Chapter 5

CHI Square tests

5.1 Introduction

In this chapter, we will consider testes for proportions when we have a qualitative variable with more than 2 categories (levels) from one or more populations, For example hair color , blood type and eye color. If we take a random sample of size n with different K categories and we obtain the observed frequencies denoted by $O_1, O_2, \dots, O_K$ and the expected frequencies denoted by $E_1, E_2, \dots, E_K$ $(with\ E_i = nP_i)\ where\ \sum O_i = \sum E_i = n.$ Also, Each category of the variable in the population has proportions $P_1, P_2, \dots, P_K$ such that $\sum_{i=1}^{k} P_i = 1$.

Our aim is to compare how the expected frequencies under the hypothesis match or fit the observed frequencies. So, such tests are known as goodness-of- fit tests which is a tests among a group of methods known as nonparametric tests. We will discuss these tests in chapter 7.

We have three cases for the proportions of K different categories of the qualitative variable, that is:

1- Different proportions are different of each other (specified), i.e.,

$$P_1 = P_{10}, P_2 = P_{20}, \dots, P_K = P_{K0}$$

2- Different proportions are equal, i.e.,

$$P_1 = P_2 = \dots = P_K = \frac{1}{K}$$

3- Specified frequencies (ratio) for different categories, i.e.,

$$f_1 = f_{10}, f_2 = f_{20}, \dots, f_K = f_{K0}$$

And then we can get the different proportion from

$$P_{i0} = \frac{f_{i0}}{\sum f} \quad where \sum f = f_{10} + f_{20} + \dots + f_{K0}$$

5.2 Goodness-of-Fit tests

1- Data $n, \alpha, O_1, O_2, \dots, O_K$

2- The hypothesis:

$$H_0: \begin{cases} P_1 = P_{10}, P_2 = P_{20}, \dots, P_K = P_{K0} \\ P_1 = P_2 = \dots = P_K = \dfrac{1}{K} \\ P_1 = \dfrac{f_1}{\sum f}, P_2 = \dfrac{f_2}{\sum f}, \dots, P_K = \dfrac{f_K}{\sum f} \end{cases}$$

$$H_1: at\ least\ one\ proportion\ is\ different\ from\ H_0$$

3-The test statistic:

$$\chi^2 = \sum_{i=1}^{k} \frac{O_i^2}{E_i} - n$$

Where $O_1\ O_2 \dots . O_K$ are the observed frequencies,

$P_1\ P_2\ \dots\ P_K$ are the specified proportions,

And $E_1, E_2, \dots, E_K$ are the expected frequencies.

4-The table value:

$$\chi^2_{1-\alpha, k-1}$$

5-the decision:

We reject H_0 if $\chi^2 > \chi^2_{1-\alpha, k-1}$

EX(1)

According to the inheritance pattern for flower's color resulting from a cross between red and yellow flowers. We obtain 25% red flowers ,50% orange flowers and 25% yellow flowers. when we apply that theory on 144 flowers, we get 30 red flowers, 78 orange flowers and 36 yellow flowers. Is this data proof the theory at $\alpha = 0.01$.

Solution

1-data: $n = 144, O_1 = 30, O_2 = 78, O_3 = 36, K = 3, \alpha = 0.01$

2- $H_0: P_1 = 0.25\ , P_2 = 0.75\ , P_3 = 0.25$

H_1: At least one proportion is different

3-the test statistic:

$$\chi^2 = \sum_{i=1}^{k} \frac{O_i^2}{E_i} - n = \left[\frac{30^2}{36} + \frac{78^2}{72} + \frac{36^2}{36}\right] - 144 = 1.5$$

Where $E_1 = nP_1 = 144 * 0.25 = 36$,

$E_2 = 144 * 0.75 = 72 \quad and \quad E_3 = 144 * 0.25 = 36$

4-the table value:

$$\chi^2_{1-\alpha\ ,k-1} = \chi^2_{0.99\ ,2} = 9.21$$

5-the decision:

We accept H_0 , since $\chi^2 = 1.5 \ngtr 9.21 = \chi^2_{0.99\ ,2}$

EX(2)

In a study of the strength of the egg's shell for a sample of white chicken eggs and obtain the following frequencies:

Weak	Moderate	Strong
37	68	45

Using $\alpha = 0.05$,

a) Test if the levels of strength of white egg shells occur with equal proportions.
b) Test if the proportions of the levels of strength are different from 1/4, 1/2, and 1/4 respectively.
c) Test if the frequency of the levels of strength are different from a 3:6:1

Solution

a) levels of strength of white egg shells occur with equal proportions:

1-data: $n = 150\ , O_1 = 37, O_2 = 68, O_3 = 45\ , K = 3, \alpha = 0.05$

2- $H_0: P_1 = P_2 = P_3 = 1/3$

H_1: At least one proportion is different

3-the test statistic:

$$\chi^2 = \sum_{i=1}^{k} \frac{O_i^2}{E_i} - n = \left[\frac{37^2}{50} + \frac{68^2}{50} + \frac{45^2}{50}\right] - 150 = 10.36$$

Where $E_1 = E_2 = E_3 = 150 * 1/3 = 50$

4-the table value:

$$\chi^2_{1-\alpha\,,k-1} = \chi^2_{0.95\,,2} = 5.991$$

5-the decision:

We reject H_0 , since $\chi^2 \;\; > \chi^2_{0.95\,,2}$ and accept H_1.

I.e., the levels of strength of white egg shells haven't equal proportions.

b) the proportions of the levels of strength are different from 1/4, 1/2, and 1/4 respectively.

1-data: $n = 150\,, O_1 = 37, O_2 = 68, O_3 = 45\,, K = 3, \alpha = 0.05$

2- $H_0: P_1 = \frac{1}{4} = 0.25, P_2 = \frac{1}{2} = 0.5, P_3 = \frac{1}{4} = 0.25$

H_1: At least one proportion is different

3-the test statistic:

$$\chi^2 = \sum_{i=1}^{k} \frac{O_i^2}{E_i} - n = \left[\frac{37^2}{37.5} + \frac{68^2}{75} + \frac{45^2}{37.5}\right] - 150 = 2.16$$

Where $E_1 = 150 * 0.25 = 37.5,$

$E_2 = 150 * 0.5 = 75, \quad and \qquad E_3 = 150 * 0.25 = 37.5$

4-the table value:

$$\chi^2_{1-\alpha,k-1} = \chi^2_{0.95,2} = 5.991$$

5-the decision:

We accept H_0 , since $\chi^2 \ngtr \chi^2_{0.95,2}$

I.e., the proportions of the levels of strength are 1/4, 1/2, and 1/4 respectively.

c)If the frequency of the levels of strength are different from a 3:6:1

1-data: $n = 150$, $O_1 = 37, O_2 = 68, O_3 = 45$, $K = 3, \alpha = 0.05$

2- $H_0: P_1 = \frac{3}{10} = 0.3, P_2 = \frac{6}{10} = 0.6, P_3 = \frac{1}{10} = 0.1$

H_1: At least one proportion is different

3-the test statistic:

$$\chi^2 = \sum_{i=1}^{k} \frac{O_i^2}{E_i} - n = \left[\frac{37^2}{45} + \frac{68^2}{90} + \frac{45^2}{15}\right] - 150 = 66.8$$

Where $E_1 = 150 * 0.3 = 45$,

$E_2 = 150 * 0.6 = 90$, and $E_3 = 150 * 0.1 = 15$

4-the table value:

$$\chi^2_{1-\alpha,k-1} = \chi^2_{0.95,2} = 5.991$$

5-the decision:

We reject H_0 , since $\chi^2 > \chi^2_{0.95,2}$

I.e., the frequency of the levels of strength are different from a 3:6:1

5.3 Independence test

When we are interested in a population which we draw a random sample of size n . Then we classifying the elements of the sample into two qualitative variables, the first variable with c levels and the second variable with r levels. Thus, we can get the observed frequencies as the following *contingency* table:

Variable 1 with c levels

Variable 2 with r levels		1	2	…	c	Row Totals
	1	O_{11}	O_{12}	……	O_{1c}	$O_{1.}$
	2	O_{21}	O_{22}	……	O_{2c}	$O_{2.}$
	⋮	⋮	⋮		⋮	⋮
	r	O_{r1}	O_{r2}	……	O_{rc}	$O_{r.}$
	Column Totals	$O_{.1}$	$O_{.2}$	……	$O_{.c}$	n

Now, the question is :

Are the two variables independent(not related) in the population?

i.e., is there a relationship between the two variables in the population.

5.3.1 The test steps

1- Data n, α, r, c

2- The hypothesis:

$H_0: variable\ 1\ is\ independent\ of\ variable\ 2$

$H_1: variable\ 1\ isnot\ independent (related)\ of\ variable\ 2$

3-The test statistic:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{O_{ij}^2}{E_{ij}} - n$$

Where O_{ij} are the observed frequencies,

And $E_{ij} = \frac{O_{i.}\ O_{.j}}{n}$ are the expected frequency.

4-The table value:

$$\chi^2_{1-\alpha,(r-1)(c-1)}$$

5-the decision:

We reject H_0 if $\chi^2 > \chi^2_{1-\alpha,(r-1)(c-1)}$

EX(3)

The use of the internet is known to help student in studying. A random sample of students from KSU University was classified by the usage level of the internet and the degree in final exam of Statistics:

Usage level of internet	A	B	C	D	Ttotal
High	25	46	30	15	116
Moderate	85	25	120	20	250
Low	40	15	15	65	135
Total	150	86	165	100	501

a) Test whether there is a relationship between internet usage and the degree in statistics. use $\alpha = 0.1$.
b) Find the observed frequency of students had degree A and use internet in low level.
c) Find the expected frequency of students had degree C and use internet in moderate level.

Solution

a) **Test whether there is a relationship between internet usage and the degree in Statistics**. use $\alpha = 0.1$

1- Data $n = 501, \alpha = 0.1, r = 3, c = 4$

2- The hypothesis:

$$H_0: internet\ usage\ is\ independent\ of\ the\ degree\ in\ Statistics$$
$$H_1: internet\ usage\ is\ dependent\ of\ the\ degree\ in\ Statistics$$

3-The test statistic:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{O_{ij}^2}{E_{ij}} - n = 12.592$$

4-The table value:

$$\chi^2_{1-\alpha\,,(r-1)(c-1)} = \chi^2_{0.9\,,6} = 10.645$$

5-the decision:

We reject H_0 and acceptH_1, since $\chi^2 = 12.592 > 10.645 = \chi^2_{0.9\,,6}$
i.e., there is a relationship between the internet usage and the degree in the final exam in Statistics.

b) Find the observed frequency of students had degree A and use internet in low level.

$$O_{31} = 40$$

c)Find the expected frequency of students had degree C and use internet in moderate level.

$$E_{23} = \frac{O_{i.} * O_{.j}}{n} = \frac{O_{2.} * O_{.3}}{n} = \frac{250 * 165}{501} = 82.335$$

5.4 Homogeneous Test

When we are interested in studying a variable with C levels in more than two populations say r . Then, we draw several independent samples of these populations , so we obtain r samples. n_1 from population 1 , n_2 from population 2,….,n_r from population r . Then we classifying each sample by the levels of the single variable. Thus, we can get the observed frequencies as the following *contingency* table:

Level of the variable

Samples from r populations	1	2	…	c	Row Totals
1	O_{11}	O_{12}	……	O_{1c}	$O_{1.}$
2	O_{21}	O_{22}	……	O_{2c}	$O_{2.}$
⋮	⋮	⋮		⋮	⋮
R	O_{r1}	O_{r2}	……	O_{rc}	$O_{r.}$
Column Totals	$O_{.1}$	$O_{.2}$	……	$O_{.c}$	n

Now, the question is :

Are these ***r*** populations homogenous with respect to the variable ?
i.e., are the proportions in each category the same for every population?

5.4.1 The test steps

1- Data n, α, r, c

2- The hypothesis:

H_0: the r populations are homogenous w.r.t. the variable

H_1: the r populations are not homogenous

3-The test statistic:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{O_{ij}^2}{E_{ij}} - n$$

Where O_{ij} are the observed frequencies,

And $E_{ij} = \frac{O_{i.}\ O_{.j}}{n}$ are the expected frequency.

4-The table value:

$$\chi^2_{1-\alpha\,,(r-1)(c-1)}$$

5-the decision:

We reject H_0 if $\chi^2 > \chi^2_{1-\alpha\,,(r-1)(c-1)}$

EX(4)

The following contingency table indicates the number of students with their result (success or fall) in three classes A,B,C:

	Success	**fall**	**Total**
A	50	5	55
B	47	14	61
C	56	8	64
Total	153	27	180

Test whether the proportions of the result are the same for the three classes. Use level of significance of 0.01.

(Are the three classes homogeneous w.r.t. the result proportions at $\alpha = 0.01$?)

Solution

d) **Test whether there is a relationship between internet usage and the degree in Statistics**. use $\alpha = 0.1$

1- Data $n = 180, \alpha = 0.01, r = 3, c = 2$

2- The hypothesis:

H_0: $the\ three\ classes\ are\ homogenuous\ w.r.t.\ the\ result$

H_1: $the\ three\ classes\ are\ not\ homogenuous\ w.r.t.\ the\ result$

3-The test statistic:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{O_{ij}^2}{E_{ij}} - n = 4.84$$

4-The table value:

$$\chi^2_{1-\alpha,(r-1)(c-1)} = \chi^2_{0.99,2} = 9.21$$

5-the decision:

We accept H_0, since $\chi^2 = 4.84 \ngtr 9.21 = \chi^2_{0.99,2}$

i.e. The three classes are homogeneous w.r.t. the proportions of success and fall.

EX(5)

Three random samples from three countries which are Saudi Arabia, Egypt, and Qatar are asked about their opinion in medical care level in their countries , we get the following frequencies:

	Excellent	Good	Acceptance	Total
Saudi Arabia	105	59	36	200
Egypt	72	46	32	150
Qatar	70	52	28	150
Total	247	157	96	500

a) Is this data indicates the homogeneous between the three countries w.r.t. medical care level at $\alpha = 0.1$ and the statistic value equals1.969.
b) Find the observed frequency of persons from Saudi Arabians with acceptance opinion .

c) Find the expected frequency of persons from Qatar with excellent opinion.

solution

1- Data $n = 500, \alpha = 0.1, r = 3, c = 3$

2- The hypothesis:

H_0: *the three countries are homogenuous w.r.t. the opinions of the medical care*

H_1: *the three classes are not homogenuous*

3-The test statistic:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{O_{ij}^2}{E_{ij}} - n = 1.969$$

4-The table value:

$$\chi^2_{1-\alpha,(r-1)(c-1)} = \chi^2_{0.9,4} = 7.779$$

5-the decision:

We accept H_0, since $\chi^2 = 1.969 \ngtr 7.779 = \chi^2_{0.9,4}$

i.e. *the three countries are homogenuous w.r.t. the opinions of the medical care.*

b) The observed frequency of persons from Saudi Arabians with acceptance opinion .

$$O_{13} = 36\, persons$$

c) Find the expected frequency of persons from Qatar with excellent opinion.

$$E_{31} = \frac{O_{3.}\ O_{.1}}{n} = \frac{150 * 247}{500} = 74.1$$