# Chapter 12 – Inference on Categorical Data

## OUTLINE

1. Goodness-of-Fit Test
2. Tests for Independence and the Homogeneity of Proportions
3. Inference about Two Population Proportions: Dependent Samples

## Putting It Together

In Chapters 9 through 11, we introduced statistical methods for testing hypotheses regarding a parameter such as *p* or .

Often, however, rather than testing a hypothesis regarding a parameter of a probability distribution, we want to test a hypothesis regarding the entire probability distribution. For example, we might test if the distribution of colors in a bag of plain M&M candies is 13% brown, 14% yellow, 13% red, 20% orange, 24% blue, and 16% green. Methods for testing such hypotheses are covered in Section 12.1.

In Section 12.2, we discuss a method for determining whether two qualitative variables are independent based on a sample. If they are not independent, the value of one variable affects the value of the other variable, and the variables are related. We conclude Section 12.2 by introducing tests for homogeneity, which compare proportions from two or more populations. This test are an extension of the two-sample *z*-test for proportions from independent samples discussed in Section 11.1.

We end the chapter with a discussion for comparing two proportions from dependent samples in Section 12.3.

## Section 12.1 Goodness-of-Fit Test

### Objective

1. Perform a Goodness-of-Fit Test

Introduction, Page 2

1. State the characteristics of the chi-square distribution.

Introduction, Page 4

1. What does  represent?

Introduction, Page 5

**Example 1 *Finding Critical Values for the Chi-Square Distribution***

Find the critical values that separate the middle 90% of the chi-square distribution from the 5% area in each tail, assuming 15 degrees of freedom.

Introduction, Page 7

1. Explain how to determine which row in Table VIII to use if the number of degrees of freedom is not in the table.

#### Objective 1: Perform a Goodness-of-Fit Test

Objective 1, Page 1

1. State the definition of a goodness-of-fit test.

Objective 1, Page 2

1. Explain how to obtain the expected counts for a goodness-of-fit test.

Objective 1, Page 3

**Example 2 *Finding Expected Counts***

One growing concern regarding the U.S. economy is the inequality in the distribution of income. The data in Table 1 represents the distribution of household income for various levels of income in 2000. An economist would like to know if income distribution is changing, so she randomly selects 1500 households and obtains the household income. Find the expected number of households in each income level assuming that the distribution of income has not changed since 2000.

**Note:** The income data has been adjusted for inflation.

**Table 1**

Distribution of income in the U.S. in 2000

| Income | Percent |
| --- | --- |
| Under $15,000 | 10.2 |
| $15,000 to $24,999 | 9.8 |
| $25,000 to $34,999 | 9.9 |
| $35,000 to $49,999 | 13.3 |
| $50,000 to $74,999 | 18.1 |
| $75,000 to $99,999 | 13.1 |
| $100,000 to $149,999 | 14.6 |
| $150,000 to $199,999 | 5.8 |
| At least $200,000 | 5.2 |

Data from U.S. Census Bureau

Objective 1, Page 6

1. State the test statistic for a goodness-of-fit test.
2. What two conditions are needed in order to use the test statistic **?

Objective 1, Page 7

1. State the five steps for a goodness-of-fit test.

Step 1

Step 2

Step 3

Step 4

Step 5

Objective 1, Page 8

**Example 3 *Conducting a Goodness-of-Fit Test***

One growing concern regarding the U.S. economy is the inequality in the distribution of income. The data in Table 2 represents the distribution of household income in 2000. An economist would like to know if income distribution is changing, so she randomly selected 1500 households and obtained the household income. Table 3 contains the results of this survey

Note: The data in Table 3 are based on the 2016 Current Population Survey and have been adjusted for inflation.

**Table 2**

| Income | Percent |
| --- | --- |
| Under $15,000 | 10.2 |
| $15,000 to $24,999 | 9.8 |
| $25,000 to $34,999 | 9.9 |
| $35,000 to $49,999 | 13.3 |
| $50,000 to $74,999 | 18.1 |
| $75,000 to $99,999 | 13.1 |
| $100,000 to $149,999 | 14.6 |
| $150,000 to $199,999 | 5.8 |
| At least $200,000 | 5.2 |

**Table 3**

| Income | Frequency |
| --- | --- |
| Under $15,000 | 168 |
| $15,000 to $24,999 | 144 |
| $25,000 to $34,999 | 141 |
| $35,000 to $49,999 | 194 |
| $50,000 to $74,999 | 254 |
| $75,000 to $99,999 | 184 |
| $100,000 to $149,999 | 210 |
| $150,000 to $199,999 | 99 |
| At least $200,000 | 106 |

Does the evidence suggest that the distribution of income has changed since 2000 at the  level of significance?

Objective 1, Page 9

In Example 3, if we compare the observed and expected counts, we see where the shift in income is occurring.

Objective 1, Page 11

**Example 4 *Conducting a Goodness-of-Fit Test***

An obstetrician wants to know whether the proportion of children born on each day of the week are the same. She randomly selects 500 birth records and obtains the data shown in Table 4 (Data from Vital Statistics of the United States).

Is there reason to believe that the day on which a child is born occurs with equal frequency at the  level of significance?

**Table 4**

| **Day of Week** | **Frequency** |
| --- | --- |
| Sunday | 46 |
| Monday | 76 |
| Tuesday | 83 |
| Wednesday | 81 |
| Thursday | 81 |
| Friday | 80 |
| Saturday | 53 |

Objective 1, Page 13

Goodness-of-Fit tests cannot be used to test whether sample data follow a specific distribution. We can only say the data is consistent with a distribution stated in the null hypothesi

## 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

#### Objective 1: Perform a Test for Independence

Objective 1, Page 1

1. What do the data presented in a contingency table measure?

Objective 1, Page 2

1. What is the chi-square test for independence used to determine? Describe the null and alternative hypotheses for a chi-square test for independence.

Objective 1, Page 3

**Example 1 *Determining the Expected Counts in a Test for Independence***

Is there a relationship between marital status and happiness? The data in Table 5 show the marital status and happiness of individuals who participated in the General Social Survey. Compute the expected counts within each cell, assuming that marital status and happiness are independent.

**Table 5**

|  |  | Marital Status | Marital Status | Marital Status | Marital Status |
| --- | --- | --- | --- | --- | --- |
|  |  | Married | Widowed | Divorced/ Separated | Never Married |
| Happiness | Very Happy | 600 | 63 | 112 | 144 |
| Happiness | Pretty Happy | 720 | 142 | 355 | 459 |
| Happiness | Not Too Happy | 93 | 51 | 119 | 127 |

Objective 1, Page 4

1. State the formula for finding expected frequencies in a chi-square test for independence.

Objective 1, Page 6

1. State the test statistic for a chi-square test for independence.

Objective 1, Page 6 (continued)

1. What two conditions are needed in order to use the test statistic **?

Objective 1, Page 7

1. State the five steps for a test for independence.

Step 1

Step 2

Step 3

Step 4

Step 5

Objective 1, Page 9

**Example 2 *Performing a Chi-Square Test for Independence***

Does one's happiness depend on one's marital status? The data in Table 5 represent the marital status and disclosed level of happiness for a random sample of adult Americans who participated in the General Social Survey. Use this data to answer the question at the  level of significance.

**Table 5**

|  |  | Marital status | Marital status | Marital status | Marital Status |
| --- | --- | --- | --- | --- | --- |
|  |  | Married | Widowed | Divorced/ Separated | Never Married |
| Happiness | Very Happy | 600 | 63 | 112 | 144 |
| Happiness | Pretty Happy | 720 142 | 355 | 459 |  |
| Happiness | Not Too Happy | 93 | 51 | 119 | 127 |

Objective 1, Page 11

**Visual Support of Conclusions of Inference**

To see the association between two qualitative variables, draw bar graphs of the conditional distributions of the response variable by the explanatory variable.

Objective 1, Page 12

**Example 3 *Constructing a Conditional Distribution and Bar Graph***

Find the conditional distribution of happiness by marital status for the data in Table 5. Then draw a bar graph that represents the conditional distribution of happiness by marital status.

**Table 5**

|  |  | Marital status | Marital status | Marital status | Marital Status |
| --- | --- | --- | --- | --- | --- |
|  |  | Married | Widowed | Divorced/ Separated | Never Married |
| Happiness | Very Happy | 600 | 63 | 112 | 144 |
| Happiness | Pretty Happy | 720 142 | 355 | 459 |  |
| Happiness | Not Too Happy | 93 | 51 | 119 | 127 |

#### Objective 2: Perform a Test for Homogeneity of Proportions

Objective 2, Page 1

1. Explain what the chi-square test for homogeneity tests.

Objective 2, Page 2

1. The procedures for the test for independence and the test of homogeneity of proportions are the same. Explain how the data differ for the test for independence and the test for homogeneity differ.

Objective 2, Page 3

**Example 4 *A Test for Homogeneity of Proportions***

Zocor is a drug (Merck and Co.) that is meant to reduce the level of LDL (bad) cholesterol and increase the level of HDL (good) cholesterol. In clinical trials of the drug, patients were randomly divided into three groups. Group 1 received Zocor, group 2 received a placebo, and group 3 received cholestyramine, a cholesterol-lowering drug currently available. Table 7 contains the number of patients in each group who did and did not experience abdominal pain as a side effect. Is there evidence to indicate that the proportion of subjects in each group who experienced abdominal pain is different at the  level of significance?

**Table 7**

|  | Group 1 (Zocor) | Group 2 (placebo) | Group 3 (cholestyramine) |
| --- | --- | --- | --- |
| Number of people who experienced abdominal pain Number of people | 51 | 5 | 16 |
| who did not experience abdominal pain Data from Merck and Co. | 1532 | 152 | 163 |

Objective 2, Page 5

1. In the requirements for performing a chi-square test are not satisfied, list two options that a researcher has.

## Section 12.3 Inference about Two Population Proportions: Dependent Samples

### Objective

1. Test Hypotheses Regarding Two Proportions from Dependent Samples

Introduction, Page 1

1. What is the name of the test used to compare two proportions with matcher-pairs data?

#### Objective 1: Test Hypotheses Regarding Two Proportions from Dependent Samples

Objective 1, Page 1

McNemar’s Test will be used to determine whether the difference in sample proportions is due to sampling error or whether there the differences are significant enough to conclude that the proportions are different. The null hypothesis is , and the alternative hypothesis is .

Objective 1, Page 2

1. To use McNemar’s test, explain how to arrange the data in a contingency table.
2. What two conditions are necessary for performing McNemar’s test?

Objective 1, Page 2 (Continued)

1. State the five steps for McNemar’s test.

Step 1

Step 2

Step 3 (By Hand)

Step 3 (Using Technology)

Step 4

Step 5

Objective 1, Page 3

**Example 1 *Analyzing the Difference of Two Proportions: Dependent Samples***

A recent General Social Survey asked the following two questions of a random sample of 1492 adult Americans under the hypothetical scenario that the government suspected that a terrorist act was about to happen:

Do you believe the authorities should have the right to tap people's telephone conversations?

Do you believe the authorities should have the right to stop and search people on the street at random?

The results of the survey are shown in Table 9.

**Table 9**

|  |  | Random Stop | Random Stop |
| --- | --- | --- | --- |
|  |  | Agree | Disagree |
| Tap Phone | Agree | 494 | 335 |
| Tap Phone | Disagree | 126 | 537 |

Do the proportions who agree with each scenario differ significantly? Use the  level of significance.