Stat 154: Elementary Statistics

Jongyun Jung

Minnesota State University, Mankato

Ch 8: **Chi-square Tests and Analysis of Variance**

jongyun.jung@mnsu.edu

April 12, 2019

Overview

Chi-square Goodness-of-fit Test

- If there is a claim that a sample is from a certain distribution, a chi-square deviance statistic is computed, if the deviance is high, the claim is refuted.
- The test statistic is

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

- where O stands for the observed sample frequency for a group (or value) and E stands for the corresponding expected frequency under the claim, the sum if for all the groups (or values) and χ^2 has a Chi-square distribution with $g - k - 1$ degrees of freedom.
- g is the number of distinct groups (or values) and k is the number of parameters estimated using the data in the process of computing the expected frequencies.

Chi-square Goodness-of-fit Test

- **Example 8.1**: There is a claim that among all the automobiles in the USA, 10% are pick-ups, 15% are vans, 12% are utility vehicles and 63% are sedans. To test the claim, a market research firm randomly selected 1,200 automobiles from the USA and found that 105 are pick-ups, 98 are vans, 114 are utility vehicles, and 883 are sedans. Test the claim using a 10% level of significance.

research firm randomly selected 1,200 automobiles from the USA and found that 105 are pick-ups, 98 are vans, 114 are utility vehicles, and 883 are sedans. Test the claim using a 10% level of significance.

Solution: H_0 : $p_1 = 0.10; p_2 = 0.15; p_3 = 0.12; p_4 = 0.63$
H_1 : at least one proportion is different than as specified

O	105	98	114	883
E	$1{,}200 \times 0.10 = 120$	$1{,}200 \times 0.15 = 180$	$1{,}200 \times 0.12 = 144$	$1{,}200 \times 0.63 = 756$

Then the test statistic

$$\chi^2 = \frac{(105-120)^2}{120} + \frac{(98-180)^2}{180} + \frac{(114-144)^2}{144} + \frac{(883-756)^2}{756}$$
$$= 1.875 + 37.356 + 6.250 + 21.335 = 66.816$$

The upper tail chi-square value at a 10% level of significance for g–k–1 = 4–0–1 = 3 degrees of freedom is 6.25, from Table 4 in Appendix II.

So, we reject the null hypothesis and conclude that the proportions are significantly different than claimed.

Note that here a one sided test is used as the higher deviance indicating the rejection of H_0.

Figure: Ex 8-1

Test for Honogeneity

- Testing for equality or homogeneity of proportions examines the differences between two or more independent proportions.
- The hypothesis is written as

$$\begin{aligned} H_0 &: \quad p_1 = p_2 = \cdots = p_k \\ H_1 &: \quad \text{at least one proportion is different} \end{aligned}$$

- where k is the number of independent populations. Under H_0, the test statistic is

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

- has a chi-square distribution with $k - 1$ degrees of freedom

$$E_i = n_i \overline{p}, \quad \text{where} \quad i = 1, 2 \ldots, k \quad \text{and} \quad \overline{p} = \frac{X_1 + X_2 + \cdots + X_k}{n_1 + n_2 + \cdots + n_k}$$

- where X_i's are the number of successes and n_k's are the number of trials in different populations.

Test for Honogeneity

EXAMPLE 8.4

A study is conducted to justify the fact that in US universities, the gender distributions in majors such as Mathematical Sciences, Biological Sciences, Businesses, Social Sciences, and Arts and Humanities are equal. Samples are taken and reveal that among 800 Mathematical Sciences majors, 356 are women; among 900 Biological Sciences majors, 498 are women; among 760 Business majors, 342 are women; among 480 Social Sciences majors, 258 are women; and among 350 Arts and Humanities majors, 254 are women. Test the claim using a 5% level of significance.

Solution: H_0 : $p_1 = p_2 = \cdots = p_5$

H_1 : at least one proportion is different

$$\bar{p} = \frac{X_1 + X_2 + \cdots + X_5}{n_1 + n_2 + \cdots + n_5} = \frac{356 + 498 + 342 + 258 + 254}{800 + 900 + 760 + 480 + 350} = \frac{1708}{3290} = 0.52$$

$$E_1 = n_1\bar{p} = 800(0.52) = 416, \quad E_2 = n_2\bar{p} = 900(0.52) = 468,$$

$$E_3 = n_3\bar{p} = 760(0.52) = 395.2, \quad E_4 = n_4\bar{p} = 480(0.52) = 249.6 \text{ and}$$

$$E_5 = n_5\bar{p} = 350(0.52) = 182.$$

Then.

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

$$= \frac{(356 - 416)^2}{416} + \frac{(498 - 468)^2}{468} + \frac{(342 - 395.2)^2}{395.2} + \frac{(258 - 249.6)^2}{249.6} + \frac{(254 - 182)^2}{182}$$

$$= 8.65 + 1.92 + 7.16 + 0.28 + 28.48 = 46.49.$$

The 5% upper tail chi-square value at 4 degrees of freedom is 9.49, from Table 4 in Appendix II. So, we reject H_0 and conclude that the proportions are significantly different.

Figure: Ex 8-4

Chi-square Test for Independence

- In testing a significant relationship between two variables having categorical measurements, a chi-square statistic is used.
- The hypothesis is written as

$$\begin{aligned} H_0 &: \text{ Two variables are not related.} \\ H_1 &: \text{ The two variables are related.} \end{aligned}$$

- The test statistic is

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

- where the sum is for all $r \times c$ cells, where r indicates the number of rows and c indicates the number of columns.
- $E_{ig} = \frac{R_i \cdot C_j}{G}$ indicates the expected frequency for the ith row and the jth column, R_i is the ith row total, C_j is the jth column total, and G is the grand total.
- χ^2 follows a chi-square distribution with $(r-1)(c-1)$ degrees of freedom.

Chi-square Test for Independence

EXAMPLE 8.5 Use the data in the following table to test the claim at the 5% level of significance that occupation is independent of whether the cause of death was homicide. The table is based on data from the US Department of Labor Statistics.

	Police	*Cashiers*	*Taxi Drivers*	*Guards*
Homicide	82	107	70	59
Cause of death other than homicide	92	9	29	42

Solution: H_0 : Cause of death and occupation are independent.
H_1 : Cause of death and occupation are dependent.

	Police	*Cashiers*	*Taxi Drivers*	*Guards*	*Total*
Homicide	82(112.92)	107(75.28)	70(64.25)	59(65.55)	318
Cause of death other than homicide	92(61.08)	9(40.72)	29(34.75)	42(34.45)	172
Total	174	116	99	101	490

The numbers in the parentheses are the expected frequencies computed using the above formula, such as, $E_{11} = R_1C_1/G = (318)(174)/490 = 112.92$, and so on.

Then the chi-square statistic is computed as 65.97. The 5% upper chi-square value for $(2 - 1)(4 - 1) = 3$ degrees of freedom is 7.81, from Table 4 in Appendix II. So, we reject the null hypothesis and conclude that homicide and occupation are significantly related.

Figure: Ex 8-5

Analysis of Variance

- In comparing several independent population means, the ratio of the variability within the groups and the variability between the groups are used and known as the analysis of variance.
- The hypothesis is written as

$$\begin{aligned} H_0 &: \quad \mu_1 = \mu_2 = \cdots = \mu_k \\ H_1 &: \quad \text{At least one mean is different.} \end{aligned}$$

- The test statistic is

$$F = \frac{\frac{\sum n_i(\overline{X}_i - \overline{X})^2}{k-1}}{\frac{\sum (n_i-1)s_i^2}{\sum (n_i-1)}}$$

- n_i is the ith sample size; $\overline{X}_i$ is the ith sample mean, $\overline{X}$ is the combined mean of all k samples and s_i^2 is the ith sample variance.
- F has F-distribution with $k-1$ and $\sum(n_i - 1)$ degrees of freedom. The higher value of F indicates rejection of the H_0. The Table 5: F-**Distribution** Table is given in Appendix II.

Analysis of Variance

A service station manager wants to estimate the mean summer weekly demand for three types of gasoline—leaded, unleaded and super unleaded. Data (in thousands of gallons) collected over an eight-week period are shown below:

Week	1	2	3	4	5	6	7	8
Leaded	2.4	2.6	2.7	2.8	3.0	2.6	2.7	2.4
Unleaded	18.2	19.7	21.3	22.4	21.5	18.0	18.0	18.8
Super unleaded	4.3	4.6	4.7	5.5	5.0	4.6	4.6	5.0

Test using the 5% level of significance that there is no significant difference of demand among the gasoline types.

Solution: H_0 : $\mu_1 = \mu_2 = \mu_3$

H_1 : At least one mean is different.

The test statistic,

$$F = \frac{\frac{\sum n_i(\bar{X}_i - \bar{X})^2}{k-1}}{\frac{\sum (n_i - 1)s_i^2}{\sum (n_i - 1)}}.$$

To compute the test statistic, the necessary information is that $n_1 = n_2 = n_3 = 8$, $\bar{X}_1 = 2.65$, $\bar{X}_2 = 19.7375$, $\bar{X}_3 = 4.7875$, $\bar{X} = 9.0583$, $s_1 = 0.2$, $s_2 = 1.7695$ and $s_3 = 0.3682$.

Then

$$F = \frac{\frac{8(2.65 - 9.0583)^2 + 8(19.7375 - 9.0583)^2 + 8(4.7875 - 9.0583)^2}{3-1}}{\frac{(8-1)(0.2)^2 + (8-1)(1.7695)^2 + (8-1)(0.3682)^2}{(8-1)+(8-1)+(8-1)}} = 630.4397.$$

The 5% critical value for F-distribution with 2 and 21 degrees of freedom is 3.4668, from Table 5 in Appendix II.

So, we reject H_0 and conclude that the demands for different gasoline types are significantly different.

Figure: Ex 8-6

References

Mezbahur Rahman, Deepak Sanjel, Han Wu. Statistics Introduction, Revised Printing

KendallHunt