Chapter 6
Chi square tests

Department of Statistics and Operations Research

February 11, 2019

Plan

1. Goodness-of-Fit Test
2. Test for independence (categorical data)
3. Test for homogeneity

Plan

1. Goodness-of-Fit Test
2. Test for independence (categorical data)
3. Test for homogeneity

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.

A **goodness-of-fit test** between observed and expected frequencies is based on the quantity

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i},$$

- The symbols o_i represents the observed frequencies for the i^{th} cell.
- The symbols e_i represents the expected frequencies for the i^{th} cell.
- χ^2 is a value of a random variable whose sampling distribution is approximated very closely by the chi-squared distribution with $\nu = k - 1$ degrees of freedom.

- If the observed frequencies are close to the corresponding expected frequencies, the χ^2 value will be small, indicating a good fit.
- If the observed frequencies differ considerably from the expected frequencies, the χ^2 value will be large and the fit is poor.
- $$\begin{cases} H_0: & \text{A good fit test;} \\ H_1: & \text{A poor fit test.} \end{cases}$$

The critical region will fall in the right tail of the chi-squared distribution. For a level of significance equal to α, we find the critical value χ^2_α from Table A.5, and then $\chi^2 > \chi^2_\alpha$ constitutes the critical region.

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

$$f(x) = \frac{1}{6},\ x = 1, 2, 3, 4, 5, 6$$

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 hypothesis, H_0 : is H_0 : the die is fair.

Solution

We find, from the table, that the χ^2-value to be

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

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

Example 2

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

Class Boundaries	Frequency
1.45-1.95	2
1.95-2.45	1
2.45-2.95	4
2.95-3.45	15
3.45-3.95	10
3.95-4.45	5
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: the data is normal $N(3.5, 0.7)$.

Solution

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

$P(-2.93 < Z < -2.21) = P(Z < -2.21) - P(Z < -2.93) = 0.0119.$

Hence, the expected frequency for the first class is

$$e_1 = 40(0.0119) = 0.5$$

Similarly, we get

$$e_2 = 40[P(-2.21 < Z < -1.5)] = 2.1$$

$$e_3 = 40[P(-1.5 < Z < -0.79)] = 5.9$$

$$e_4 = 40[P(-0.79 < Z < -0.07)] = 10.3$$

$$e_5 = 40[P(-0.07 < Z < 0.64)] = 10.7$$

$$e_6 = 40[P(0.64 < Z < 1.36)] = 7.0$$

$$e_7 = 40[P(1.36 < Z < 2.07)] = 3.5$$

The χ^2-value is then given by

$$\begin{aligned}\chi^2 &= \frac{(2-0.5)^2}{0.5}+\frac{(1-2.1)^2}{2.1}+\frac{(4-5.9)^2}{5.9}+\frac{(15-10.3)^2}{10.3}\\ &+ \frac{(10-10.7)^2}{10.7}+\frac{(5-7)^2}{7}+\frac{(3-3.5)^2}{3.5}=3.05\end{aligned}$$

Since the computed χ^2-value is less than $\chi^2_{0.05} = 12.592$ for 6 degrees of freedom, 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.

Plan

The chi-squared test procedure is used to test the hypothesis H_0 of independence of two variables of classification. The general rule for obtaining the expected frequency of any cell is given by the following formula:

$$e_{ij} = \frac{(j^{th} \text{ column total}) \times (i^{th} \text{ row total})}{\text{grand total}}$$

A simple formula providing the correct number of degrees of freedom is

$$\nu = (c-1) \times (r-1)$$

The test statistic:

$$\chi^2 = \sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}.$$

Example

Suppose that we wish to determine whether the opinions of the voting residents of the state of Illinois concerning a new tax reform are independent of their levels of income. Members of a random sample of 1000 registered voters from the state of Illinois are classified as to whether they are in a low, medium, or high income bracket and whether or not they favor the tax reform. The observed frequencies are presented in Table 6.2.

Tax Reform	Low	Medium	High	Total
For	182	213	203	598
Against	154	138	110	402
Total	336	351	313	1000

Our aim is to test the hypothesis:

1. H_0: the two random variables (voter's opinion concerning the tax reform and his or her level of income) are independent.
2. H_1: voter's opinion concerning the tax reform and his or her level of income are not independent.

Solution

Table 6.3: Observed and Expected Frequencies

	Income Level			
Tax Reform	Low	Medium	High	Total
For	182 (200.9)	213 (209.9)	203 (187.2)	598
Against	154 (135.1)	138 (141.1)	110 (125.8)	402
Total	336	351	313	1000

Hence, the expected frequency for the first class is

$$e_{11} = (336 \times 598)/1000 = 200.9.$$

- $\nu = (2-1)(3-1) = 2$ degrees of freedom.
- $\alpha = 0.05$.

- The test statistic:

$$
\begin{aligned}
\chi^2 &= \sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = \frac{(182 - 200.9)^2}{200.9} + \frac{(213 - 209.9)^2}{209.9} \\
&+ \frac{(203 - 187.2)^2}{187.2} + \frac{(154 - 135.1)^2}{135.1} + \frac{(138 - 141.1)^2}{141.1} \\
&+ \frac{(110 - 125.8)^2}{125.8} = 7.85
\end{aligned}
$$

- P-value $\simeq 0.02$.
- From chi-square table we find that $\chi^2_{0.05} = 5.991$ for $\nu = (2-1)(3-1) = 2$ degrees of freedom. The null hypothesis is rejected and we conclude that a voter's opinion concerning the tax reform and his or her level of income are not independent.

Plan

this test determines if two or more populations have the same distribution of a single categorical variable. The general rule for obtaining the expected frequency of any cell is given by the following formula:

$$e_{ij} = \frac{(j^{th}\text{ column total}) \times (i^{th}\text{ row total})}{\text{grand total}}$$

A simple formula providing the correct number of degrees of freedom is

$$\nu = (c - 1) \times (r - 1)$$

The test statistic:

$$\chi^2 = \sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}.$$

The null hypothesis is

$$H_0 : \text{For each row } i,\ p_{i1} = \cdots = p_{ic},$$

where p_{ij} are the proportions. The alternative hypothesis H_1 is that at least one of the null hypothesis statements is false. This test is called a test for homogeneity.

Example

Suppose, for example, that we decide in advance to select 200 Democrats, 150 Republicans, and 150 Independents from the voters of the state of North Carolina and record whether they are for a proposed abortion law, against it, or undecided.

Abortion Law	Democrat	Republican	Independent	Total
For	82	70	62	214
Against	93	62	67	222
Undecided	25	18	21	64
Total	200	150	150	500

We want to test the hypothesis that the proportions of Democrats, Republicans, and Independents favoring the abortion law are the same, the proportions of each political affiliation against the law are the same, and the proportions of each political affiliation that are undecided are the same.

Solution

The observed and the expected frequencies are given by

Abortion Law	Democrat	Republican	Independent	Total
For	82(85.6)	70(64.2)	62(64.2)	214
Against	93(88.8)	62(66.6)	67(66.6)	222
Undecided	25(25.6)	18(19.2)	21(19.2)	64
Total	200	150	150	500

- H_0: For each opinion, the proportions of Democrats, Republicans, and Independents are the same.
- H_1: For at least one opinion, the proportions of Democrats, Republicans, and Independents are not the same.
- $\alpha = 0.05$.
- Critical region: $\chi^2 > \chi^2_\alpha$ with $\nu = (3-1)(3-1) = 4$ degrees of freedom. Then the critical region: $\chi^2 > 9.488$.
- We obtain that $\chi^2 = 1.53$.
- Decision: Do not reject H_0. There is insufficient evidence to conclude that the proportions of Democrats, Republicans, and Independents differ for each stated opinion.