Chapter 7

CHI-SQUARE TESTS

Chi-square Tests (page 555)

- Chi-square tests involve the comparison of the observed frequencies in a one-way or two-way table with the expected frequencies if the null hypothesis were true. The test statistic used will follow a chi-square distribution when the null hypothesis is true. Table B.3 (page 606) presents the $100(1-\alpha)^{th}$ percentiles of the chi-square distribution

- Three tests discussed in book
  - Goodness-of-fit test: used when one wishes to test whether or not a hypothesized probabilistic model provides a good fit for the unknown distribution of the variable of interest. For example, Ho: X follows a normal distribution vs Ha: X does not follow a normal distribution.
  - Test for independence: used to determine whether or not a relationship exists between two categorical variables (nominal or ordinal usually with only a few distinct categories).
  - Test for homogeneity (page 574): used to determine if *c* populations differ with respect to their relative frequency distributions.

Test for Independence: An example

Example 17.4: (page 569) A study was conducted to determine whether the leader-follower tendency of a person is associated with his height. In this study, a sample of 95 people were selected. Based on the information collected, each one in the sample was classified according to their leader-follower tendency and height.

The categories of leader-follower tendency are:

(1) follower – a person who tends to follow

(2) in-between – a person who sometimes tend to follow but other times tend to lead

(3) leader – a person who tends to lead

The categories of height are:

(1) short

(2) tall.

Test the hypothesis that there is a relationship between leader-follower tendency and height at 0.01 level of significance

Steps in Performing a Test for Independence (page 368)

Step 1: State the null and alternative hypotheses.

Ho: The random variables , X and Y, are independent.

Ha: X and Y are not independent.

Ho: Leader-follower tendency (X) and height (Y) are independent/not related/not associated.
Ha: Leader-follower tencency and height are not independent/related/associated.

Step 2: Choose the level of significance, α. α=.01

Step 3. Collect the data. We measure the two variables from each element in the random sample.

Steps in Performing a Test for Independence (cont'd)

Step 4. Construct the r x c contingency table. Compute for the row totals and column totals. Let:

O_{ij} = observed number of elements whose realized value for X is the i^{th} category and whose realized value for Y is the j^{th} category, where $i = 1,2,\ldots,r$ and $j=1,2,\ldots,c$

r = number of rows

c = number of columns

R_i = total number of elements belonging in the i^{th} row, $i = 1,2,\ldots,r$

C_j = total number of elements belonging in the j^{th} column, $j=1,2,\ldots,c$

n = sample size = $\sum_{i=1}^{r} R_i = \sum_{j=1}^{c} C_j$

Leader-Follower Tendency	***Height of Person***		**Total**
	Short	*Tall*	
Follower	($O_{1,1}$) 22	($O_{1,2}$) 14	(R_1) 36
In-between	($O_{2,1}$) 9	($O_{2,2}$) 6	(R_2) 15
Leader	($O_{3,1}$) 12	($O_{3,2}$) 32	(R_3) 44
Total	(C_1) 43	(C_2) 52	**(n) 95**

Steps in Performing a Test for Independence (cont'd)

Step 5. Compute for the expected frequencies using the formula:

$$E_{ij} = \frac{R_i C_j}{n} \text{ for } i=1,2,\ldots,r \text{ and } j=1,2,\ldots,c$$

Leader-Follower Tendency	***Height of Person***		**Total**
	Short	*Tall*	
Follower	22 [16.3]	14 [19.7]	36
In-between	9 [6.8]	6 [8.2]	15
Leader	12 [19.9]	32 [24.1]	44
Total	43	52	**95**

Steps in Performing a Test for Independence (cont'd)

Step 6. Compute the test statistic, given by*:*

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

Leader-Follower Tendency	***Height of Person***		**Total**
	Short	*Tall*	
Follower	22 [16.3]	14 [19.7]	36
In-between	9 [6.8]	6 [8.2]	15
Leader	12 [19.9]	32 [24.1]	44
Total	43	52	**95**

$$X^2 = \frac{(22-16.3)^2}{16.3} + \frac{(14-19.7)^2}{19.7} + \frac{(9-6.8)^2}{6.8} + \frac{(6-8.2)^2}{8.2} + \frac{(12-19.9)^2}{19.9} + \frac{(32-24.1)^2}{24.1}$$

$$= 10.67$$

Steps in Performing a Test for Independence (cont'd)

Step 7. Make the statistical decision.

Decision rule: Reject Ho if $X^2 > c_a^2(v = (r - 1)x(c - 1))$.

Reject Ho if $X^2 > c_{0.01}^2(v = (3 - 1)x(2 - 1) = 2)$.

Reject Ho if $X^2 > 9.21$.

Computed $X^2 = 10.67$. What is the decision?

Modified Chi-Square Test for 2x2 Tables (page 571)

- In the Chi-square test for independence, we approximate the distribution of a discrete random variable involving cell frequencies by the continuous chi-square distribution. This approximation is good so long as the degrees of freedom is more than 1. For 2x2 tables, the degrees of freedom is equal to 1 so we need to incorporate a correction for continuity as follows:

x	***y***		***Row Total***
	0	1	
0	*a*	*b*	*a*+*b*
1	*c*	*d*	*c*+*d*
Column Total	*a*+*c*	*b*+*d*	*a*+*b*+*c*+*d*= *n*

$$X^2 = \frac{n\left(|ad-bc|-\frac{n}{2}\right)^2}{a+b \quad c+d \quad a+c \quad b+d}$$

Example (page 572)

Example 17.5: A service company is classified as small if the number of its employees is at most 200, and it is classified as big otherwise. Profit in sales of services, such as training and consulting, is classified as either low or high.

A sample of 200 service companies produced the following contingency table:

Profit Level	***Size of Company***	
	Small	*Big*
Low service profit	30	63
High service profit	75	32

Test whether the size of the service company is independent of the level of profit in sales of services at the 0.05 level of significance.

Ho: The profit level is independent of the size of the service company.
Ha: The profit level is not independent of the size of the service company.

From the contingency table above, we have *a*=30, *b*=63, *c*=75, *d*=32, and *n*=200. The value of the test statistic is:

$$X^2 = \frac{n\left(|ad-bc|-\frac{n}{2}\right)^2}{(a+b)(c+d)(a+c)(b+d)} = \frac{200\left(|30\cdot 32-63\cdot 75|-100\right)^2}{93\cdot 107\cdot 105\cdot 95} = 27.064$$

Decision rule: Reject the null hypothesis if $X^2 > \chi^2_{0.05}\ \nu = 1\ = 3.841.$

Since the test statistic value of 27.064 is greater than 3.841, we reject the null hypothesis. Therefore, at the 0.05 level of significance, we conclude that the size of the service company is related to the level of profit in sales of services.

Remark 1 on the Use of the Test (page 572)

- The chi-square test requires that the expected frequencies are not very small or else the chi-square distribution will not be a good approximation of the null distribution of the test statistic. The general rule is that an expected frequency that is less than 5 in a cell is too small to use. The test should not be used if more than 20% of the expected frequencies are less than 5 or when any expected frequency is less than 1.

Example: Exercise 1 for Section 17.2 (page 576)

How many cells may be allowed to have expected frequencies less than 5 (but at least 1) for the chi-square test for independence to be valid if you have a contingency table of dimension

- 2x2
- 2x3
- 3x5
- 4x4
- 5x5

Remark 2 on the Use of the Test (page 573)

- The chi-square test for independence is quite sensitive to the sample size. If the sample size is too small, the value is overestimated. On the other hand, if the sample size is too large, the value is underestimated. As a result, even those relationships that are too weak to take into account for practical purposes are very likely to be detected by the test when the sample size is large. To overcome this problem, it is recommended that we estimate some measure of association so that we can assess how strong the relationship is.

  Cramer's V is a statistic that measures the strength of association or dependency between two (nominal/ordinal) categorical variables in a contingency table.

$$V = \sqrt{\frac{\chi^2}{n \ \min(r-1, c-1)}}$$

  $0 \le V \le 1$. The closer V is to 0, the weaker the association between the two variables. The closer V is to 1, the stronger the association between the two variables.

Example: The association between leader-follower tendency and height in the earlier example is

$$V = \sqrt{\frac{10.67}{95 \min(2,1)}} = \sqrt{\frac{10.67}{95(1)}} = 0.335$$

Remark 3 on the Use of the Test (page 573)

- It is not appropriate to perform pairwise chi-square test for independence in studying the relationship of 3 or more categorical variables.

  Example: We may be interested in determining whether a family's socio-economic classification (lower, middle, or upper class), level of parental encouragement (low or high), and whether or not a child performs well in school (good or bad performance) are associated.

  If we perform pairwise chi-square test for independence, it is possible that the test leads us to conclude that socio-economic classification and a child's performance in school are related. The detected relationship may be spurious and misleading since it could be attributed to their relationship to parental encouragement which was not considered in the pairwise analysis. The proper approach is to study these relationships simultaneously. This will be discussed in Stat 149. Introduction to Categorical Data Analysis.

Test for Homogeneity (page 574)

- Problem: We wish to determine if *c* populations differ with respect to their relative frequency distributions.
- Approach: We take a random sample of size n_j, j=1,2,…,c, from each of them to test the hypotheses:

  Ho: The relative frequency distributions of the c populations are the same.

  Ha: The relative frequency distribution of at least one of the c populations is different.

  Construct the rxc contingency table and perform the chi-square test using the same test statistic and region of rejection as the test for independence.
- Remark: The major difference between the test for independence and test for homogeneity is in the sampling design. In the test for independence, we take a random sample of size n and measure the two variables from each element. Consequently, the column totals and row totals are all random variables. In the test for homogeneity, we take c random samples of size $n_j=C_j$=jth column total so that the column totals are fixed by design and it is just the row totals that are random variables. Despite this difference, the tests are exactly the same.

Example 17.6 (page 575)

Example 17.6: In a masteral thesis by Ayroso (2004), a sample of faculty members and students was selected from each one of the three different types of university food services: (1) school-owned, (2) concessionaire, and (3) cooperative. Part of the data analyzed for the study was the frequency of eating lunch of students and faculty members in the three different types of university foodservices. The collected data are summarized as follows:

Frequency of Eating Lunch	*Type of Foodservice*			Total
	School-owned	*Concessionaire*	*Cooperative*	
Once	36	26	33	95
Twice	19	13	29	61
3-4 times a week	11	15	26	52
Daily	14	14	27	55
Total	80	68	115	**263**

Given these data, test whether the distributions of frequency of eating lunch are different for the three types of food service. Use α = 0.05.

Example 17.6 (Solution)

Ho: The distributions of the frequency of eating lunch are the same for the three types of food service.

Ha: The distribution of the frequency of eating lunch is different for at least one of the three types of food service.

Frequency of Eating Lunch	*Type of Foodservice*			Total
	School-owned	*Concessionaire*	*Cooperative*	
Once	36 *28.9*	26 24.6	33 41.5	95
Twice	19 18.6	13 15.8	29 26.7	61
3-4 times a week	11 15.8	15 13.4	26 22.7	52
Daily	14 16.7	14 14.2	27 24.0	55
Total	80	68	115	**263**

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{O_{ij}^2}{E_{ij}} - n = \sum_{i=1}^{4}\sum_{j=1}^{3}\frac{O_{ij}^2}{E_{ij}} - 263$$

$$= \frac{36^2}{28.9}+\frac{26^2}{24.6}+\frac{33^2}{41.5}+\frac{19^2}{18.6}+\frac{13^2}{15.8}+\frac{29^2}{26.7}+\frac{11^2}{15.8}+\frac{15^2}{13.4}+\frac{26^2}{22.7}+\frac{14^2}{16.7}+\frac{14^2}{14.2}+\frac{27^2}{24.0}-263 = 270.311 - 263 = \boxed{\mathbf{7.311}}$$

Decision rule: Reject Ho if $X^2 > \chi^2_{0.05}$ $v = (4-1)(3-1) = 6$ $= 12.592$.

Since $7.311 \not> 12.592$, we do not reject Ho. We do not have sufficient evidence at 0.05 level of significance to say that the frequency of eating lunch differ for the three types of university food service.

Using PhStat to Create Two-Way Table

Step 1. Encode measurements taken from ith element in the ith row. Encode the values of the first variable in one column and the values of the second variable in another column.

Step 2. Select *Descriptive Statistics*, then choose *Two Way Tables and Chart.*

Step 3. Fill up dialogue box.

Using PhStat to Perform the Chi-square Test

Step 1. Create two-way table.

Step 2. Select *Multiple-Sample Tests*, then choose *Chi-square Test.*

Step 3. Indicate dimension of table (number of rows and number of columns).

Step 4. Fill-up the Table of Observed Frequencies.

Assignment

A researcher wishes to determine whether there exists a relationship between music preference and IQ. He collected data from a random sample of 480 college students on their IQs and music preference. Below is the contingency table of the collected data:

Music Preference	IQ		
	High	Medium	Low
Classical	40	26	17
Pop	47	59	25
Rock	83	104	79

a) State Ho and Ha.
b) Write the formula of the test statistic to be used.
c) State the decision rule at 0.05 level of significance
d) Present the table of expected frequencies.
e) Compute for the value of the test statistic.
f) Is there sufficient evidence at 0.05 level of significance to conclude that there is a relationship between music preference and IQ?
g) Compute for Cramer's V