Chapter 11: Inference for Distributions of Categorical Data

Section 11.2
Inference for Relationships

The Practice of Statistics, 4th edition – For AP*
STARNES, YATES, MOORE

Chapter 11
Inference for Distributions of Categorical Data

- 11.1 Chi-Square Goodness-of-Fit Tests
- **11.2 Inference for Relationships**

Section 11.2
Inference for Relationships

Learning Objectives

After this section, you should be able to…

- ✓ COMPUTE expected counts, conditional distributions, and contributions to the chi-square statistic
- ✓ CHECK the Random, Large sample size, and Independent conditions before performing a chi-square test
- ✓ PERFORM a chi-square test for **homogeneity** to determine whether the distribution of a categorical variable differs for several populations or treatments
- ✓ PERFORM a chi-square test for **association/independence** to determine whether there is convincing evidence of an association between two categorical variables
- ✓ EXAMINE individual components of the chi-square statistic as part of a follow-up analysis
- ✓ INTERPRET computer output for a chi-square test based on a two-way table

Introduction

There are two types of chi-square tests for inference for relationships, (homogeneity and independence). Both the chi-square test for homogeneity and the chi-square test for association/independence start with a two-way table of observed counts. They even calculate the test statistic, degrees of freedom, and *P*-value in the same way. *The questions that these two tests answer are different, however.*

- A chi-square test for **homogeneity** tests whether the distribution of a categorical variable is the same for each of several populations or treatments.
- The chi-square test for **association/independence** tests whether two categorical variables are associated in some population of interest.

Instead of focusing on the question asked, it's much easier to look at how the data were produced.

✓If the data come from two or more independent random samples or treatment groups in a randomized experiment, then do a chi-square test for homogeneity.

✓If the data come from a single random sample, with the individuals classified according to two categorical variables, use a chi-square test for association/independence.

Example: Comparing Conditional Distributions

Market researchers suspect that background music may affect the mood and buying behavior of customers. One study in a supermarket compared three randomly assigned treatments: no music, French accordion music, and Italian string music. Under each condition, the researchers recorded the numbers of bottles of French, Italian, and other wine purchased. Here is a table that summarizes the data:

	Music			
Wine	**None**	**French**	**Italian**	**Total**
French	30	39	30	**99**
Italian	11	1	19	**31**
Other	43	35	35	**113**
Total	**84**	**75**	**84**	**243**

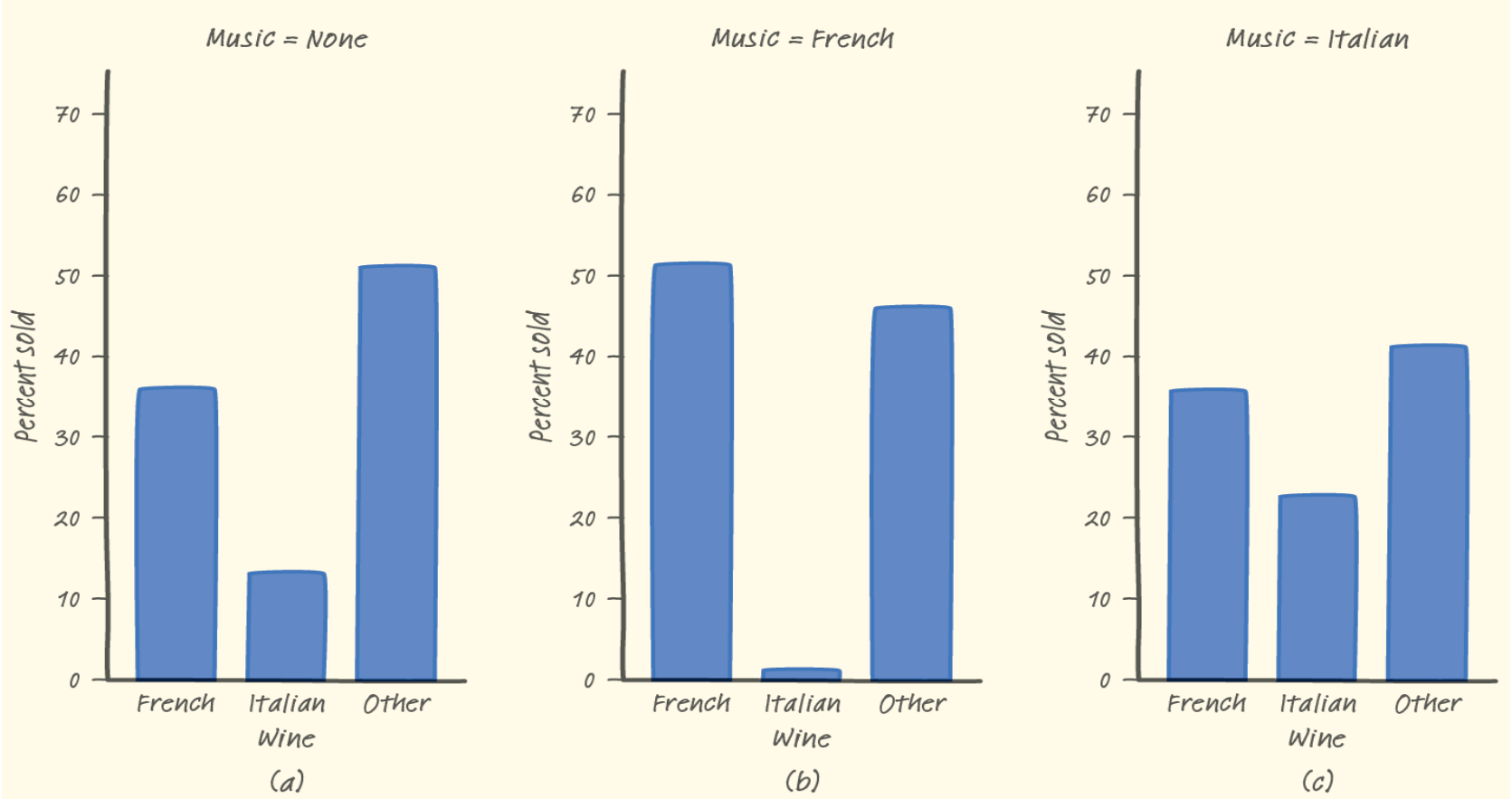
PROBLEM:

(a) Calculate the conditional distribution (in proportions) of the type of wine sold for each treatment.

(b) Make an appropriate graph for comparing the conditional distributions in part (a).

(c) Are the distributions of wine purchases under the three music treatments similar or different? Give appropriate evidence from parts (a) and (b) to support your answer.

Example: Comparing Conditional Distributions



The type of wine that customers buy seems to differ considerably across the three music treatments. Sales of Italian wine are very low (1.3%) when French music is playing but are higher when Italian music (22.6%) or no music (13.1%) is playing. French wine appears popular in this market, selling well under all music conditions but notably better when French music is playing. For all three music treatments, the percent of Other wine purchases was similar.

The Chi-Square Test for Homogeneity

Chi-Square Test for Homogeneity

Suppose the Random, Large Sample Size, and Independent conditions are met. You can use the **chi-square test for homogeneity** to test

H_0: There is no difference in the distribution of a categorical variable for several populations or treatments.
H_a: There is a difference in the distribution of a categorical variable for several populations or treatments.

Start by finding the expected counts. Then calculate the chi-square statistic

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

where the sum is over all cells (not including totals) in the two-way table. If H_0 is true, the χ^2 statistic has approximately a chi-square distribution with degrees of freedom = (number of rows – 1) (number of columns - 1). The *P*-value is the area to the right of χ^2 under the corresponding chi-square density curve.

Example: Does Music Influence Purchases?

State:

H_0: There is no difference in the distributions of wine purchases at this store when no music, French accordion music, or Italian string music is played.

H_a: There is a difference in the distributions of wine purchases at this store when no music, French accordion music, or Italian string music is played.

Do:

Finding Expected Counts

The expected count in any cell of a two-way table when H_0 is true is

$$\text{expected count} = \frac{\text{row total} \cdot \text{column total}}{\text{table total}}$$

The values in the calculation are the row total for French wine, the column total for no music, and the table total. We can rewrite the original calculation as:

Observed Counts

	Music			
Wine	**None**	**French**	**Italian**	**Total**
French	30	39	30	**99**
Italian	11	1	19	**31**
Other	43	35	35	**113**
Total	**84**	**75**	**84**	**243**

$$\frac{\bullet}{\qquad} = 34.22$$

Calculating The Chi-Square Statistic

The tables below show the observed and expected counts for the wine and music experiment. Calculate the chi-square statistic.

Observed Counts

	Music			
Wine	**None**	**French**	**Italian**	**Total**
French	30	39	30	**99**
Italian	11	1	19	**31**
Other	43	35	35	**113**
Total	**84**	**75**	**84**	**243**

Expected Counts

	Music			
Wine	**None**	**French**	**Italian**	**Total**
French	34.22	30.56	34.22	**99**
Italian	10.72	9.57	10.72	**31**
Other	39.06	34.88	39.06	**113**
Total	**84**	**75**	**84**	**243**

For the French wine with no music, the observed count is 30 bottles and the expected count is 34.22. The contribution to the χ^2 statistic for this cell is

$$\frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} = \frac{(30 - 34.22)^2}{34.22} = 0.52$$

The χ^2 statistic is the sum of nine such terms:

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} = \frac{(30 - 34.22)^2}{34.22} + \frac{(39 - 30.56)^2}{30.56} + \ldots + \frac{(35 - 39.06)^2}{39.06}$$

$$= 0.52 + 2.33 + \ldots + 0.42 = 18.28$$

Example: Does Music Influence Purchases?

To find the *P*-value, use χ^2cdf with df=(3-1)(3-1)=4.

So, the *P*-value for a test based on our sample data is χ^2cdf(18.28, 1e99, 4) = .00109.

Conclude:

The small *P*-value gives us convincing evidence to reject H_0 and conclude that there is a difference in the distributions of wine purchases at this store when no music, French accordion music, or Italian string music is played. Furthermore, the random assignment allows us to say that the difference is caused by the music that's played.

Follow-up Analysis

The chi-square test for homogeneity allows us to compare the distribution of a categorical variable for any number of populations or treatments. If the test allows us to reject the null hypothesis of no difference, we then want to do a follow-up analysis that examines the differences in detail.

Start by examining which cells in the two-way table show large deviations between the observed and expected counts. Then look at the individual components to see which terms contribute most to the chi-square statistic.

Minitab output for the wine and music study displays the individual components that contribute to the chi-square statistic.

Looking at the output, we see that just two of the nine components that make up the chi-square statistic contribute about 14 (almost 77%) of the total $\chi^2 = 18.28$.

We are led to a specific conclusion: *sales of Italian wine are strongly affected by Italian and French music*.

Chi-Square Test: None, Franch, Italian

Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

	None	French	Italian	Total
1	30	39	30	99
	34.22	30.56	34.22	
	0.521	2.334	0.521	
2	11	1	19	31
	10.72	9.57	10.72	
	0.008	7.672	6.404	
3	43	35	35	113
	39.06	34.88	39.06	
	0.397	0.000	0.422	
Total	84	75	84	243

Chi-Sq = 18.279, DF = 4, P-Value = 0.001

Comparing Several Proportions

Many studies involve comparing the proportion of successes for each of several populations or treatments.

- The two-sample z test from Chapter 10 allows us to test the null hypothesis H_0: $p_1 = p_2$, where p_1 and p_2 are the actual proportions of successes for the two populations or treatments.
- The chi-square test for homogeneity allows us to test H_0: $p_1 = p_2 = \ldots = p_k$. This null hypothesis says that there is no difference in the proportions of successes for the k populations or treatments. The alternative hypothesis is H_a: at least two of the p_i's are different.

Caution:
Many students *incorrectly state* H_a as "all the proportions are different."

Think about it this way: the opposite of "all the proportions are equal" is "some of the proportions are not equal."

The Chi-Square Test for Association/Independence

Chi-Square Test for Association/Independence

Suppose the Random, Large Sample Size, and Independent conditions are met. You can use the **chi-square test for association/independence** to test

H_0: There is no association between two categorical variables in the population of interest.
H_a: There is an association between two categorical variables in the population of interest.

Or, alternatively

H_0: Two categorical variables are independent in the population of interest.
H_a: Two categorical variables are dependent in the population of interest.

Start by finding the expected counts. Then calculate the chi-square statistic

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

where the sum is over all cells (not including totals) in the two-way table. If H_0 is true, the χ^2 statistic has approximately a chi-square distribution with degrees of freedom = (number of rows – 1) (number of columns - 1). The *P*-value is the area to the right of χ^2 under the corresponding chi-square density curve.

Relationships Between Two Categorical Variables

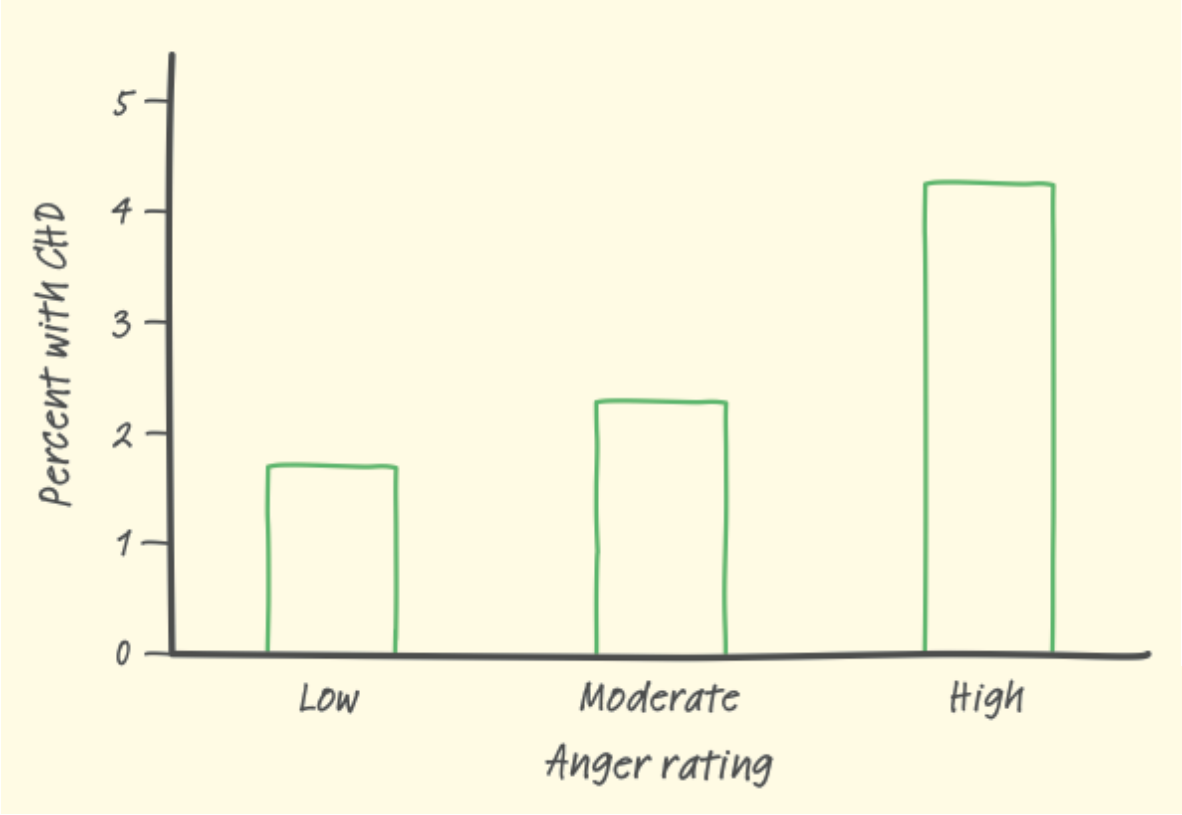
Another common situation that leads to a two-way table is when a *single* random sample of individuals is chosen from a *single* population and then classified according to two categorical variables. In that case, our goal is to analyze the relationship between the variables.

A study followed a random sample of 8474 people with normal blood pressure for about four years. All the individuals were free of heart disease at the beginning of the study. Each person took the Spielberger Trait Anger Scale test, which measures how prone a person is to sudden anger. Researchers also recorded whether each individual developed coronary heart disease (CHD). This includes people who had heart attacks and those who needed medical treatment for heart disease. Here is a two-way table that summarizes the data:

	Low anger	Moderate anger	High anger	Total
CHD	53	110	27	190
No CHD	3057	4621	606	8284
Total	3110	4731	633	8474

Example: Angry People and Heart Disease

We're interested in whether angrier people tend to get heart disease more often. We can compare the percents of people who did and did not get heart disease in each of the three anger categories:



There is a clear trend: as the anger score increases, so does the percent who suffer heart disease. A much higher percent of people in the high anger category developed CHD (4.27%) than in the moderate (2.33%) and low (1.70%) anger categories.

The Chi-Square Test for Association/Independence

We often gather data from a random sample and arrange them in a two-way table to see if two categorical variables are associated. The sample data are easy to investigate: turn them into percents and look for a relationship between the variables.

Our null hypothesis is that there is no association between the two categorical variables. The alternative hypothesis is that there is an association between the variables. For the observational study of anger level and coronary heart disease, we want to test the hypotheses

H_0: There is no association between anger level and heart disease in the population of people with normal blood pressure.
H_a: There is an association between anger level and heart disease in the population of people with normal blood pressure.

No association between two variables means that the values of one variable do not tend to occur in common with values of the other. That is, the variables are independent. An equivalent way to state the hypotheses is therefore

H_0: Anger and heart disease are independent in the population of people with normal blood pressure.
H_a: Anger and heart disease are dependent in the population of people with normal blood pressure.

Example: Angry People and Heart Disease

Here is the complete table of observed and expected counts for the CHD and anger study side by side. Do the data provide convincing evidence of an association between anger level and heart disease in the population of interest?

State: We want to perform a test of

H_0: There is no association between anger level and heart disease in the population of people with normal blood pressure.

H_a: There is an association between anger level and heart disease in the population of people with normal blood pressure.

We will use $\alpha = 0.05$.

Example: Angry People and Heart Disease

Plan: If the conditions are met, we should conduct a chi-square test for association/independence.

- *Random* The data came from a random sample of 8474 people with normal blood pressure.
- *Large Sample Size* All the expected counts are at least 5, so this condition is met.
- *Independent* Knowing the values of both variables for one person in the study gives us no meaningful information about the values of the variables for another person. So individual observations are independent. Because we are sampling without replacement, we need to check that the total number of people in the population with normal blood pressure is at least 10(8474) = 84,740. This seems reasonable to assume.

Example: Angry People and Heart Disease

Do: Since the conditions are satisfied, we can perform a chi-test for association/independence. We begin by calculating the test statistic.

	Observed			Expected		
	Low	**Moderate**	**High**	**Low**	**Moderate**	**High**
CHD	53	110	27	69.73	106.08	14.19
No CHD	3057	4621	606	3040.27	4624.92	618.81

Test statistic:

$$\chi^2 = \sum \frac{(\text{Observed-Expected})^2}{\text{Expected}} = 16.077$$

***P*-Value**:

The two-way table of anger level versus heart disease has 2 rows and 3 columns. We will use the chi-square distribution with df = (2 - 1)(3 - 1) = 2 to find the *P*-value.

The command χ^2cdf(16.077,1e99,2) gives 0.00032.

Conclude: Because the *P*-value is clearly less than $\alpha = 0.05$, we reject H_0 and conclude that anger level and heart disease are associated in the population of people with normal blood pressure.

Section 11.2
Inference for Relationships

Summary

In this section, we learned that…

- ✓ We can use a two-way table to summarize data on the relationship between two categorical variables. To analyze the data, we first compute percents or proportions that describe the relationship of interest.

- ✓ If data are produced using independent random samples from each of several populations of interest or the treatment groups in a randomized comparative experiment, then each observation is classified according to a categorical variable of interest. The null hypothesis is that the distribution of this categorical variable is the same for all the populations or treatments. We use the **chi-square test for homogeneity** to test this hypothesis.

- ✓ If data are produced using a single random sample from a population of interest, then each observation is classified according to two categorical variables. The **chi-square test of association/independence** tests the null hypothesis that there is no association between the two categorical variables in the population of interest. Another way to state the null hypothesis is H_0:The two categorical variables are independent in the population of interest.

Section 11.1
Chi-Square Goodness-of-Fit Tests

Summary

- The expected count in any cell of a two-way table when H_0 is true is

$$\text{expected count} = \frac{\text{row total} \cdot \text{column total}}{\text{table total}}$$

- The chi-square statistic is

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

where the sum is over all cells in the two-way table.

- The chi-square test compares the value of the statistic χ^2 with critical values from the chi-square distribution with df = (number of rows - 1)(number of columns - 1). Large values of χ^2 are evidence against H_0, so the *P*-value is the area under the chi-square density curve to the right of χ^2.

Section 11.1
Chi-Square Goodness-of-Fit Tests

Summary

- ✓ The chi-square distribution is an approximation to the distribution of the statistic χ^2. You can safely use this approximation when all expected cell counts are at least 5 (the Large Sample Size condition).

- ✓ Be sure to check that the Random, Large Sample Size, and Independent conditions are met before performing a chi-square test for a two-way table.

- ✓ If the test finds a statistically significant result, do a follow-up analysis that compares the observed and expected counts and that looks for the largest components of the chi-square statistic.

Looking Ahead…

Homework…

Chapter 11: