Chapter 12 Inference on Categorical Data

Section 12.1 Goodness-of-Fit Test

Objectives

1. Perform a Goodness-of-fit test

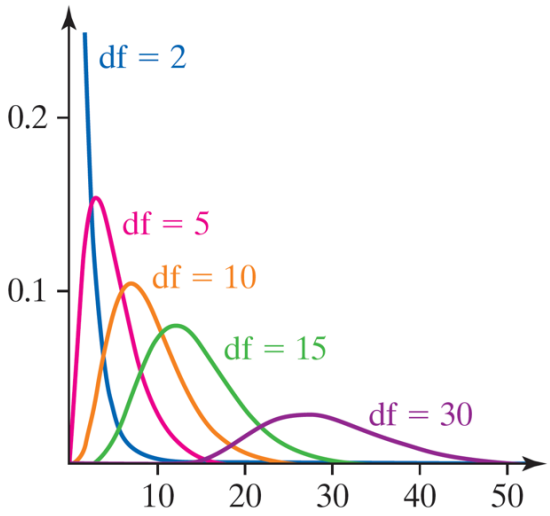
1 Perform a Goodness-of-Fit Test

We begin by introducing a new distribution, called the chi-square distribution.

Characteristics of the Chi-Square Distribution

1. It is not symmetric.
2. Its shape depends on the degrees of freedom, just like Student's t-distribution.
3. As the number of degrees of freedom increases, it becomes more symmetric, as illustrated in Figure 1.
4. The values of χ^2 are nonnegative. That is, the values of χ^2 are greater than or equal to 0.

Figure 1
Chi-square distributions



A **goodness-of-fit test** is an inferential procedure used to determine whether a frequency distribution follows a specific distribution.

Expected Counts

Suppose there are n independent trials of an experiment with $k \geq 3$ mutually exclusive possible outcomes. Let p_1 represent the probability of observing the first outcome and E_1 represent the expected count of the first outcome, p_2 represent the probability of observing the second outcome and E_2 represent the expected count of the second outcome, and so on. The expected counts for each possible outcome are given by

$$E_i = \mu_i = np_i \quad \text{for} \quad i = 1, 2, \ldots, k$$

Example Finding Expected Counts

For any number representing a positive count, the first digit cannot be 0. We might guess that the values 1, 2, 3, 4, 5, 6, 7, 8, and 9 might all be equally likely to be the first digit, but Benford's Law states that in many situations where counts accumulate over time, the first digit in the count follow a particular pattern. The Bloomberg's web site (www.bloomberg.com) gives information on the stock price and trading volume of the members of the S&P 500 stock index. A random sample of 100 stocks were selected, and the first digit of the daily stock volume was recorded. The volume represents the total number of shares traded in a given day.

To get some intuition on Benford's Law, suppose that 100,000 shares of a particular stock have been traded in the first part of a business day. To reach the 200,000 mark, it will take approximately a 100% increase in time. Then, to go from 200,000 to 300,000 shares will take a 50% increase in time. After that, increasing from 300,000 to 400,000 represents a 33% increase in time. In these types of settings, the common first digit is 1. The next most common is 2, and so on. The least frequently occurring first digit is 9. The relative frequencies for the first digit are given in the table below.

First digit	1	2	3	4	5	6	7	8	9
Rel. freq.	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

Source: T.P. Hill, The First Digit Phenomenon, *American Scientist*, July-August, 1998

Below are the counts of the number of times each of the digits 1-9 appeared as the first digit in the stock volume on a randomly selected day.

First digit	1	2	3	4	5	6	7	8	9
Frequency	28	16	18	8	7	6	6	5	6

Source: www.bloomberg.com/markets/stocks/movers_index_spx.html

Find the expected counts for each first digit assuming first digits follow Benford's Law.

Test Statistic for Goodness-of-Fit Tests

Let O_i represent the observed counts of category i, E_i represent the expected counts of category i, k represent the number of categories, and n represent the number of independent trials of an experiment. Then

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \qquad i = 1, 2, \ldots, k$$

approximately follows the chi-square distribution with $k - 1$ degrees of freedom, provided that

1. all expected frequencies are greater than or equal to 1 (all $E_i \geq 1$) and
2. no more than 20% of the expected frequencies are less than 5.

Note: $E_i = np_i$ for $i = 1, 2, \ldots, k$.

The Goodness-of-Fit Test

To test hypotheses regarding a distribution, use the steps that follow.

CAUTION!
If the requirements in Step 3(b) are not satisfied, one option is to combine two or more low-frequency categories into a single category.

Step 1 Determine the null and alternative hypotheses:

H_0: The random variable follows a certain distribution.
H_1: The random variable does not follow the distribution in the null hypothesis.

Step 2 Decide on a level of significance, α, depending on the seriousness of making a Type I error.

Step 3

(a) Calculate the expected counts, E_i, for each of the k categories: $E_i = np_i$ for $i = 1, 2, \ldots, k$, where n is the number of trials and p_i is the probability of the ith category, assuming that the null hypothesis is true.

(b) Verify that the requirements for the goodness-of-fit test are satisfied.

1. All expected counts are greater than or equal to 1 (all $E_i \geq 1$).
2. No more than 20% of the expected counts are less than 5.

Classical Approach

Step 3 (continued)

(c) Compute the **test statistic**

$$\chi_0^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Note: O_i is the observed count for the ith category.

Step 4 Determine the critical value using Table VIII. All goodness-of-fit tests are right-tailed tests, so the critical value is χ_α^2 with $k - 1$ degrees of freedom. See Figure 2.

P-Value Approach

By Hand Step 3 (continued)

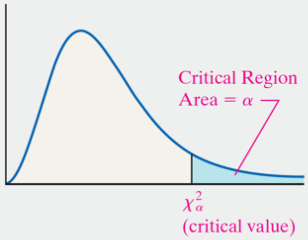
(c) Compute the **test statistic**

$$\chi_0^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Note: O_i is the observed count for the ith category.

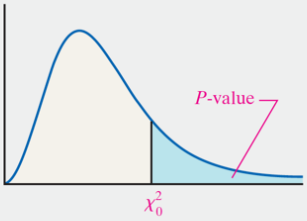
(d) Use Table VIII to approximate the P-value by determining the area under the chi-square distribution with $k - 1$ degrees of freedom to the right of the test statistic. See Figure 3.

Figure 2



Compare the critical value to the statistic. If $\chi_0^2 > \chi_\alpha^2$, reject the null hypothesis.

Figure 3



Technology Step 3 (continued)

(c) Use a statistical spreadsheet or calculator with statistical capabilities to obtain the *P*-value. The directions for obtaining the *P*-value using the TI-84 Plus graphing calculators, Minitab, Excel, and StatCrunch are given in the Technology Step-by-Step on pages 586–587.

Step 4 If *P*-value $< \alpha$, reject the null hypothesis.

Step 5 State the conclusion.

Example Perform a Goodness-of-Fit Test

Using the 0.05 level of significance, test whether the first digits in the stock price follow the relative frequencies given by Benford's Law.

Example Perform a Goodness-of-Fit Test

Is there a birthday effect for men's college baseball? The following data show the birth month of current college baseball players when the cut-off date for determining age was August 1. So, a male born in August would be the oldest for that age-group.

Month	Jan	Feb	March	April	May	June	July	Aug	Sept	Oct	Nov	Dec
Number of Players	96	74	80	76	73	73	77	73	81	110	66	108

Source: NCAA

Does the data suggest that birth month is not equally distributed across the twelve months? If birth month is not equally distributed, which months have more players than expected? What does this suggest? Use a level of significance of 0.05.

Section 12.2 Tests for Independence and the Homogeneity of Proportions

Objectives

1. Perform a test for independence
2. Perform a test for homogeneity of proportions

1 Perform a Test for Independence

The **chi-square test for independence** is used to determine whether there is an association between a row variable and column variable in a contingency table constructed from sample data.

- The null hypothesis is that the variables are not associated, or independent.
- The alternative hypothesis is that the variables are associated, or dependent.

The idea behind testing these types of hypotheses is to compare actual counts to the counts we would expect assuming that the variables are independent (that is, assuming that the null hypothesis were true). If a significant difference between the actual counts and expected counts exists, we have evidence against the null hypothesis.

If two events are independent, then

$$P(E \text{ and } F) = P(E)P(F)$$

We can use the Multiplication Principle for Independent Events to obtain the expected proportion of observations within each cell under the assumption of independence and multiply this result by n, the sample size, in order to obtain the expected count within each cell.

Example Determining the Expected Counts in a Test for Independence

In a poll, 883 males and 893 females were asked "If you could have only one of the following, which would you pick: money, health, or love?" Their responses are presented in the table below. Determine the expected counts within each cell assuming that gender and response are independent.

	Money	Health	Love
Men	82	446	355
Women	46	574	273

Expected Frequencies in a Chi-Square Test for Independence

To find the expected frequency in a cell when performing a chi-square independence test, multiply the cell's row total by its column total and divide this result by the table total. That is,

$$\text{Expected frequency} = \frac{(\text{row total})(\text{column total})}{\text{table total}} \quad \textbf{(1)}$$

Test Statistic for the Test of Independence

Let O_i represent the observed number of counts in the ith cell and E_i represent the expected number of counts in the ith cell. Then

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

approximately follows the chi-square distribution with $(r - 1)(c - 1)$ degrees of freedom, where r is the number of rows and c is the number of columns in the contingency table, provided that (1) all expected frequencies are greater than or equal to 1 and (2) no more than 20% of the expected frequencies are less than 5.

Chi-Square Test for Independence

To test hypotheses regarding the association between (or independence of) two variables in a contingency table, use the steps that follow:

Step 1 Determine the null and alternative hypotheses.

H_0: The row variable and column variable are independent.
H_1: The row variable and column variable are dependent.

Step 2 Choose a level of significance, α, depending on the seriousness of making a Type I error.

Step 3
(a) Calculate the expected frequencies (counts) for each cell in the contingency table using Formula (1).
(b) Verify that the requirements for the chi-square test for independence are satisfied:

1. All expected frequencies are greater than or equal to 1 (all $E_i \geq 1$).
2. No more than 20% of the expected frequencies are less than 5.

Classical Approach

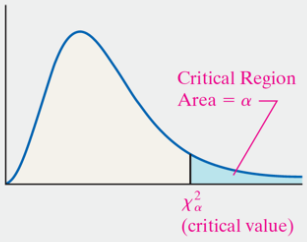
Step 3 (continued)
(c) Compute the **test statistic**

$$\chi_0^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Note: O_i is the observed frequency for the ith cell.

Step 4 Determine the critical value using Table VIII. All chi-square tests for independence are right-tailed tests, so the critical value is χ_α^2 with $(r-1)(c-1)$ degrees of freedom, where r is the number of rows and c is the number of columns in the contingency table. See Figure 9.

Figure 9



Compare the critical value to the test statistic. If $\chi_0^2 > \chi_\alpha^2$, reject the null hypothesis.

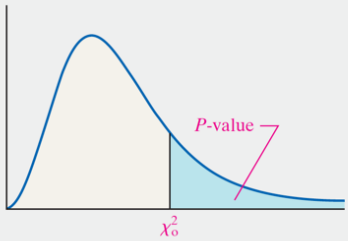
P-Value Approach

By Hand Step 3 (continued)
(c) Compute the **test statistic**

$$\chi_0^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Note: O_i is the observed frequency for the ith cell.
(d) Use Table VIII to determine an approximate P-value by determining the area under the chi-square distribution with $(r-1)(c-1)$ degrees of freedom to the right of the test statistic, where r is the number of rows and c is the number of columns in the contingency table. See Figure 10.

Figure 10



Technology Step 3 (continued)
(c) Use a statistical spreadsheet or calculator with statistical capabilities to obtain the P-value. The directions for obtaining the P-value using the TI-83/84 Plus graphing calculators, Minitab, Excel, and StatCrunch are in the Technology Step-by-Step on pages 602–603.

Step 4 If P-value $< \alpha$, reject the null hypothesis.

Step 5 State the conclusion.

Example Performing a Chi-Square Test for Independence

Are gender and response to, “If you could have only one of the following, which would you pick: money, health, or love?“ independent?

To see the relation between response and gender, we draw bar graphs of the conditional distributions of response by gender. Recall that a conditional distribution lists the relative frequency of each category of a variable, given a specific value of the other variable in a contingency table.

Example Constructing a Conditional Distribution and Bar Graph

❷ Perform a Test for Homogeneity of Proportions

In a **chi-square test for homogeneity of proportions**, we test whether different populations have the same proportion of individuals with some characteristic.

The procedures for performing a test of homogeneity are identical to those for a test of independence.

While the procedures for the test for independence and the test of homogeneity of proportions are the same, the data differ. How? In the test for independence, we are measuring two variables (such as marital status and level of happiness) on each individual. In other words, a single population is segmented based on the value of two variables. In the test of homogeneity of proportions, we consider whether the proportion of individuals among different populations have the same value.

So, if you have a single population in which two variables are measured on each individual to assess whether one variable might be associated with another, conduct a test of independence. If you have two or more populations in which you want to determine equality of proportions among the populations, conduct a test of homogeneity of proportions.

Example Perform a Chi-Square Test for Homogeneity of Proportions

Do you believe everyone has an equal opportunity to obtain a quality education in the United States? The results of this General Social Survey question are presented next by level of education.

	Highest Degree				
	Less Than High School	High School	Junior College	Bachelor	Graduate
Yes	302	551	29	100	46
No	83	200	24	71	31

Source: General Social Survey

a. Researchers wanted to determine whether there is a difference in the proportion of individuals who feel everyone has an equal opportunity to obtain quality education in the United States. Explain why this data should be analyzed by homogeneity of proportions.
b. Does the evidence suggest that the proportion of individuals who feel everyone has an equal opportunity to obtain a quality education in the United States is equal for each level of education? Use the $\alpha = 0.05$ level of significance.
c. Construct a conditional distribution of opinion by level of education and draw a bar graph of the conditional distribution. What role does education appear to play in beliefs about access to quality education?