05,3} = 7.815, m = 0$ (the μ, σ are given and are not estimated), we have no reason to reject the null hypothesis and conclude that the normal distribution with mean $\mu = 3.5$ and standard deviation $\sigma = 0.7$ provides a good fit for the distribution of battery lives.

Example 4:
The frequency distribution table of 120 women weights is:

Class interval	42.5-47.5	47.5-52.5	52.5-57.5	57.5-62.5	62.5-67.5	67.5-72.5	72.5-77.5
Frequencies (O_i)	5	16	28	32	23	11	5
Mid Class (x_i)	45	50	55	60	65	70	75

Does this data suggest that the women weight is Normally distributed at level 0.05?

Solution:

The mean μ and the variance σ^2 of the weight population are estimated by $\bar{X} = \sum_{i=1}^{k} \frac{O_i \times x_i}{n} = \sum_{i=1}^{7} \frac{O_i \times x_i}{120} = 59.4$ and $S^2 = 51.9$.

The hypothesis:

H_0 : Women weight follows N(59.4,51.9),

H_1 : Women weight does not follow N(59.4,51.9).

Under that H_0 is correct, the expected number in each class interval (a_i, b_i) is:

$$
\begin{aligned}
E_i = \quad & n \times p_i, \\
& p_i = P(a_i \leq X \leq b_i) = \big(F(b_i) - F(a_i)\big) \\
E_1 = \quad & 120 \times P(42.5 \leq X \leq 47.5) = 4.76 \\
E_2 = \quad & 120 \times P(47.5 \leq X \leq 52.5) = 4.38 \\
E_3 = \quad & 120 \times P(52.5 \leq X \leq 57.5) = 27.24 \\
E_4 = \quad & 120 \times P(57.5 \leq X \leq 62.5) = 32.48 \\
E_5 = \quad & 120 \times P(62.5 \leq X \leq 67.5) = 24.37 \\
E_6 = \quad & 120 \times P(67.5 \leq X \leq 72.5) = 11.51 \\
E_7 = \quad & 120 \times P(72.5 \leq X \leq 77.5) = 3.42
\end{aligned}
$$

We may summarize our calculation as in the following table:

Class interval	42.5-47.5	47.5-52.5	52.5-57.5	57.5-62.5	62.5-67.5	67.5-72.5	72.5-77.5
Observed (O_i)	5	16	28	32	23	11	5
Expected (E_i)	4.76	14.38	27.24	32.48	24.37	11.51	3.42

We note that the expected values $E_1 = 4.76$ and $E_7 = 3.42$ be less than 5, so we cancel these intervals and (in the time) to add the observed and expected values to their neighbors, and we have now the next table to be used for our calculations:

Class interval	42.5-52.5	52.5-57.5	57.5-62.5	62.5-67.5	67.5-77.5
Observed (O_i)	21	28	32	23	16
Expected (E_i)	19.14	27.24	32.48	24.37	14.93

$$\chi_0^2 = \sum_{i=1}^{5} \frac{(O_i - E_i)^2}{E_i} = 0.363.$$

The degrees of freedom of χ_0^2 is:
$\nu = k - m - 1 = 5 - 2 - 1 = 2$, and $m = 2$ as we estimated 2 parameters. The R.R is: $(\chi^2_{0.05,2}, \infty) = (5.991, \infty)$, so H_0 is accepted at the level 0.05 and the data follows $N(59.4, 51.9)$.

For independence test:
If we select a random sample from a population, we may need to classify the units according to two (A and B) classification random variables "or phenomena". For instance, to classify a random sample of students according to;

- The weight and the advanced in the academic level.
- The ability in Mathematics and the ability in Statistics.
- The sex (male, female) and the advanced in the academic level.
- The smoking habit and the smoking habit of their parents.

We assume that (A) having $r \geq 2$ different levels (or categories), to be assigned in r rows and (B) having $c \geq 2$ different levels (or categories) to be assigned in c columns. The table of this classification is called "$r \times c$ association table" of the form:

Measure "A"	Measure "B"					
	1	2	3	j	C	Total
1st level A_1	O_{11}	O_{12}	O_{13}	O_{1j}	O_{1C}	$O_{1.}$
2nd level A_2	O_{21}	O_{22}	O_{23}	O_{2j}	O_{2C}	$O_{2.}$
.....	.....	.....	.....	.....	.....	.....
.. ..	O_{i1}	O_{i2}	O_{i3}	O_{ij}	O_{iC}	$O_{i.}$
....	.....	.....	.....	.....	.....	.....
r^{th} level A_r	O_{r1}	O_{r2}	O_{r3}	O_{rj}	O_{rC}	$O_{r.}$
Total	$O_{.1}$	$O_{.2}$	$O_{.3}$	$O_{.j}$	$O_{.C}$	$O_{..}$

Where
$O_{.j} = \sum_{i=1}^{r} O_{ij}, j = 1, \ldots, c, O_{i.} = \sum_{j=1}^{c} O_{ij}, i = 1, \ldots, r$ and
$O_{..} = \sum_{i=1}^{r} \sum_{j=1}^{c} O_{ij} = n.$

The aforesaid logic of the $r \times c$ association table indicates that sampling to be carried out before classification. Moreover, the determination of the totality for both the horizontal rows and for the vertical columns are depending only on chance. We aim behind this logic at testing the null hypothesis;

$$H_0 : \quad \text{A and B are independent,}$$
$$H_1 : \quad \text{A and B are not independent.}$$

For homogeneity:
If there are certain r populations, and we select from each of them independent samples. For instance, a researcher was inspecting devices in three different institutes due to a certain ministry and from each a sample is selected. Here, the total number in each row of the phenomena A is known in advance. Then, classification according to the phenomena B is completed in to c levels. So, the $r \times c$ association table indicates that classification to be carried before sampling. We aim behind this logic at testing Homogeneity between A and B, as following:

$$\begin{aligned} H_0 &: \text{ A and B are homogeneous,} \\ H_1 &: \text{ A and B are not homogeneous.} \end{aligned}$$

Under that H_0 is true (Independence or Homogeneity), the expected value E_{ij} is shown to be determined by:

$$\begin{aligned} E_{ij} &= n \times P(\text{getting } O_{ij}) \\ &= n \times \frac{O_{i.}}{O_{..}} \times \frac{O_{.j}}{O_{..}} \\ &= O_{..} \times \frac{O_{i.}}{O_{..}} \times \frac{O_{.j}}{O_{..}} \\ &= \frac{O_{i.} \times O_{.j}}{O_{..}} \end{aligned}$$

The test statistic:

$$\chi_0^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)\times(c-1)},$$

and R.R is $(\chi^2_{\alpha,(r-1)\times(c-1)}, \infty)$.

Example 5:
In a planned study to investigate the relationship between the weights (A) and the advanced in the academic level (B) for the student, in a sample of 120 students we got the following table:

	Excellent	Very Good	Pass	Fail	Total
Skinny	14	11	10	5	40
Average	10	16	16	14	56
Fat	3	4	7	10	24
Total	27	31	33	29	120

Does this data suggest that there is a relationship between weight (phenomena A) and the advanced in the academic level (phenomena B)? Use $\alpha = 0.05$.

Solution:
The null hypothesis is

H_0 : the two random variables (phenomena's, A and B) are independent
H_1 : (A) And (B) are not independent.

The test statistic is:

$$\chi_0^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi_\nu^2, \nu = (r-1) \times (c-1) = 2 \times 3 = 6$$

$$\text{and R.R} = (\chi^2_{0.05,6}, \infty) = (12.592, \infty)$$

Now, we add the expected values $E_{ij} = \frac{O_{i.} \times O_{.j}}{O_{..}}$ in the following table for our calculation:

	Excellent	Very Good	Pass	Fail	Total
Skinny	14 9	11 10.33	10 11	5 9.67	40
Average	10 12.6	16 14.47	16 15.4	14 13.53	56
Fat	3 5.4	4 6.2	7 6.6	10 5.8	24
Total	27	31	33	29	120

And therefore, the calculated test tool will be:

$$\begin{aligned}\chi_0^2 &= \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \\ &= \frac{(14-9)^2}{9} + \frac{(11-10.33)^2}{10.33} + \ldots + \frac{(10-5.8)^2}{5.8} = 10.82\end{aligned}$$

As χ_0^2 lies outside R.R, so we cannot reject H_0 given $\alpha = 0.05$ and the independence of the phenomena's, A and B cannot be rejected.

Example 6:
A researcher was inspecting devices in three different institutes due to a certain ministry in order to check the similarity of them. He selected three samples to study their quality (Low, Average and High), and the following table shows results obtained:

Institutes(A)	Quality of the Devices(B)			Total
	Low	Average	High	
(I)	12	23	89	124
(II)	8	12	62	82
(III)	25	30	119	174
Total	45	65	270	380

Does this data give an evidence to say the qualities of the devices in the three institutes are the same? Use $\alpha = 0.01$.

Solutions:
Note that this is an application for testing weather that the two (phenomena's, A and B) are Homogeneity? We are testing The Null hypothesis:

H_0 : Homogeneity property is satisfied in the two phenomena's, A and B

H_1 : Homogeneity property is not satisfied between the two phenomena's, A and B.

The test statistic is:

$$\chi_0^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi_\nu^2, \nu = (r-1) \times (c-1) = 2 \times 2 = 4$$

$$\text{and R.R} = (\chi^2_{0.05,4}, \infty) = (13.277, \infty)$$

Now, we add the expected values $\chi_0^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$ in the following table for our calculation:

Institutes(A)	Quality of Devices(B)			Total
	Low	Average	High	
(I)	12 14.68	23 21.21	89 88.11	124
(II)	8 9.71	12 14.03	62 58.26	82
(III)	25 20.61	30 29.76	119 123.63	174
Total	45	65	270	380

$$\chi_0^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$
$$= \frac{(12-14.68)^2}{14.68} + \frac{(23-21.21)^2}{21.21} + \ldots + \frac{(119-123.63)^2}{123.63} = 2.59$$

χ_0^2 lies outside R.R, so we cannot reject H_0 and the Homogeneity property is accepted.